Competition platforms

Andrey Ustyuzhanin ANDREY.USTYUZHANIN@CONSTRUCTOR.ORG
Constructor University, Bremen,
Campus Ring 1, 28759, Germany

Harald Carlens HARALD@MLCONTESTS.COM
ML Contests

Reviewed on OpenReview:

Abstract

The ecosystem of artificial intelligence contests is diverse and multifaceted, encompassing several platforms that each host numerous competitions and challenges annually, alongside many specialized websites dedicated to individual contests. These platforms manage the overarching administrative responsibilities inherent in orchestrating contests, thus allowing organizers to allocate greater attention to other aspects of their contests. Notably, these platforms exhibit considerable variety in their features, economic models, and communities. This chapter conducts an extensive review of the leading services in this space and explores alternative methods facilitating the independent hosting of such contests. We provide hints and tips on choosing the right platform for your challenge at the end.

Keywords: competition platform, challenge hosting services, service comparison

1 Platforms for AI contests

The majority[1] of AI contest organisers use a third-party platform to host their contest rather than building and maintaining their own infrastructure.

The choice of platform is driven by various considerations. Before we introduce these, and the role we expect a platform to fulfil, it's helpful to return to a definition of the types of contests we're considering. We can ground our expectations of a platform in the Common Task Framework (Donoho, 2017) (CTF), which lays out the key ingredients of an AI challenge:

1. A publicly available training dataset, involving a list of feature measurements and a class label for each observation;

2. A set of enrolled competitors whose common task is to infer a class prediction rule from the training data;

3. A scoring referee, to which competitors can submit their prediction rule. The referee runs the prediction rule against a testing dataset which is not made available to competitors. The referee objectively and automatically reports the score achieved by the submitted rule.

1. 317 of 367 contests in 2023 were hosted on a third-party platform (Carlens, 2024). The universe of contests considered here includes only those with meaningful prize money (over $1,000) or a conference affiliation.

While these "ingredients" are somewhat specific to supervised learning problems, it is not too difficult to see how they would generalise to other fields - such as reinforcement learning style challenges, where data sets are replaced with environments. They also generalise to our broader definition of "contests", which includes these measurable challenges as well as competitions where performance is evaluated by a panel of judges[2]. In the case of a subjectively-judged competition, the judged output generally includes a written document or working prototype in addition to, or instead of, a simple prediction rule. From the above ingredients we can get a list of responsibilities, to be shared between organisers and platforms:

- **Design**: framing a problem in a way that is amenable to a CTF-style contest
- **Data**: gathering and cleaning data for training and test datasets
- **Discovery**: notifying potential participants
- **Admin**: publishing rules and making training data available
- **Engagement**: enabling participants to be productive and to collaborate
- **Scoring**: accepting submissions, evaluating them, and updating leaderboards
- **Dissemination**: sharing insights from submissions and contest outcomes

It is possible for all of these responsibilities to be undertaken by the contest organisers, or for them all to be outsourced to a platform, but in most cases the responsibilities are shared. The decision of which responsibilities are to be outsourced is of primary importance in the choice of platform, as some are better suited to certain responsibilities than others[3]. Secondly, the exact requirements for each responsibility will further determine the choice of platform [4]. The remaining components of platform choice come down to budget, familiarity with different platforms, and geographical considerations.

2. Throughout this article, we aim to conform to the nomenclature used by the other chapters in the "AI Competitions and Benchmarks" book, of which this article will make up Chapter 10. The common definitions for contest, competition, challenge, and benchmark are as follows:

   Contest: A contest is an event created by organizers, governed by rules, and directed to a group of participants, offering the opportunity to win an award or a prize.

   Competition: A competition is a skill-based scientific contest, with a limited time duration, involving the submission of proposals, project propositions, project outcomes, and/or prototypes that are evaluated by a panel of judges.

   Challenge: A challenge is a skill-based scientific contest, with a limited time duration, ending by a total ranking of participants according to a pre-defined scoring metric, and the selection of winners.

   Benchmark: A benchmark refers to on-going evaluations of methods or models in well-defined conditions, for the purpose of making standardized comparisons.

   There will be occasional inevitable exceptions, including where products use names that conflict with our definitions - e.g. "Kaggle Competitions".
3. For example, Codabench is able to fulfil most of these, but does not provide support with design or data preparation.
4. For example, is the contest aiming to reach a broad audience, or is it targeted at a niche research community who can all be reached through a single mailing list? Is it a straightforward supervised learning problem, or does it incorporate adversarial elements or reinforcement-learning style environments for scoring?

With these responsibilities in mind, in the next section we lay out more detailed criteria and review the main features of several leading platforms:

- AIcrowd (Mohanty et al., 2017),
- Codabench (Xu et al., 2022) by Université Paris-Saclay,
- CodaLab (Pavao et al., 2023) by Université Paris-Saclay,
- DrivenData (DrivenData, 2014),
- EvalAI (Yadav et al., 2019) by CloudCV,
- Kaggle (Goldbloom and Hamner, 2010) by Alphabet Inc,
- Tianchi (Group, 2014) by Alibaba,
- Zindi (Zindi, 2018).

This list is not intended to be comprehensive, and is focused on generalist platforms with active communities as of the end of 2022. We give a separate overview of several non-generalist platforms that target specific domains or follow a complementary pattern that doesn't strictly adhere to the CTF setup, as well as some non-English language platforms.

2 Platform comparison criteria

We outline the main characteristics that we use for comparison, which are provided roughly in order of the responsibilities listed above.

Design support: platforms vary in the amount of support they are able to provide to organisers in designing a contest. Here we are defining the "design" process to cover the initial problem formulation, decisions around the train/test split, and choice of evaluation metrics. This tends to be most important for companies with little in-house data science expertise, and not so relevant for researchers with specific problems in mind.

Data support: some platforms help organisers gather and clean data, as well as transforming it into a format that is convenient for participants to use.

Registered users: the total number of users registered on a platform gives a good indication of the size of the audience that can be reached. This is particularly important for organisers looking to reach a broad audience, or to reach participants who are not already familiar with the problem area or organiser.

Code sharing: some platforms allow structured code-sharing through notebooks which can be hosted and executed on the platform. Others allow participants to embed their solution as an external notebooks or code repositories. This functionality allows other participants to easily reproduce and build on others' solutions. Open community collaboration in this way can be a valuable feature for complicated or novel challenges.

Submission code evaluation: the most straightforward way to run a challenge is to ask participants to submit a set of predictions, and compare those against some "ground truth" values using a loss metric. Many platforms allow for challenges where participants submit code that is then run on the platform side to generate predictions against unseen

data. This allows organisers to do things like impose compute budget constraints on submissions, and vet submissions for compliance with the rules. It also changes the nature of the challenge, since participants have less knowledge about the distribution of test set examples than they do in the case where they have access to test set features. Most platforms that support code submissions can support it in any language, though support for Python and R tends to be better than for other languages.

Custom metrics: some platforms or offerings are able to support only "common" metrics like mean squared error or cross-entropy loss. The ability to implement custom metrics is important for many challenges, especially those looking to capture particular trade-offs. Some platforms allow choosing just one among many predefined metrics; some allow for custom implementations. Some platforms charge an additional cost for implementing non-standard metrics.

Staged contests: contests from within a niche domain can initially look inscrutable to the wider community, and it can help to split the contest into smaller chunks of gradually increasing complexity. A preliminary trial stage can also help to mitigate risks of data leakage. In addition to this, it has been shown that pre-filtering participants in a trial run can help reduce over-fitting on winner selection. (Pavao et al., 2022)

Private evaluation: Data privacy is a sensitive issue. Some platforms allow participants' solutions evaluation using an organizer's dedicated machines. With the help of such a feature, one can set up a challenge without needing to share restricted code or datasets with anyone, even with platform owners. This feature can also be used to support unusual evaluation procedures - for example, those needing to run on specific hardware managed by organisers, or on a physical robot in a lab.

Reinforcement Learning (RL) evaluation: running participants' RL agents on the platform's side is inherently more complex than running a metric evaluation script across a vector of predictions, and not all platforms support this. Computational cost for these types of challenges is not only often higher than for supervised learning problems, but also more unpredictable - since RL evaluation episode lengths can depend on the success of an agent. Supporting multi-agent environments or tournament-style evaluations are an additional challenge, and we do not evaluate this ability in our analysis.

Judging panel: some contests do not use a simple scoring metric, and instead are evaluated by a panel of judges. These competitions sometimes incorporate more open-ended elements of data exploration and discovery, or require participants to develop prototype solutions or products which are not easily evaluated in an automated and measurable way.

Human-in-the-loop (HITL) evaluation : some contests do not have ground-truth labels in the data, and require large-scale human evaluation for comparison. For example, a dialogue bot evaluation requires communication with a living person. Some platforms enable the use of human-evaluation platforms, such as Amazon Mechanical Turk (see below).

Run for free: most platforms charge a fee. The exact cost usually depends on the range of services offered. Some platforms offer a free "self-service" offering, allowing organisers to set up a completely self-managed contest.

Open-source: for some platforms, the code that runs them is open-source. In most cases a platform fulfils a service, and organisers do not have an interest in changing the platform's functionality. However, being able to access a platform's source code can help organisers assess the pace of development on the platform, verify details of the platform

Criteria	AIcrowd	Codabench	CodaLab	DrivenData	EvalAI	Kaggle	Tianchi	Zindi
Design support	✓	-	-	✓	-	✓	✓	✓
Data support	✓	-	-	✓	-	✓	✓	✓
Registered users	140k+	5k+	55k+	100k+	40k+	16m+	1.4m+	70k+
Code sharing	✓	✓	✓	✓[6]	✓	✓	✓	✓
Code evaluation	✓	✓	✓	✓	✓	✓	✓	-
Custom metrics	✓	✓	✓	✓	✓	✓	✓	✓
Staged contests	✓	✓	✓	✓	✓	-	✓	-
Private evaluation	✓	✓	✓	-	✓	-	✓	-
RL-friendly	✓	✓	✓	-	✓	✓	-	✓[7]
Judging panel	✓	-	-	✓	-	✓	✓	✓
HITL evaluation	✓	-	-	-	✓	-	-	-
Run for free	-	✓	✓	-	✓	✓[8]	✓	✓[9]
Open-source	-	✓	✓[10]	-	✓[11]	-	-	-
Established	2017	2023	2013	2014	2017	2010	2014	2018

Table 1: Platform overview

evaluation mechanics, or allow organisers to run their instance on their own premises for a local event with private datasets. It also allows organisers to add features to the platform themselves.

3 Platform Comparison

An overview of platforms as measured by the criteria above is presented in Table 1[5]. Features which we were able to verify as being supported are marked as ✓, and where possible these were confirmed with the team running the platform. In some cases where we could not find public documentation of a feature and we did not receive any response from the platform operators, it is possible that we have incorrectly marked features as unavailable. Estimates for the number of users and typical number of entries reflect activity in 2023 (Carlens, 2024).

Here are some highlights of the platforms included in the comparison.

AIcrowd started as a research project at EPFL, and has since run a large variety of competitions. It has hosted several official NeurIPS competitions including many reinforcement learning challenges.

Codabench is an open-source platform, with an instance maintained by Université Paris-Saclay. Anyone can sign up and host or take part in a contest. Free CPU resources are available for inference, and organisers can supplement this with their own hardware. Codabench is friendly to a variety of challenges: from online data science classes/hackathons to contests affiliated with leading conferences, and can also be used for ongoing benchmarks. Codabench is suitable to organisers who have a clear idea of the contest they want to run, and can be self-sufficient when it comes to technical and marketing aspects.

5. One of the authors maintains an updated version of this table at `https://mlcontests.com/platforms/`.

CodaLab is the predecessor of Codabench, and is maintained by the same team. Where possible, the team recommends that organisers use the newer Codabench platform. The only exception to this is for challenges which require ranking participants based on an aggregate of multiple different scores, a feature which is supported by CodaLab but not yet by Codabench as of the time of writing.

DrivenData focuses on running contests with social impact, and has run competitions for NASA and other organisations. DrivenData stands out for its thorough reports detailing participants' approaches, and permissively licensed solution code publication[12].

EvalAI is built by a team of open source enthusiasts working at CloudCV, a consulting company that aims to make AI research reproducible and easily accessible. With the platform's help, they reduce the entry barrier for research and make it easier for researchers, students, and developers to design and use state-of-the-art algorithms as a service. It is known for running many competitions involving human-in-the-loop evaluations.

Kaggle was acquired by Google in 2017 and has the largest community of all the platforms, with over 16 million registered users. As well as hosting contests, Kaggle allows users to host datasets, notebooks, and models. Kaggle's progression system [13] provides additional incentives for users users to compete, collaborate, share code, and contribute to community discussions. It is possible to run a "Community Competition" for free, with limitations around discoverability, evaluation metrics, and participant incentives.

Tianchi is a platform run by Alibaba, including running kernels and earning points. Contests can quickly gain several thousand participants. The primary audience is Chinese, though many contests also include English documentation.

Zindi is focused on connecting organisations with data scientists in Africa. As well as online contests, Zindi also runs in-person hackathons and community events.

The table above only lists a few of the largest existing platforms. These are some other general-purpose platforms worth exploring:

bitgrit[14] is an AI contest and recruiting platform founded in 2017, with over 55,000 registered users.

Hugging Face launched its Competitions[15] platform in February 2023, alongside its well-established Model Hub and widely-used open source machine learning repositories.

Humyn.ai[16] hosts contests as well as facilitating deeper engagements between businesses and its user-base of data scientists.

There is a long tail of platforms, and we expect that there are other relevant platforms which are as yet unknown to us.

4 Non-English language platforms

The comparison above is focused on English-language platforms. While the authors are less familiar with platforms in other languages, this section is an attempt at covering platforms in regions where the main common language of their audience is not English. As already

12. https://github.com/drivendataorg/competition-winners
13. https://www.kaggle.com/progression
14. https://bitgrit.net/competition/
15. https://huggingface.co/competitions
16. https://humyn.ai

mentioned, the most notable **Chinese** platform is Tianchi. Other Chinese platforms worth mentioning are: Data Castle[17], Kesci[18], Bien Data[19], and Data Fountain[20]. The **Japanese** platform Signate[21] and the company behind it collaborate with industries, government agencies, and research institutes in various domains to resolve social issues. The **Russian** community, Open Data Science[22] runs contests, as well as including organizing events and finding joint projects for researchers, engineers, and developers around Data Science. Other Russian websites listing contests include DS Works[23] and Yandex Cup[24]. All these platforms above have a reasonably developed community; however, to join those, one needs to be fluent in the corresponding language.

5 Domain-specific platforms

Several platforms host regular challenges on domains in a specific branch of science or industry, or within a more narrow scope than the Common Task Framework. We list a few examples here.

DREAM Challenges[25] has been running biomedical challenges since 2006, with now over 30,000 registered users.

Grand Challenge[26] is a platform for the end-to-end development of machine learning solutions in biomedical imaging. It has successfully run over a hundred challenges, and allows researchers to host custom algorithms that can be used for performance assessment on new datasets and crowd-sourcing activities called *reader studies*.

Makridakis Open Forecasting Center (MOFC)[27] conducts research on forecasting, and has been running the "M Competitions", a series of forecasting challenges, since 1987. The most recent M Competition was the M6 Financial Forecasting Competition, running from 2022 until 2023.

NASA Tournament Lab[28] (NTL) facilitates the use of crowd-sourcing to tackle NASA challenges. NASA's researchers, scientists, and engineers have launched numerous crowd-sourcing projects through the NTL, seeking novel ideas or solutions to accelerate research and development efforts in support of the NASA mission. The NTL offers a variety of open innovation platforms that engage the crowd-sourcing community to improve solutions for specific, real-world problems being faced by NASA and other Federal Agencies.

Numerai[29] is a fund that draws its strategy from crowd-sourced predictions submitted to regular tournaments. Participants aim to predict stock market movements from obfuscated data. Numerai states that it has paid out over $48m to its data scientist collab-

17. https://challenge.datacastle.cn/v3/cmptlist.html
18. https://www.kesci.com/
19. https://www.biendata.xyz/
20. https://www.datafountain.cn/
21. https://signate.jp/
22. https://ods.ai/
23. https://dsworks.ru/
24. https://yandex.com/cup/ml/
25. https://dreamchallenges.org
26. https://grand-challenge.org/
27. https://trustii.io
28. https://www.nasa.gov/coeci/ntl
29. https://numer.ai/

orators. It is worth noting that reward eligibility in Numerai tournaments requires staking Numerai's NMR cryptocurrency token, exposing participants to potential losses, unlike most other platforms listed here.

Onward[30] is a platform run by Shell, which is focused on enabling innovation in the energy sector. Many of the contests run on this platform so far have been targeted at solving specific business problems, and so the winning solutions are not generally shared publicly.

Solafune[31] was founded in 2020, and focuses on competitions using satellite and geospatial data.

Trustii[32] is a platform established in 2020, primarily focused on healthcare competitions. Winners' code and solutions are shared on GitHub.

Unearthed[33] is a platform that hosts contests aimed at making the energy and resources industry more efficient and sustainable. Challenges often involve a mixture of domain knowledge and data science skills.

6 Alternative approaches and adjacent services

The platforms above are the most notable ones implementing contests broadly in line with the Common Task Framework (Donoho, 2017). However, they are far from the only options for collaborative research. Below is a list of platforms and services that rely on different assumptions and implement interaction protocols that turn out to be suitable for research goals in some scientific domains, or that can aid in running CTF-style competitions in a role other than a competition platform.

Amazon Mechanical Turk (AMT)[34]**:** a marketplace for completion of virtual tasks that require human intelligence. Businesses or academic researchers regularly use it to label data that can later

DataCamp[35] a data science education platform which hosts occasional competitions targeted at beginners.

Dynabench[36]**:** a platform for dynamic data collection and benchmarking that aims to address issues with static benchmarks through human-in-the-loop benchmarking.

InnoCentive[37]**:** is an innovative hub for a new kind of problem-solving. It describes the framework of "Challenge Driven Innovation" (CDI) that helps reformulate a task or opportunity at hand into a series of modules or challenges addressed later by a network of participants. CDI framework is much broader than CTF. Thus Innocentive enjoys various challenges, including Brainstorming, Design, Prototyping, and Algorithm development. The platform has been around for over a decade. It links over half a million solvers and spans dozens of industries.

30. https://thinkonward.com
31. https://solafune.com
32. https://trustii.io
33. https://www.unearthed.solutions
34. https://www.mturk.com/
35. https://www.datacamp.com
36. https://dynabench.org/
37. https://www.innocentive.com/

LMSYS Chatbot Arena[38] is a crowd-sourced open platform for evaluating large language models through pairwise comparison.

Google Colab[39]**:** a hosted notebook solution with support for CPU/GPU/TPU accelerators and sharing via GitHub or Google Drive, Google's Colab service enables interactive code-sharing and eases reproducibility. It significantly lowers the bar for researchers to interact with code or libraries that are not within their domain of expertise, by enabling them to run and edit code without needing to worry about maintaining environments or installing libraries. It can be a useful place for organisers to share code examples with potential participants, allowing them a frictionless way to explore a contest.

ML Experiment Tracking Tools: Tools like MLflow[40] (open source), W&B[41], Comet[42], Neptune[43] enable distributed research teams to easily share their experimental results within their team or to a public audience. These can serve as a more useful record of experimental results than local tools like TensorBoard or simple text-based logs. be used for training ML algorithms. AMT has been around for more than 15 years. Major companies like Google and Microsoft have similar versions of such marketplaces.

ML Collective[44]: (MLC) is an independent, nonprofit organization with a mission to make research opportunities accessible and free by supporting open collaboration in machine learning (ML) research. Jason Yosinski and Rosanne Liu founded MLC at Uber AI Labs in 2017 and, in 2020, it moved outside Uber. The group aims to build a culture of open, cross-institutional research collaboration among researchers of diverse and non-traditional backgrounds. Thus, the outcome of the cooperation is the natural growth of participating researchers through discussion and publishing process participation. As of mid-2022, the community is more than 3 thousand ML researchers sharing collaborative research values.

ML Contests[45]**:** is primarily a contest discovery platform, with a listing page that shows currently active contests across many platforms. Organisers can add their contest to the listing page for free. Alongside this, ML Contests also publishes research on competitive machine learning.

MLCommons[46]**:** an AI engineering consortium, bringing together groups from industry and academia to foster collaboration and set standards. As well as maintaining the MLPerf benchmarks, MLCommons runs working groups on benchmarks, AI safety, data, and research.

OpenChallenges[47]**:** a centralised hub for biomedical challenges across various platforms, maintained by Sage Bionetworks.

OpenML[48]**:** an online machine learning platform for sharing and organizing data, machine learning algorithms, and experiments. Thus they have created a service that allows

38. `https://chat.lmsys.org/`
39. `https://colab.research.google.com/`
40. `https://mlflow.org/`
41. `https://wandb.ai/`
42. `https://www.comet.com/`
43. `https://neptune.ai/`
44. `https://mlcollective.org/`
45. `https://mlcontests.com`. Note: ML Contests is maintained by one of the authors of this article.
46. `https://mlcommons.org/`
47. `https://openchallenges.io`
48. `https://www.openml.org/`

running an algorithm across several datasets and systematically comparing its performance. While there are no private leaderboards, every check is systematically performed via system API and protocol. Thus new experiments are immediately compared to state of the art without always having to rerun other people's experiments. The recent development of OpenML involves the design of an AutoML evaluation framework for a broad spectrum of datasets.

Papers With Code[49]**:** organizes access to scientific papers from the leading Machine Learning conferences and links to known implementations of the methods described in such articles. The service also compares different methods of solving several tasks in the form of a leaderboard where entries are linked to particular implementations. The diversity of such leaderboards has grown immensely in the past few years. With the help of this platform, one can find the most current state of the art to the problem of interest and read details of the method in the companion paper.

Seasonal events: there are many yearly data analysis events organized around the world. Usually, those are hosted by universities and attract quite a significant number of participants. International Data Analysis Olympiad (IDAO)[50] is just a single example among many others[51],[52]. IDAO has engaged several thousand participants across almost a hundred countries each year since 2019. Besides reaching out to a big community, organizers usually run a series of events, including online and offline interactions with the participants.

Zooniverse[53]**:** Zooniverse builds a community of people interested in contributing their efforts and intelligence to scientific research advances. It provides participants with unlabelled datasets from various scientific branches: biology, climate, history, physics, etc. Those datasets require human intelligence to label and understand the scientific assumptions of the domain and phenomena presented. Participation in real-science research can motivate people quite significantly. In some cases, discussions between scientists and Zooniverse participants lead to new scientific discoveries (Clery, 2011).

Other: There are many different venues for interactions between science and citizens. In his book "Reinventing Discovery: The New Era of Networked Science" (Nielsen, 2020), Michael Nielsen gives a good overview. An interesting example of such interaction is the design of a network of micro-prediction agents that follow a specific question-answering protocol. Authors of those agents get rewards for providing correct answers. Such protocol incentivizes the participants to come up with better algorithms and suitable external data sources (Cotton, 2019). A broader list of citizen-science projects is, of course, available at Wikipedia (wik).

7 Independently hosted contests

As we've seen, most organisers choose to host their contests on a platform. However, others have shown that it's still possible to "self-host" contests. Here we give a few brief examples of independently hosted contests.

49. `https://paperswithcode.com/`
50. `https://idao.world`
51. Data Mining Cup, `https://www.data-mining-cup.com/`
52. ASEAN Data Science Explorers `https://www.aseandse.org/`
53. `https://www.zooniverse.org/`

MIT Battlecode[54] is an annual competitive real-time strategy game where players need to write code to manage a robot army. The first iteration took place in 2003, predating all currently active contest platforms. Anyone can participate, but only student teams (from any university) are eligible for prizes. Recent sponsors include game studios and quantitative trading firms. MIT students participating in Battlecode are eligible for credits, as it is a registered course.

Real Robot Challenge (Bauer et al., 2022)[55] is a contest involving dexterous manipulation tasks using robot hands. Evaluation takes place on physical robots. Participants are provided with software simulation environments to train their policies, and are able to submit their policies for physical evaluation. The organising team had to do significant development work in order to be able to accept submissions to run on their physical robots, and they decided to self-host the whole contest since the additional work to build their own leaderboard was deemed easier than integrating with an existing platform.

The ARCathon[56] is an ongoing abstract reasoning benchmark that spun out of the 2020 Kaggle Abstraction and Reasoning Challenge. It maintains the same closely-guarded test set, and their website provides tools for exploring the training set manually as well as crowdsourcing new training examples.

The Humanoid Robot Wrestling Competitions[57] are a series of simulated robotics challenges. The organisers built their own leaderboard management framework on top of GitHub Actions, enabling anyone with a GitHub account to take part in the challenge. Evaluations are run automatically on a dedicated server managed by the organisers whenever a competitor pushes a code change to their GitHub repository. Participants' code can stay hidden from other participants; participants just need to add the challenge organiser's GitHub account as a collaborator on their repository. The organisers helpfully shared their challenge template[58] under a generous open-source licence, enabling others to run challenges like this with minimal additional setup.

The Vesuvius Challenge[59] aims to decipher millenia-old carbonised papyrus scrolls by using computer vision algorithms on high-resolution non-invasive x-ray scans. Alongside the main prizes for reading characters or passages from the scrolls, the organisers offer various prizes for preliminary progress, and contribtions to the community through tool-building or information sharing. The organisers hosted an image segmentation challenge on Kaggle for the subproblem of ink detection, but most of the prizes require submission through a Google form.

8 Choosing the right platform

Given the set of platforms available, choosing the one best suited to a particular competition or challenge is not trivial. We hope that table 1 can be a helpful resource for contest organisers. In addition to this, we can provide some general advice.

54. `https://battlecode.org`
55. `https://real-robot-challenge.com`
56. `https://lab42.global/arcathon/`
57. `https://webots.cloud/competition`
58. `https://github.com/cyberbotics/competition-template`
59. `https://scrollprize.org/`

For companies with limited in-house data science expertise or tech resources, it makes sense to choose a platform which offers support with challenge design and data preparation. While these platforms can require a larger budget than alternative options, they are often able to leverage their existing significant user-base to engage desirable and capable competitors, resulting in more and higher-quality submissions than might otherwise be possible. This reduces the pressure on organisers to promote the contest themselves.

Contest organisers with a limited budget will generally have to take on the challenge design and data preparation work themselves. In these cases, unless an additional contest sponsor can be found, using a platform with free contest hosting options will likely be desirable. In order to aid with discoverability on the free hosting options - helping potential participants find the contest - organisers might want to try to get their competition mentioned in relevant newsletters, or submit their contest to a contest listing site like ML Contests if they are trying to reach a broad audience.

Teams organising contests with particular requirements - reinforcement learning environments, data privacy restrictions, or human-in-the-loop evaluation - are more restricted in their choice of platforms than "vanilla" supervised learning contests. It's worth noting that even if platforms don't officially list certain features, sometimes they are able to accommodate additional requirements - so it can be worth having an exploratory conversation before ruling out any platforms, as long as sufficient budget is available to compensate platforms for any additional development they might need to do.

Contests targeted at niche communities might benefit from the relevant exposure they would get on a domain-specific platform (see section 5). Similarly, contests targeting participants with certain language skills or located in particular geographic areas might take this into account when choosing a platform (see section 4).

Only organisers of the most idiosyncratic contests or those with significant in-house resources would likely find it preferable to run a contest without making use of any platform. We mention some examples of these in section 7.

9 Conclusion

This chapter presents an overview of the most popular AI contest platforms. It gives a summary of each of the platforms, introduces key criteria for platform comparison, and uses these to provide a simple comparison table that we hope will be a useful reference for any contest organiser looking to find the most suitable service for running their contest and maximising its potential impact.

Acknowledgments and Disclosure of Funding

The work presented in this book chapter was undertaken as a community collaboration and did not receive any external funding.

References

List of crowdsourcing projects. https://en.wikipedia.org/wiki/List_of_crowdsourcing_projects.

S. Bauer, M. Wüthrich, F. Widmaier, A. Buchholz, S. Stark, A. Goyal, T. Steinbrenner, J. Akpo, S. Joshi, V. Berenz, V. Agrawal, N. Funk, J. Urain De Jesus, J. Peters, J. Watson, C. Chen, K. Srinivasan, J. Zhang, J. Zhang, M. Walter, R. Madan, T. Yoneda, D. Yarats, A. Allshire, E. Gordon, T. Bhattacharjee, S. Srinivasa, A. Garg, T. Maeda, H. Sikchi, J. Wang, Q. Yao, S. Yang, R. McCarthy, F. Sanchez, Q. Wang, D. Bulens, K. McGuinness, N. O'Connor, R. Stephen, and B. Schölkopf. Real robot challenge: A robotics competition in the cloud. In D. Kiela, M. Ciccone, and B. Caputo, editors, *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pages 190–204. PMLR, 06–14 Dec 2022. URL https://proceedings.mlr.press/v176/bauer22a.html.

H. Carlens. State of competitive machine learning in 2023. *ML Contests Research*, 2024. https://mlcontests.com/state-of-competitive-machine-learning-2023.

D. Clery. Galaxy zoo volunteers share pain and glory of research. *Science*, 2011. ISSN 1095-9203. doi: 10.1126/science.333.6039.173.

P. Cotton. Self organizing supply chains for micro-prediction: Present and future uses of the roar protocol. 2019.

D. Donoho. 50 years of data science. volume 26, pages 745–766. Taylor & Francis, 2017. doi: 10.1080/10618600.2017.1384734. URL https://doi.org/10.1080/10618600.2017.1384734.

DrivenData. Drivendata. https://www.drivendata.org/competitions/, 2014.

A. Goldbloom and B. Hamner. Kaggle. https://kaggle.com/competitions, 2010.

A. Group. Tianchi. https://tianchi.aliyun.com/competition, 2014.

S. Mohanty, S. Khandelwal, and M. Salathe. Aicrowd. https://www.aicrowd.com/challenges, 2017.

M. Nielsen. *Reinventing discovery: the new era of networked science*, volume 70. Princeton University Press, 2020.

A. Pavao, Z. Liu, and I. Guyon. Filtering participants improves generalization in competitions and benchmarks. In *ESANN 2022 - European Symposium on Artificial Neural Networks*, Bruges, Belgium, Oct. 2022. URL https://inria.hal.science/hal-03869648.

A. Pavao, I. Guyon, A.-C. Letournel, D.-T. Tran, X. Baro, H. J. Escalante, S. Escalera, T. Thomas, and Z. Xu. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6, 2023. URL http://jmlr.org/papers/v24/21-1436.html.

Z. Xu, S. Escalera, A. Pavão, M. Richard, W.-W. Tu, Q. Yao, H. Zhao, and I. Guyon. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543, 2022. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2022.100543. URL https://www.sciencedirect.com/science/article/pii/S2666389922001465.

D. Yadav, R. Jain, H. Agrawal, P. Chattopadhyay, T. Singh, A. Jain, S. B. Singh, S. Lee, and D. Batra. Evalai: Towards better evaluation systems for ai agents. 2019.

Zindi. Zindi. `https://zindi.africa/competitions`, 2018.