Academic Competitions

Hugo Jair Escalante HUGOJAIR@INAOEP.MX
Instituto Nacional de Astrofísica,
Óptica y Electrónica,
Tonantzintla, 72840, Puebla, Mexico
The University of Texas at El Paso,
500 W University Ave, El Paso, TX 79968

Aleksandra Kruchinina ALEKSANDRA.KRUCHININA@UNIVERSITE-PARIS-SACLAY.FR
Université Paris Saclay
Paris, France

Reviewed on OpenReview: *https://openreview.net/forum?id=Oc1taS2iYd*

Abstract

Competitions comprise effective means for (i) advancing the state of the art, (ii) putting in the spotlight of a scientific community specific topics and problems, as well as (iii) closing the gap for under represented communities in terms of accessing and participating in the shaping of research fields. Competitions can be traced back for centuries and their achievements have had great influence in our modern world. Recently, they (re)gained popularity, with the overwhelming amounts of data that is being generated in different domains, as well as the need of pushing the barriers of existing methods, and available tools to handle such data. This chapter provides a survey of academic challenges in the context of machine learning and related fields. We review the most influential competitions in the last few years and analyze challenges per area of knowledge. The aims of scientific challenges, their goals, major achievements and expectations for the next few years are reviewed. An associated repository is available here: `https://hugojair.github.io/challenges-survey/`

Keywords: Academic competitions and challenges, Survey of academic challenges, Impact of academic competitions.

1 Introduction

Competitions are nowadays a key component of academic events, as they comprise effective means for making rapid progress in specific topics. By posing a challenge to the academic community, competition organizers contribute to pushing the state of the art in specific subjects and/or to solve problems of practical importance. In fact, challenges are a channel for the reproducibility and validation of experimental results in specific scenarios and tasks.

We can distinguish two types of competitions: those associated to industry or aiming at solving a practical problem, and those that are associated to a research question (academic competitions). While sometimes it is difficult to typecast competitions in these two categories, one can often identify a tendency to either variant. This chapter focuses on academic competitions, although some of the reviewed challenges are often associated to industry too. An academic competition can be defined as a *contest that aims to answer a scientific question via crowd sourcing where participants propose innovative solutions, ideally the challenge will push the state-of-the-art and have a long-lasting impact and/or an*

established benchmark. In this context, academic competitions relying on data have been organized for a while in a number of fields like natural language processing (Harman, 1993), machine learning (Guyon et al., 2004) and knowledge discovery in databases[1], however, their spread and impact has considerably increased during the last decade, see Figure 1 for statistics of the CodaLab platform (Pavao et al., 2023).

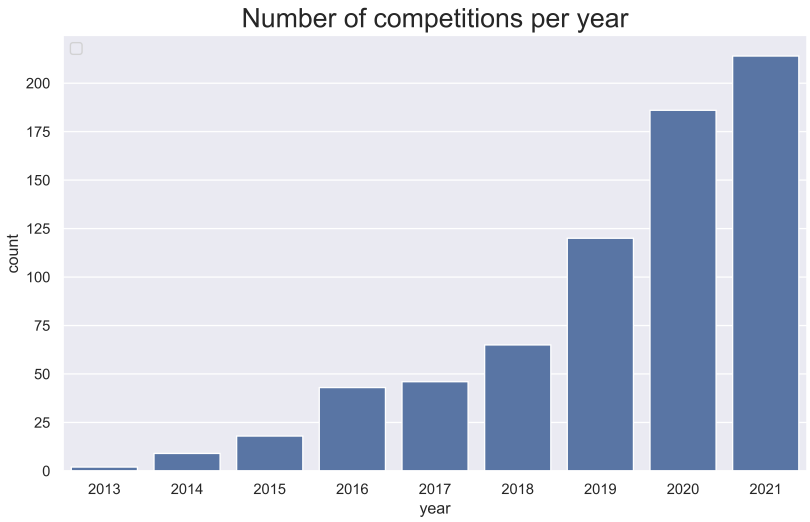

Figure 1: Evolution of the number of competitions each year. Data gathered from *CodaLab Competitions* (Pavao et al., 2023), a platform with a community focused on academic competitions recognized consecutive years as the platform with most organized competitions (Carlens, 2023, 2024, 2025).

As a consequence of this growth, we can witness the permeation and influence that competitions have had in a number of fields. This chapter aims to survey academic competitions and their impact in the last few years. The objective is to provide the reader with a snapshot of the rise and establishment of academic competitions, and to outline open questions that could be addressed with support of contests in the near future. We have focused on machine learning competitions with emphasis on academic challenges. Nevertheless, competitions from other related fields are also briefly reviewed.

The remainder of this chapter is organized as follows. Next section provides a brief historical review of competitions in the context of academia and their impact in different fields. Then, in Section 3, we review academic competitions in terms of the associated field. Section 4 elaborates on ongoing trends in the era of foundational models. Finally in Section 5 we outline some thoughts and ideas on the future of academic competitions.

1. `https://www.kdd.org/kdd-cup/view/kdd-cup-1997`

2 A review of academic challenges: past and present

This section provides a survey on academic challenges in the context of machine learning and related fields.

2.1 Historical review

While it is a daunting task to provide a comprehensive timeline of the evolution of challenges in machine learning and related fields, this section aims at providing a generic overview. Perhaps the first memorable *challenge* is the Longitude Act issued in 1714. It asked participants to develop a method to determine longitude up to a half degree accuracy (i.e., about 69 miles in distance if one is placed in the Meridian). After years of milestones and fierce competition, Thomas Harrison was acknowledged as the winner of this *challenge*. The main incentive, in addition to scientific curiosity, was a monetary prize offered by the British crown that today would be equivalent to millions of pounds.

This form of incentive has guided several other competitions organized by governments[2], for example the DARPA (Defense Advanced Research Projects Agency) grand challenge[3] series that for years organized competitions for building an all-terrain autonomous vehicle. These type of challenges are still being organized nowadays, not only by governments but also by other institutions and even the private sector. Consider for instance the funded challenges organized by the National Institute of Standards and Technology[4] (NIST) and the latest editions of the X-Prize Challenge[5] and the Longitude Prize[6], both targeting critical health problems via challenges in their most recent editions. This same model of making progress via crowd sourcing has been adopted by academy for a while now. The first efforts in this direction arose in the 90s, it was in that decade that the first RoboCup, ICDAR (International Conference on Document Analysis and Recognition), KDD Cup (Knowledge Discovery and Data Mining Tools Competition) and TREC (Text Retrieval Conference) competitions were organized. Such challenges are still being organized on a yearly basis, and they have helped to guide the progress in their respective fields.

RoboCup initially focused on the development of robotic systems able to eventually *play* Soccer at human level (Kitano et al., 1998). With currently more than 25 editions, RoboCup has evolved in the type of tasks addressed in the context of the challenge. For instance, the 2026 edition[7] comprises leagues on rescue robots, service robots, soccer playing robots, industrial robots and even a junior league for kids, where each league has multiple tracks. RoboCup competition model has motivated progress on different sub fields within robotics, from hardware to robot control and multi agent communication among others, see (Visser, 2016) for a survey on the achievements of this first 20 editions of RoboCup. Together with the DARPA challenge, RoboCup has largely guided the progress of autonomous robotic agents that interact in physical environments.

2. https://www.nasa.gov/solve/history-of-challenges
3. https://www.darpa.mil/news-events/2014-03-13
4. https://www.nist.gov/
5. https://www.xprize.org/challenges
6. https://longitudeprize.org/
7. https://2022.robocup.org/

Organized by NIST, TREC is another of the *long-lived* evaluation forums that arose in the early 90s (Harman, 1993). TREC initially focused on text retrieval tasks. Unlike RoboCup, where solutions were tested lively during the event, TREC asked participants to submit *runs* of their retrieval systems in response to a series of queries. By that time this represented a great opportunity for participants to evaluate their solutions in large scale and realistic retrieval scenarios. This evaluation model actually is still popular among text-based evaluation forums (see e.g., SemEval[8]). The TREC forum has evolved and now it focuses on a diversity of tasks around information retrieval (e.g., retrieval of clinical treatments based on patients' cases). Additionally, TREC gave rise to a number of efforts like CLEF (Conference and Labs of the Evaluation Forum), ImageCLEF and TRECVID. They split from TREC to deal with specific sub problems such as: question answering, image and video retrieval, respectively.

In terms of OCR, there were also efforts aiming to boost research in this open problem during the 90s (Garris et al., 1997). The first ICDAR conference took place in 1991, although well documented competitions started in the early 00s (see, e.g., (Lucas et al., 2003)), it seems that competitions associated to digital document analysis were associated to ICDAR since the early 90s, see (Matsui et al., 1993). By that time, NIST released a large dataset of handwritten digits (Grother, 1995) with detailed instructions on preprocessing, evaluation protocols and reference results. While this was not precisely an academic competition, this effort allowed reproducibility in times where the world was starting to benefit from information spread throughout the internet. The impact of this effort has been such that, in addition to motivating breakthroughs in OCR, established the MNIST benchmark as a reference problem for supervised learning (see e.g., Yann Lecun's site[9] on results in a subset of this benchmark). Please note that MNIST is considered a *biased* dataset, but other versions exist, including QMNIST (Yadav and Bottou, 2019), where the authors reconstructed the MNIST test set with 60,000 samples.

Another successful challenge series is the KDD Cup, with its first edition taking place in 1997[10]. KDD Cup has focused on challenges on data mining bridging industry and academy, with a variety of topics being covered with time, from retailing, recommendation and customer analysis to authorship analysis and student performance evaluation[11]. While KDD Cup has been more application-oriented, findings from this competition have resulted in progress in the field without any doubt. KDD Cups are reviewed in the next chapter.

The first decade of the 2000 was critical for the consolidation of challenges as a way to solve tough problems with the help from the community. It was during this time that the popular Netflix prize[12] was organized, granting a 1M dollar prize to the team able to improve the performance of their *in-house* recommendation method. The winning team improved by $\approx 10\%$ the reference model (Koren, 2009). Also, one of the long-lived competition programs in the context of machine learning arose in this decade[13]: the *ECML/PKDD Discovery Challenge series*. Organized since 1999, this forum has released a number of

8. https://semeval.github.io/
9. http://yann.lecun.com/exdb/mnist/
10. https://www.kdd.org/kdd-cup/view/kdd-cup-1997
11. https://kdd.org/kdd-cup
12. https://www.netflixprize.com/
13. https://sorry.vse.cz/~berka/challenge/PAST/

datasets, although it is now an established competition track, in the early years, competitions consisted of releasing data and asking participants to build and evaluate solutions by themselves. The NeurIPS 2003 feature selection challenge took place[14] in this decade too, being this one of the oldest machine learning competitions in which test data was withheld from participants (Guyon et al., 2004).

In that same decade, the first edition of evaluation efforts that are still being run were launched, for instance, the first: CLEF[15] (2000), ImageCLEF[16] forum (2003), TRECVID[17] conference (2003), PASCAL VOC[18] (2005) challenges. All of these efforts and others that evolved over the years (e.g., the model selection[19] and performance prediction[20] challenges (2006) that laid the foundation for AutoML challenges), set the basis for the settlement of academic competitions.

The 2000s not only were fruitful in terms of the number and variety of long lasting challenges that emerged, but also because of the establishment of organizations. It was in 2009 that Kaggle[21] was founded, initially focused on challenges as a service, nowadays Kaggle also offers training-learning, hiring and data-code sharing options. From the academic side, in 2011 ChaLearn[22], the *Challenges in Machine Learning Organization* was founded as well. ChaLearn is a non-profit organization that focuses on the organization and dissemination of academic challenges. ChaLearn provides support to potential organizers of competitions and regularly collaborates with a number of institutions and research groups, likewise, it focuses on research associated to challenge organization in general, this book is a product of such efforts.

From 2010 and on challenges have been established as one of the most effective way of boosting research in a specific problem to get practical solutions rapidly. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) featured from 2010 to 2017 has been among the most successful challenges in computer vision, as it witnessed the rise of CNNs for solving image classification tasks, see next section. Likewise, the VOC challenge organized until 2012, contributed to the development of object detection techniques like YOLO (Redmon et al., 2016). The AutoML challenge series (from 2015) proved that long term contests with code submission could lead to progress on the automation of model design at different levels. As a result, nowadays, top-conferences and venues from different fields have their competition track. Table 1 shows representative competition programs associated to major conferences and related organizations.

This table illustrates that many scientific communities have acknowledged the importance of academic competitions, and highly value these by dedicating resources towards organizing such competitions. Please note that there are top tier venues that do not have an *official* competition track, and therefore they were not included in this table. However,

14. http://clopinet.com/isabelle/Projects/NIPS2003/
15. https://www.clef-initiative.eu/web/clef-initiative/
16. https://www.imageclef.org/
17. https://trecvid.nist.gov/
18. http://host.robots.ox.ac.uk/pascal/VOC/
19. http://clopinet.com/isabelle/Projects/NIPS2006/home.html
20. http://www.modelselect.inf.ethz.ch/
21. https://kaggle.com/
22. http://chalearn.org/

Table 1: Competition tracks of main conferences in machine learning and related fields. Column four shows the number of tasks organized in the latest edition of the associated track (# Tasks LE) as of 2025 or 2026. Acronyms are as follows: Machine Learning (ML), Data Mining (DM), Computational Intelligence (CI), Pattern Recognition (PR), Robotics (RO), MIR (Multimedia Information Retrieval), Multimedia Information Processing (MIP), Information Retrieval (IR), Natural Language Processing (NLP), Artificial Intelligence (AI), Evolutionary Computation (EC), Medical Image Analysis (MI), Signal Processing (SP), Image Processing (IP), Miscellaneous (MS). The last four rows of this table shows institutions and organizations associated with challenges.

Venue	Field	Since	# Tasks LE	URL
TREC	IR	1993	7	https://trec.nist.gov/
ICDAR	PR	1993	8	https://icdar2026.org/
KDD	DM	1997	2	https://kdd.org/
ECML	ML	1999	3	https://ecmlpkdd.org/
RoboCup	RO	1997	5	https://www.robocup.org/
PAN-CLEF†	NLP	2000	5	https://pan.webis.de/
TrecVid	MIR	2003	2	https://trecvid.nist.gov/
ImageCLEF†	MIP	2003	5	https://www.imageclef.org
MediaEval	MIP	2003	6	https://multimediaeval.github.io/
GECCO	EC	2004	12	https://gecco-2026.sigevo.org/HomePage
WCCI	CI	2006	4	https://attend.ieee.org/wcci-2026/
MICCAI	MI	2007	38	https://conferences.miccai.org/2022/en/
Interspeech	SP	2008	7	https://interspeech2026.org/en-AU
ICRA	RO	2008	9	https://2026.ieee-icra.org/
ACM Multimedia	MIP	2009	25	https://2026.acmmm.org/
ICPR	PR	2010	5	https://icpr2026.org/
SemEval	NLP	2010	13	https://semeval.github.io/
IROS	RO	2012	8	https://2026.ieee-iros.org/
ICMI	MIP	2013	3	https://icmi.acm.org/2026/
ICASSP	SP	2014	14	https://2026.ieeeicassp.org/
ICME	MIP	2015	8	https://2026.ieeeicme.org/
CIKM	DM	2017	1	https://cikm2026.diag.uniroma1.it/
ICIP	IP	2017	4	https://2026.ieeeicip.org/
NeurIPS	ML	2018	18	https://neurips.cc/Conferences
IJCAI	AI	2018	14	https://2026.ijcai.org/
AutoML	ML	2022	1	https://automl.cc/
Loingitude Prize	MS	1714*	1	https://longitudeprize.org/
XPrize*	MS	1996	2	https://www.xprize.org/
Kaggle	MS	2009	-	https://www.kaggle.com/
ChaLearn	ML	2011	-	http://chalearn.org/

these venues have hosted workshops associated to competitions that have had great impact. Just to name a few: CVPR, ICCV, ECCV, ICML, ICLR, EMNLP, ACL.

2.2 Progress driven by academic challenges

As previously mentioned, challenges are now established mechanisms for dealing with complex problems in science and industry. This is not fortuitous, but a response from the community to a number of accomplishments in different fields. This section aims to briefly summarize the main achievements of selected challenges that have motivated other researchers and fields to organize competitions. We focused on a representative machine learning challenge (AutoML) and two evaluation campaigns from the two fields where more contests are organized, see Figure 2.

- **AutoML challenges.** AutoML is the sub field of machine learning that aims at automating as much as possible all of the aspects of the design cycle (Hutter et al., 2018). While people were initially skeptical of the potential of this sort of methods, nowadays AutoML is a trending research topic within machine learning (there is a dedicated AutoML conference with a competition track[23] since 2022). This is in large part due to the achievements obtained in the context of AutoML challenges. Back in 2006 early efforts in this direction were the prediction performance challenge (Guyon et al., 2006) and the agnostic *vs.* prior knowledge challenge (Guyon et al., 2008). These contests asked participants to build methods for automatically or manually building classification models. They became the predecessors of the AutoML challenge series that ran from 2015 to 2018 (Guyon et al., 2019), and all of the follow up events that are still organized. Initially, the AutoML challenge series focused on tabular data, but it then evolved to deal with raw heterogeneous data in the AutoDL[24] challenge series(Liu et al., 2021b), whose latest edition is the Cross-Domain MetaDL challenge 2022 [25] (El Baz et al., 2021a,b; Carrión-Ojeda et al., 2022). A number of methods (e.g., AutoSKLearn (Feurer et al., 2019)), evaluation protocols, AutoML mechanisms (e.g., Fast Augmentation Learning methods (Baek et al., 2020)) and improvements arose in the context of these challenges including the evaluation of submitted code, cheating prevention mechanisms, the progressive automation of different types of tasks (e.g., from binary classification to regression, to multiclass classification, to neural architecture search) and the use of different data sources (from tabular data, to raw images, to raw heterogeneous datasets). The result is an established benchmark that is widely used by the community.

- **ImageNet Large Scale Visual Recognition Challenge.** The so called, ImageNet challenge asked participants to develop image classification systems for 1,000 categories and using millions of images as training data (Russakovsky et al., 2015). At the time of the first edition of the challenge, object recognition, image retrieval and classification datasets were dealing with problems involving thousands of images and dozens of categories (see e.g., (Escalante et al., 2010)). While the scale made participants struggle in the first two editions of the challenge, the third round witnessed the renaissance of convolutional neural networks, when AlexNet reduced drastically the error rate for this dataset (Krizhevsky et al., 2012). In the following editions of the challenge other landmark CNN-based architectures for image classification were proposed including: VGG (Simonyan and Zisserman, 2015), GoogLeNet (Szegedy et al., 2015) and ResNet (He et al., 2015). These architectures comprised important contributions to deep learning, including residual connections/blocks and inception-based networks, the establishment of regularization mechanisms like dropout, pretraining and fine tuning and the efficient usage of GPUs for training large models. While the challenge itself did not provoke the aforementioned contributions, it was the catalyst and solid test bed for the rise of deep learning in computer vision.

23. `https://automl.cc/`
24. `https://autodl.chalearn.org/`
25. `https://metalearning.chalearn.org/`

- **Text Retrieval Evaluation Conference.** TREC initially focused on the evaluation of information retrieval systems (text) (see (Voorhees and Harman, 2005; Rowe et al., 2010) for an overview of the early editions of TREC), but it rapidly evolved to include novel tasks and evaluation scenarios in the forthcoming years. This led to include? tasks that involved information sources from multiple languages, and eventually images and videos. Other tasks that have been widely considered in the TREC campaign are: question answering, adaptive filtering, text summarization, indexing, among many others. Thanks to this effort the information retrieval and text mining fields were consolidated and boosted the progress in the development of search engines and related tools that are quite common nowadays. Well known retrieval models and related mechanisms for efficient indexing, query expansion, relevance feedback, arose in the context of TREC or were validated in this forum. Another important contribution of TREC through the years is that it has evolved to give rise to numerous tasks and application scenarios that have defined the text mining field.

We surveyed a few representative challenges and outlined the main benefits that they bring into their respective communities. While these are very specific examples and while we have chosen breaking through competitions, similar outcomes can be drawn from challenges organized in other fields. In Section 3 we review challenges from a wider variety of domains.

2.3 Pros and cons of academic challenges

We have learned so far that challenges are beneficial in a number of ways, and have boosted progress in a variety of domains. However, it is true that there are some limitations and undesired effects of challenges that deserve to be pointed out. This section briefly elaborates on benefits and limitations of academic challenges.

2.3.1 Benefits of academic challenges

The main benefit of challenges is the solution of complex problems via crowd sourcing, advancing the state of the art and the establishment of benchmarks. There are, however, other benefits that make them appealing to both participants and organizers, these include:

- **Training and learning through challenges.** Competitions are an effective way to learn new skills, they *challenge* participants to gain new knowledge and put in practice known concepts for solving relevant problems in research and industry. Even if participants do not win a challenge or a series of them, they progressively improve their problem solving skills.

- **Challenges are open to anyone.** Apart of political restrictions that may be applied for some organizations, competitions target anyone with the ability to approach the posted problem. This is particularly appealing to underrepresented groups and people with limitations to access the cutting edge problems, data and resources. For instance, most competitions adopting code submission provide cloud-based computing to participants. Likewise, challenges can be turned into ever lasting benchmarks and they contribute to making data available to the public.

- **Engagement and motivation.** The engagement offered by competitions is priceless. Whether the reward is economic, academic (e.g., publication or talk in a workshop, professional recognition in the field), competitiveness, or just fun, participants find challenges motivating.

- **Reproducibility.** This cannot be emphasized enough, benchmarks associated to challenges not only provide the task, data and evaluation protocols. In most cases resources, starting-kits, others' participants code and computing resources are given as well. This represents an easy way to get into competitions to participants, which can directly compete with state-of-the-art solutions. At the same time, competitions having these features guarantee reproducibility of results which is clearly beneficial to the progress in the field.

2.3.2 Pitfalls of academic challenges

Despite the benefits of challenges, they are not risk-free, therefore, there are certain limitations that should be taken into account.

- **Performance improvement vs. scientific contribution.** Academic challenges often ask participants to build solutions that achieve the best performance according to a given metric. Although in most cases there is a research question associated to a challenge, participants may end up building solutions that optimize the metric but that do not necessarily result in new knowledge. This gives challenges a bitter-sweet taste, as often new findings are overshadow by super-tuned off-the-shell solutions.

- **Stagnation.** An undesirable outcome for a challenge is stagnation, this is often the result of wrong challenge design decisions, that result in either a problem that is too hard to be solved with current technology or unattractive to participants. While it is not possible to anticipate how far the community can go in solving a task, the implementation of (strong) baselines, starting kits and appealing datasets, or rewards could help to avoid stagnation.

- **Data Leakage.** It refers to the use of target (or any other relevant information that is supposed to be withheld from participants) information by participants to build their solutions (Kaufman et al., 2011). This is a common issue when datasets are re-used or when datasets are build from external information (e.g., from social networks). Anonymization and other mechanisms as those exposed in (Kaufman et al., 2011) could be adopted for avoiding this problem.

- **Privacy and rights on data.** *"Data is the new oil"* has been a popular say recently[26], while this is debatable, it is true that data is a valuable asset that must be *handled with care*. Therefore copyright infringement should be avoided to the uttermost end. Likewise, failing to guarantee privacy is an important issue that must be addressed by organizers as this could lead into legal issues. Anonymization mechanism should be applied to data before its release, making sure it is not possible to track users identity or other important and confidential information.

26. https://www.forbes.com/sites/forbestechcouncil/2019/11/15/data-is-the-new-oil-and-thats-a-good-thing/?sh=381ec30d7304

Additional Downsides from the Perspective of Participants We now list additional drawbacks from the participants' perspective that should be taken into account when designing and running an academic competition.

- **The cost-benefit asymmetry of participation.** Participating in a competition imposes substantial costs on participants. However, only few participants are *rewarded* or *recognized*.
- **The compute gap.** Some competitions do not provide cloud-computing resources for participants. In this scenarios there is a clear asymmetry between well-resourced and under resourced participants.
- **Benchmark contamination and data leakage in the LLM era.** The widespread practice of relying on pretrained models introduces the risk that some test-set documents or samples were considered as pretraining data.
- **Gaming the leaderboard.** Overfitting of the learboard can occur on competitions implementing public leaderboards and unlimited submissions.
- **Psychological and well-being concerns.** Competitive environments can generate significant psychological pressure, particularly for participants who have invested heavily in a challenge.
- **Intellectual property and credit attribution.** When participants submit code or models as part of a competition, the intellectual property (IP) arrangement is not always clearly defined.
- **The saturation problem from a participant perspective.** Once a benchmark is saturated (i.e., top-performing systems approach or exceed human-level performance) the marginal scientific value of further improvement is low.

These shortcomings suggest that a more participant-centered approach to competition design (attending to compute equity, reproducibility, psychological well-being, and clear IP frameworks) would help to ensure that the benefits of academic challenges are distributed more equitably and that the participation experience reinforces, rather than undermines, participants' long-term engagement with research.

2.4 What makes academic challenges successful?

Having reviewed competitions, their benefits and pitfalls/limitations, this section elaborates on characteristics that we think make a challenge successful. While it is subjective to define a successful challenge, the following guidelines associate success to high participation, quantitative performance and novelty of top ranked solutions.

- **Scientific rigor.** The design and the analysis of the outcomes of a competition are critical for its success. Following scientific rigor as "to ensure robust and unbiased experimental design, methodology, analysis, interpretation and reporting of results" (Hofseth, 2017) is necessary and helps to avoid some of the limitations mentioned above. Adopting statistical testing for the analysis of results, careful designing

of evaluation metrics, establishing theoretical bounds on these, running multiple tests before releasing the data/competition, formalizing the problem formulation, performing ablation studies are all critical actions that impact on the outcomes of academic challenges.

- **Rewarding and praising scientific merit and novelty of solutions.** It is worth mentioning that novel methods do not always make it to the top of the leaderboard, but these new ideas may be great seeds and serve as an inspiration to others for further fruitful research. Therefore, rewarding and acknowledging scientific merit and novelty of solutions is very important. There are several ways of doing this, for instance, having a *prize* for the most original/novel submission or granting a best paper award that is not entirely based on quantitative performance.

- **Publication and dissemination of results** are good practices with multiple benefits. Participants are often invited to fill out *fact sheets* and write workshop papers in order to document their solutions. Similarly, organizers commonly publish overview papers that summarize the competition, highlighting the main findings and analyzing results in detail. Associating a special issue of a journal with competitions is a good idea as it is motivating for participants, and at the same time it is a *product* that organizers can report in their work evaluations.

- **Associating the competition with an top tier venue** (e.g., conferences, summits, workshops, etc.) makes a challenge more attractive to participants, as they associate the quality of associated venues and competitions. Also, physically attending the competition session is more appealing if participants can also attend top tier events.

- **Organization of panels and informal discussion sessions** involving both participants and organizers is valuable for sensing perception of people associated to the event. This is critical when organizing challenges that run for several editions.

- **Establishing benchmarks** should be an underlying goal of every competition. Therefore, curated data, fail safe evaluation protocols, and adequate platforms for maintaining competitions as long term evaluation test beds are essential. Likewise, the use of open data and open source code for the purposes of reproducibility and so that everyone can benefit and continue their own research.

2.4.1 Academic vs. industrial challenges

Industrial challenges are described in detail in the next chapter. In this section we outline the main differences of industry and scientific competitions.

The main objective of industrial challenges is the economic advantage from the winning model that will potentially increase profits and improve business model, meaning it should be an end-to-end solution. The organizers care much less about scientific publications, being scientifically rigorous neither about the results being statistically significant. These types of contest do not single out scientific questions, that is not the priority for them. They aim at specific business problems, usually posses big not preprocessed datasets and evidently can provide more often big prizes. Up till now the direct positive correlation

between these big rewards and qualitative contributions has not been proven. But it was observed that big prizes might attract many participants, create "big splash" in the news for the company-organizer and cause a lot of noise in the leaderboard, potentially leading to gaining by chance. While the winners and contributors of academic challenges get scientific recognition, the top performers at the industrial contests can receive job offers and be hired by the organizers.

Another important aspect of industrial challenges is that due to their nature and the concurrent market, the company-organizers prefer to keep the data and the submitted code private, which is in the opposition with the scientific mentality, because it prevents to benefit from the latest break-through and get inspiration from the newest ideas.

3 Academic challenges across different fields

This section briefly reviews challenges across different fields. We focus on fields that have long tradition in challenges. In order to identify such fields of knowledge, we surveyed competitions organized in the CodaLab platform (Pavao et al., 2023). Figure 2 shows a distribution of CodaLab challenges across fields of knowledge. Clearly NLP and Computer vision challenges dominate, this could be due to the explosion of availability of visual and textual data of the last few years. One should note that most of the competitions shown in that plot have a strong machine learning component. In the remainder of this section we briefly survey competitions organized in a subset of selected fields.

3.1 Challenges in Machine Learning

Machine learning is a transversal field of knowledge that has been present in most challenges regardless of the application field (e.g., computer vision, OCR, NLP, time series analysis, and so on). Therefore, it is not easy to cast a challenge as a ML competition. For that reason, in this section we review as a representative sample the competition track of the NeurIPS conference. The track has run regularly since 2017, although challenges organized with the conference date back to the early 2000s (Guyon et al., 2004). Overview papers for the NeurIPS competition track from 2019 to 2023 can be found in (Escalante and Hadsell, 2019; Escalante and Hofmann, 2020; Kiela et al., 2022; Ciccone et al., 2023).

Figure 3 shows the number of competitions that have been part of the NeurIPS competition track. There has been an increasing number of competitions organized each year, see also (Carlens, 2023) for more details. The topics of challenges are quite diverse, with deep reinforcement learning (DRL) prevailing since the very beginning of the track. The first competition in the program around this topic was the Learning to Run challenge[27] that asked participants to build an human-like agent to navigate an environment with obstacles (Kidzinski et al., 2018), this challenge was run for two more editions, the last one being the Learn to Move - Walk Around[28] challenge. DRL-based competitions addressing other challenging navigation scenarios are the Animal Olympics[29] and MineRL series, see below. DRL challenges addressing different tasks are the Real robot challenge[30] series with

27. https://www.aicrowd.com/challenges/nips-2017-learning-to-run
28. https://www.aicrowd.com/challenges/neurips-2019-learn-to-move-walk-around
29. http://animalaiolympics.com/AAI/
30. https://real-robot-challenge.com/

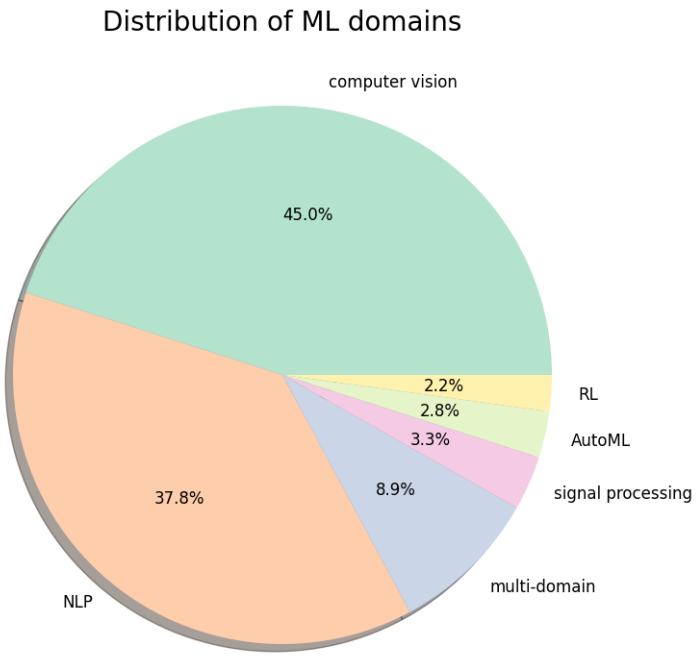

Figure 2: Distribution of competitions with different machine learning domains. Data gathered from CodaLab Competitions (Pavao et al., 2023)

two editions, the Learning to run a power network competition[31] and the two editions of the Pommerman[32] competition where the goal was to develop agents to compete to each other in a bomberman-game-like scenario. The presence of DRL in the challenge track as been growing in the last editions.

Another popular topic in the NeurIPS competition track is AutoML: since 2018, at least one competition associated to this topic has been part of the NeurIPS competition track. These include the AutoML@NeurIPS (Escalante et al., 2019) and AutoDL (Liu et al., 2021b) challenges, the black-box optimization competition (Turner et al., 2021), the predicting generalization in deep learning challenge [33], two editions of the Meta-DL challenge (El Baz et al., 2021b; Carrión-Ojeda et al., 2022) and the AutoML Decathlon[34].

Specific challenges that have been part of the competition track for more than 2 editions are the following:

- **Traffic4cast**[35]. Organizing variants of challenges aiming to predict traffic conditions under different settings and scenarios, see (Kreil et al., 2020; Kopp et al., 2021; Eichenberger et al., 2022).

31. https://l2rpn.chalearn.org/
32. https://www.pommerman.com/
33. https://sites.google.com/view/pgdl2020
34. https://www.cs.cmu.edu/~automl-decathlon-22/
35. https://www.iarai.ac.at/traffic4cast/

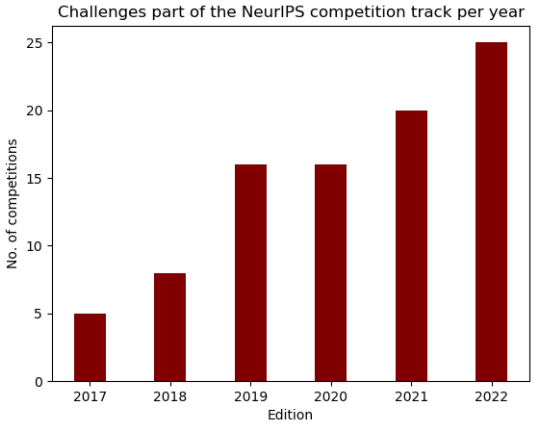

Figure 3: Number of challenges organized as part of the NeurIPS competition program.

- **The AI Driving Olympics (AI-DO).** Aiming to build autonomous driving systems running in simulation and small physical vehicles tested live during the competition track[36].
- **MineRL**[37] A competition series focusing on building autonomous agents that using minimal resources are able to solve very complex tasks in a MineCraft environment. In the first two editions agents were asked to find a diamond with limited resources, see (Milani et al., 2020; Guss et al., 2021). In the most recent editions tasks have been varied and more specific (Shah et al., 2022).
- **Reconnaissance Blind Chess.** Challenges participants to build agents able to play a chess variant in which a player cannot see her opponent's pieces but can learn about them through private, explicit sensing actions. Three editions of this competition have run in the track (Gardner et al., 2020).

It is difficult to summarize the number and variety of topics addressed in challenges part of the NeurIPS competition, however, we have reviewed a representative sample. Nevertheless, please note that most challenges reviewed in the remainder of this section also include an ML component.

3.2 Challenges in Computer Vision

Together with machine learning, computer vision has been greatly benefited from challenges. As previously mentioned, The PASCAL Object detection challenge series boosted research on object detection and semantic segmentation (Everingham et al., 2015). The ImageNet large scale classification challenge is another landmark competition that served as platform for the renaissance of convolutional neural networks (Russakovsky et al., 2015). In addition to these landmark competitions there have been a number of efforts that have pushed further the state-of-the-art, these are reviewed in the following lines.

36. `https://www.duckietown.org/research/AI-Driving-olympics`
37. `https://minerl.io/`

The ChaLearn Looking at People (ChaLearn LAP[38]) series has organized academic challenges around the analysis of human behavior from visual information. More than 20 competitions on the topic have been organized so far, see (Escalera et al., 2017a) for a (outdated) review. Among the organized competitions several of the datasets have become a reference for different tasks, and are used as benchmarks. These include: the gesture recognition challenges (Escalera et al., 2013, 2014, 2017b; Wan et al., 2017), the personality recognition challenge series (Escalante et al., 2017, 2022; Palmero et al., 2021), the age estimation challenge series (Escalera et al., 2015, 2016) and the face anti-spoofing challenge series (Liu et al., 2019; Wan et al., 2020; Liu et al., 2021a). A wide diversity of related topics have been studied in the context of ChaLearn LAP challenges, including: action recognition and cultural event recognition (Baró et al., 2015; Escalera et al., 2015), sign language understanding (Sincan et al., 2021), identity preserving human analysis (Clapés et al., 2020) among others. Undoubtedly, these challenges have advanced the state of the art in a number of directions within computer vision and affective computing.

The Common Objects in COntext (COCO[39]) challenge series that emerged after the end of the Pascal VOC challenge. This effort continued benchmarking object detection methods, but also started evaluating the so called *image captioning* task. Early efforts for the evaluation of this task emerged in the ImageCLEF forum (Clough et al., 2010; Escalante et al., 2010), where the goal was associating keywords to images. The COCO challenge was more ambitious by asking participants to describe the content of an image with a more *human-like* description. Running from 2015-2020 this benchmark was critical for the consolidation of the image captioning task, with major contributions being reported at the beginning of the series, see (Bai and An, 2018; Stefanini et al., 2021). Today, COCO is an established benchmark in a number of tasks related to vision and language, see (Lin et al., 2014).

Other efforts in the field of computer vision are the NTIRE challenge, focused on image restoration, super resolution and enhancement (Timofte et al., 2017; Gu et al., 2022) , the visual question answering competition[40] running from 2016 to 2021, the fine grained classification workshop [41] that has run a competition program since 2017, the EmotioNet[42] recognition challenge that ran in 2020 and is now a testbed for emotion recognition, the ActivityNet challenge [43] organized since 2016 and targeting action recognition in video, among several others.

3.3 Challenges in Natural Language Processing

The development of the natural language processing (NLP) field, in particular for text mining and related tasks, has been largely driven by competitions, also known in the NLP jargon as *shared tasks*. In fact, one of the oldest evaluation forums across all computer science is one focusing in NLP, that is TREC. It initially focused on the evaluation of information retrieval systems (text), but it rapidly evolved to include novel tasks and evaluation scenar-

38. https://chalearnlap.cvc.uab.cat/
39. https://cocodataset.org
40. https://visualqa.org/
41. https://sites.google.com/view/fgvc9
42. https://cbcsl.ece.ohio-state.edu/enc-2020/index.html
43. http://activity-net.org/challenges/2022/

ios in the forthcoming years (Voorhees and Harman, 1998, 2005; Rowe et al., 2010). This lead to consider tasks that involved information sources from multiple languages (Harman, 1998), and eventually, speech signals (Garofolo et al., 2000) and visual information (Awad et al., 2021). Other tasks that have been considered in the TREC campaign are: question answering (Voorhees, 2001), adaptive filtering (Harman, 1995), text summarization [44], among many others. Thanks to this effort the information retrieval and text mining fields were consolidated and boosted the progress in the development of search engines and related tools that are quite common nowadays.

Several well known evaluation campaigns evolved from TREC and consolidated on their own. Most notably, the TRECVid (Awad et al., 2021) and Cross-Language Evaluation Forum (Braschler, 2001) (CLEF) campaigns. The former focusing on tasks related to video retrieval, indexing and analysis. The academic and economic impact of TRECVid has been summarized already. Showing the relevance that such forum has had into the progress of video search technology. CLEF is another forum that initially focused on cross-lingual text analysis tasks. Now it is a conference that comprises several shared tasks, called labs. This include ImageCLEF, PAN among others. Likewise, there are forums dedicated to specific languages, for example, Evalita[45] (for Italian), IberLEF[46] (for Spanish) and GermEval[47].

In terms of speech, there were several efforts from DARPA (Marcus, 1992; Black and Eskenazi, 2009) and NIST[48] in organizing competitions as early as the late 80s. These long term efforts have helped to shape ASR and related fields. More recently, after the deep learning empowering, several challenges focusing on speech have been proposed, these are often associated to major conferences in the field (e.g. Interspeech and ICASSP), see Table 1. There is no doubt that competitions have played a key role for the shaping the wide field of NLP.

3.4 Challenges in Biology

Biology is a field of knowledge that has benefited from competitions considerably. In terms of medical imaging, the premier forum is the grand challenge series associated to the MICCAI conference, running since 2007 [49]. A number of important challenges have been organized in this context, where most competitions deal with medical imagery segmentation or reconstruction of different organs, body parts and input type, see e.g., (Scully et al., 2008; Marak et al., 2009; Andrearczyk et al., 2022). In recent editions the challenge scenarios and approached tasks have been increasing difficulty and the potential impact of solutions. In its last edition, the MICCAI grand challenge series has 38 competitions running in parallel. This is an indicator of success among the medical imaging community.

Other challenges associated to medical image analysis have been presented in forums associated to image processing and computer vision as well. For instance, in 2019 during The IEEE International Symposium on Biomedical Imaging (ISBI), nine challenges were

44. http://trecrts.github.io/
45. https://www.evalita.it/campaigns/evalita-2022/
46. https://sites.google.com/view/iberlef2022
47. https://germeval.github.io/
48. https://www.nist.gov/itl/iad/mig/past-hlt-evaluation-projects
49. https://cause07.grand-challenge.org/Results/

organized[50]. In 2020 a challenge on Image processing on real-time distortion classification in laparoscopic videos was organized with ICIP 2020[51]. In the context of ICCV, challenges on remote measurement of physiological signals from videos (RePSS) were organized: one on measurement of inter-beat-intervals (IBI) from facial videos, and another one on respiration measurement from facial videos (Li et al., 2020, 2021).

It is worth mentioning that there are platforms associated with challenges in biology and medical sciences. The Grand Challenge[52] platform being perhaps the oldest one and the most representative in terms of imagery: *more than 150 competitions are listed in the platform*, most of which are associated to medical image analysis. A related effort is that of the DREAM challenges[53] a platform that has organized more than 60 challenges in biology and medicine. The variety of topics covered by DREAM challenges is vast (Stolovitzky et al., 2009): from systems biology modelling (Meyer and Saez-Rodriguez, 2021), to prevention (Tarca et al., 2020) and monitoring (Sun et al., 2022) damage caused by certain conditions, to disease susceptibility[54], to analyzing medical documents with NLP[55], to drug analysis and combination[56] and many other relevant topics. As seen in Chapter 5, platforms play a key role in challenge success, biology is a field where excellent platforms are available and this has been critical for the advancement of state of the art in this relevant field.

Protein structure modelling was officially introduced in 1994 at the biennial large-scale experiment Critical Assessment of protein Structure Prediction (CASP), and ever since it attracted more than 100 teams to tackle the problem, see (J et al., 2014). Only almost 20 years later, two teams presented breaking through solutions to protein folding task (Kryshtafovych et al., 2021): DeepMind with their AlphaFold2 (Jumper et al., 2021) and scientists of the University of Washington with RoseTTAFold (Baek et al., 2021). Alphafold uses multiple neural networks that feed into each other in two stages. It starts with a network that reads and folds the amino acid sequence and adjusts how far apart pairs of amino acids are in the overall structure. Then goes the structure model network that reads the produced data, creates a 3D structure, and makes the needed adjustments (Evans et al., 2018; Jumper et al., 2021). RoseTTAFold adds a simultaneous third neural network, which tracks where the amino acids are in 3D space as the structure folds, alongside the 1D and 2D information (Baek et al., 2021). The solution of Washington University is less accurate but uses less computational and time resources than AlphaFold2. Without the existence of the CASP experiment, achieving the outstanding performance of these methods would have taken much more time.

As we can see, advancements of machine learning in biology are of crucial importance, that's why there are numerous competitions in this domain. Researchers and practitioners are trying to deal with biological and related domain (medicine, agriculture, and others) challenges using various machine learning solutions like computer vision, NLP and signal processing.

50. https://biomedicalimaging.org/2019/challenges/
51. https://2020.ieeeicip.org/challenge/real-time-distortion-classification-in-laparoscopic-videos/
52. https://grand-challenge.org/challenges/
53. https://dreamchallenges.org/closed-challenges/
54. https://dreamchallenges.org/respiratory-viral-dream-challenge/
55. https://dreamchallenges.org/electronic-medical-record-nlp-dream-challenge/
56. https://dreamchallenges.org/astrazeneca-sanger-drug-combination-prediction-dream-challenge/

3.5 Challenges in Autonomous Driving

DARPA Grand Challenge is considered as one of the first long distance race for autonomous driving cars, it was organised in 2004 with more than 100 teams. None of the robot vehicles managed to finish the 240 km route, only one member covered 11.78 km and then got stuck. Next year there were 195 teams, the distance of the challenge was of 212 km, and five vehicles successfully completed the course. These first courses were challenging but vehicles "operated in isolation", their interaction was not required, and there was no traffic either. So the next Urban challenge was held in 2007 in a city area, the objective was to complete 96km in 6 hours and it included "driving on roads, handling intersections and maneuvering in zones" (Urmson et al., 2007). Six teams managed to complete the course.

The basics were laid, and DARPA pursued their competitions: Robotics Challenge in 2012, 2013 - Fast Adaptable Next-Generation Ground Vehicle Challenge, 2013 – 2017 Subterranean Challenge on "autonomous systems to map, navigate, and search underground tunnel, urban, and cave spaces" [57].

Being able to test autonomous driving cars "in the wild" is important and expensive. In order to fine-grain the algorithms at a less cost one needs to test them virtually. Hopefully, there are different simulators: CARLA [58], VISTA 2.0 [59], NVIDIA DRIVE Sim [60] and others.

Several challenges have been organised based on CARLA simulator, "an open-source simulator for autonomous driving research", which is used to study "a classic modular pipeline, a deep network trained end-to-end via imitation learning, and a deep network trained via reinforcement learning" (Dosovitskiy et al., 2017).

Autonomous driving has numerous interesting challenges, and object detection is one of them. Most of the current research concentrates around camera images, but it is not the best sensor under certain conditions like bad weather, poor lighting. Radar information can help to overcome these inconveniences. It is more reliable, cost-efficient and might potentially lead to better object detection. ROD2021 Challenge is the first competition of its' kind, which proposes object detection task on radar data, and was held in the ACM International Conference on Multimedia Retrieval (ICMR) 2021. Organisers developed their own baseline: "radar object detection pipeline, which consists of two parts: a teacher and a student. Teacher's pipeline fuses the results from both RGB and RF images to obtain the object classes and locations in RF images. Student's pipeline utilizes only RF images as the input to predict the corresponding ConfMaps under the teacher's supervision. The LNMS as post-processing is followed to calculate the final radar object detection results." (Wang et al., 2021b).

This challenge attracted more than 260 participants among 37 teams with around 700 submissions. The winning team, affiliated to Baidu, submitted paper "DANet: Dimension Apart Network for Radar Object Detection" (Ju et al., 2021), where they presented their results. "This paper proposes a dimension apart network (DANet), including a lightweight dimension apart module (DAM) for temporal-spatial feature extraction. The proposed DAM extracts features from each dimension separately and then concatenates the features

57. https://www.darpa.mil/about-us/subterranean-challenge-final-event
58. https://carla.org/
59. https://vista.csail.mit.edu
60. https://www.nvidia.com/en-us/self-driving-cars/simulation/

together. This module has much smaller number of parameters, compared with RODNet-HGwI, so that significant reduction of the computational cost can be achieved. Besides, a vast amount of data augmentations are used for the network training, e.g., mirror, resize, random combination, Gaussian noise and reverse temporal sequence. Finally, an ensemble technique is implemented with a scene classification for a more robust model. The DANet achieves the first place in the ROD2021 Challenge. This method has relatively high performance but with less computational cost, which is an impressive network model. Besides, this method shows data augmentation and ensemble techniques can greatly boost the performance of the radar object detection results" (Wang et al., 2021a).

Another interesting and pioneering challenge is OmniCV (Omnidirectional Computer Vision) in conjunction with IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2021). The objective was to evaluate semantic segmentation techniques targeted for fisheye camera perception. It attracted 71 teams and a total of 395 submissions. Organisers proposed their baseline "a PSPNet network with a ResNet50 backbone finetuned on WoodScape Dataset", which "achieved a score of 0.56 (mIoU 0.50, accuracy 0.67) excluding void class". The top teams managed to get significantly better scores and proposed interesting solutions. The winning team implemented full Swin-transformer Encoder-Decoder approach, with a score of 0.84 (mIoU 0.86, accuracy 0.89) (Ramachandran et al., 2021).

4 Academic Competitions in the Era of Foundation Models

The preceding sections have focused on academic competitions from their early years through their consolidation as a mainstream mechanism for advancing machine learning and related fields. The landscape is, however, far from static. Three developments in particular are reshaping the nature, scope, and methodology of academic challenges in ways that warrant dedicated attention. This section elaborates on these critical aspects that represent the frontier of academic competitions.

4.1 Challenges in the Large Language Models era

The rise and establishment of large language models (LLMs) has generated an entirely new family of academic challenges, while simultaneously reshaping the evaluation landscape of existing ones. Prior to the availability of effective instruction-following models, NLP competitions largely evaluated task-specific fine-tuned systems on well-defined benchmarks (see Section 3.3). With the emergence of models such as GPT, Claude, Gemini, LLaMA, among many others, the community has had to rethink what evaluation means, how leaderboards can be kept meaningful, and what kinds of problems are worth posing to the crowd. In the remainder of this section we elaborate on relevant aspects for competitions in the context of LLMs.

4.1.1 Benchmark-style LLM evaluations

Some of the most influential LLM "competitions" are not hosted on a challenge platform in the traditional sense, but rather take the form of *open leaderboards* where any team can submit model outputs for evaluation against a held-out test set. The General Language Un-

derstanding Evaluation (GLUE) benchmark and its successor SuperGLUE were pioneers on this format (Wang et al., 2018, 2019), quickly becoming the reference evaluation suite for natural language understanding. Both benchmarks were saturated, with models surpassing human-level performance within a few years of their introduction. Illustrating the capacity of competitions to drive rapid progress but also the challenge of keeping evaluation meaningful once the community converges on a solution. Big-Bench (Srivastava et al., 2022) and HELM (Liang et al., 2022) subsequently attempted to address saturation by broadening evaluation to hundreds of tasks and multiple evaluation axes (accuracy, calibration, robustness, fairness, efficiency), respectively.

4.1.2 Safety, alignment, and red-teaming challenges

LLM competitions are paying explicit attention to safety and alignment. For instance, the Trojan Detection Challenge[61] at NeurIPS 2022 asked participants to identify backdoor triggers hidden in fine-tuned language models, targeting robustness concerns that had been difficult to study without a curated competition framework. Similarly, the ARC-Evals / METR[62] evaluation effort has motivated the community to design challenges that test not only capability but also the degree to which models are alingned with human values. On the other hand, red-teaming competitions, where participants attempt to elicit harmful, biased, or other undesirable outputs from a target model, have emerged as a crowd-sourcing mechanism for safety research. In this context, the Generative AI red teaming challenge[63] collocated with DEFCON aimed to identify failure modes in several frontier LLMs, producing a dataset of adversarial prompts that is now used in safety research. This form of competition is arguably a natural extension of the penetration-testing competitions long established in the cybersecurity community.

4.1.3 Text and code generation

The evaluation of free-form text generation has prompted novel competition designs. Notably, the ChatBot Arena (Chiang et al., 2024), introduced a crowd-sourced evaluation model in which human annotators rank the outputs of two anonymized models in pairwise comparisons. The resulting Elo-style leaderboard has become an influential reference for conversational LLM quality and has inspired similar arena-style evaluations for code generation and multimodal models. Coding challenges have proven to be a fruitful testbed for LLM evaluation. HumanEval (Chen et al., 2021) and the Mostly Basic Python Problems benchmark (Austin et al., 2021) established functional correctness as a primary metric, enabling automatic evaluation of free-form code produced by language models. SWE-bench (Jimenez et al., 2024) raised the bar further by posing real GitHub issues from open-source repositories as tasks, requiring models to produce patches that pass the associated test suites. These benchmarks have many of the structural properties of traditional competitions, a fixed task definition, a held-out evaluation set, and a public leaderboard, and have guided several generations of code-specialized LLMs.

61. https://trojandetection.ai/
62. https://metr.org/
63. https://humane-intelligence.org/get-involved/events/defcon-2023-overview/

Overall, the LLM era has brought both opportunities and challenges for the competition community. On one hand, it has democratized participation, while on the other hand it has raised new questions around evaluation validity, benchmark contamination, and the separation of genuine generalization from memorization of training data.

4.2 Challenges on AI Agents and Agentic AI

An autonomous agent is defined as an entity that perceives its environment, makes plans, and takes sequences of actions to achieve a goal. This notion has motivated competitions since the early days of RoboCup and the DARPA Grand Challenges (see Section 2.1). However, the convergence of LLMs with tool usage, memory, and multi-step planning has given rise to a new generation of *agentic AI* challenges that differ substantially from their predecessors in both scope and evaluation methodology.

4.2.1 Web and computer use agents

WebArena (Zhou et al., 2023) and WorkArena (Drouin et al., 2024) present agent benchmarks in which participants build systems capable of completing realistic web-browsing tasks in a sandboxed environment. The associated leaderboards have attracted considerable attention and revealed the significant gap between human performance and the best current agents. Mind2Web and VisualWebBench extend the evaluation to more diverse web environments and introduce vision as a primary input modality (Deng et al., 2023; Liu et al., 2024). OSWorld broadens the scope further by evaluating agents that interact with a full desktop operating system across a wide range of application software, constituting one of the most challenging open-ended agent benchmarks available (Xie et al., 2024).

4.2.2 Tool use and function calling

The evaluation of LLM tool usage has become a distinct sub-field within agent benchmarks. ToolBench evaluates whether an agent can correctly select and invoke tools from catalogs Xu et al. (2023), while BFCL (Berkeley Function Calling Leaderboard) provides a structured leaderboard for function-calling accuracy across programming languages and API formats (Patil et al., 2025). These benchmarks are particularly relevant for enterprise and production use cases where agents are expected to interact reliably with external systems.

4.2.3 Multi-agent challenges

An important dimension of agentic AI is the interaction between multiple agents, whether cooperative, competitive, or mixed. The Melting Pot benchmark evaluates the social generalisation of reinforcement learning agents across a diverse suite of multi-agent substrates (Leibo et al., 2021). Concordia extends this to LLM-based agents in social simulation settings (Vezhnevets et al., 2023). Multi-agent challenges introduce the additional complexity of emergent communication, coordination, and strategic reasoning, and several NeurIPS competition track entries have explored these themes (see Section 3.1). Cooperative multi-agent settings are of particular relevance for agentic AI given the growing

deployment of LLM-orchestrated pipelines in which multiple specialised agents collaborate to complete a task.

Summarizing, agent challenges represent one of the most rapidly evolving frontiers in academic competitions. They inherit and extend the tradition of interactive evaluation established by RoboCup and the DARPA series, while introducing new methodological demands around long-horizon evaluation, multi-modal perception, tool use, and safety. The connection to LLM challenges is intimate: most state-of-the-art agents today are built on top of foundation models, and progress on agent benchmarks is tightly coupled to progress in LLM capability and alignment.

4.3 Data-Centric Competitions

Historically, the dominant paradigm in machine learning competitions has been *model-centric*, that is, participants are given a fixed dataset and are asked to build the model that achieves the best score on a held-out test set. While effective, it has also been criticized for incentivizing marginal architectural improvements and hyper-parameter tuning at the expense of deeper scientific contributions, see Section 2.3.2. *Data-centric AI* inverts this relationship by fixing the model and asking participants to improve the dataset itself.

4.3.1 Label quality and annotation

Dataperf a benchmark suite developed under the MLCommons umbrella, formalizes data-centric evaluation across multiple tasks including image classification, keyword spotting, and hate speech detection (Mazumder et al., 2023). Dataperf has introduced the notion of *data selection* challenges, where a budget on the number of training examples is imposed and participants must select the most informative subset. This directly implements active learning, core-set selection, and curriculum learning research within a competition framework.

4.3.2 Dataset creation as a competition task

An emerging and particularly innovative competition format asks participants not merely to improve an existing dataset but to *create* a new one. The Dynabench [64] platform implements a human-and-model-in-the-loop evaluation paradigm in which annotators are asked to create examples that fool a current best model, and the resulting examples are added to an ever-growing dynamic benchmark. Competitions hosted on Dynabench for natural language inference, sentiment analysis, and question answering have produced datasets that are systematically harder than those assembled by conventional annotation pipelines, because adversarial human annotators deliberately probe model weaknesses.

Data-centric competitions are still maturing as a genre. Several open methodological questions remain: how to define a fair fixed-model baseline, how to prevent participants from effectively performing model selection under the guise of data selection, and how to credit contributions that improve generalisation on future test distributions rather than a fixed held-out set. Nevertheless, data-centric challenges represent an important complement to traditional model-centric competitions, and their emphasis on dataset quality is particularly

64. `https://dynabench.org/`

timely given the growing recognition that foundation models are only as good as the data on which they are trained.

5 Discussion

Academic challenges have been decisive for the consolidation of fields of knowledge. This chapter provided an historical review and an analysis of benefits and limitations of challenges, while it is true that competitions can have undesired effects, there is palpable evidence that they have boosted research across a number of fields. In fact there are several examples of breakthrough discoveries that have arosen in the context of academic competitions.

While we are witnessing the establishment of academic competitions as a way to advance the state of the art, the forthcoming years are promising. Specifically, we consider that the following lines of research will be decisive in the next few years:

- **Multimodal LLM challenges.** The integration of vision and language in modern foundation models has given rise to a new wave of multimodal challenges. However, challenges targeting heterogeneous an not common modalities (e.g., spectra, graphs, etc.) will emerge in the following years.

- **Multi-task and reasoning challenges.** Reasoning is a quality that is highly subjective to evaluate in LLMs, designing robust and highly reliable competitions for this topic represents a major opportunity.

- **Cooperative competitions.** Coopetitions is a form of crowd sourcing in which participants compete to build the best solution for a problem, but they cooperate with other participants in order to obtain an additional reward (e.g., information from other participants, higher scores, etc.).

- **Challenges for education.** Exploiting the full potential of challenges in education is a challenge itself, but we think this is a valuable resource for reaching wider audiences with assignments that require solving practical problems.

- **Academic challenges for good.** This is a topic being pursued and encouraged by evaluation forums and competition tracks, consider for instance the NeurIPS competition track (Escalante and Hadsell, 2019; Escalante and Hofmann, 2020; Kiela et al., 2022).

- **Dedicated publications for challenges.** There are few dedicated forums in which results of challenges are published (consider for instance the Challenges in Machine Learning series[65]). We foresee more dedicated venues will be available in the next few years.

65. https://www.springer.com/series/15602

Acknowledgments and Disclosure of Funding

The authors are grateful with Isabelle Guyon and Adrien Pavao for comments and suggestions that improved the manuscript.

References

Vincent Andrearczyk, Valentin Oreiller, Sarah Boughdad, Catherine Cheze Le Rest, Hesham Elhalawani, Mario Jreige, John O. Prior, Martin Vallières, Dimitris Visvikis, Mathieu Hatt, and Adrien Depeursinge. Overview of the hecktor challenge at miccai 2021: Automatic head and neck tumor segmentation and outcome prediction in pet/ct images. In Vincent Andrearczyk, Valentin Oreiller, Mathieu Hatt, and Adrien Depeursinge, editors, *Head and Neck Tumor Segmentation and Outcome Prediction*, pages 1–37, Cham, 2022. Springer International Publishing. ISBN 978-3-030-98253-9.

Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program synthesis with large language models. *CoRR*, abs/2108.07732, 2021. URL `https://arxiv.org/abs/2108.07732`.

George Awad, Asad A. Butt, Keith Curtis, Jonathan G. Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, Lukas L. Diduch, Jeffrey Liu, Alan F. Smeaton, Yvette Graham, Gareth J. F. Jones, Wessel Kraaij, and Georges Quénot. TRECVID 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains. *CoRR*, abs/2104.13473, 2021. URL `https://arxiv.org/abs/2104.13473`.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Lee, Jue Wang, Qian Cong, Lisa Kinch, Richard Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb Glassman, Andy Degiovanni, Jose Pereira, Andria Rodrigues, Alberdina Dijk, Ana Ebrecht, and David Baker. Accurate prediction of protein structures and interactions using a 3-track network, 06 2021.

Woonhyuk Baek, Ildoo Kim, Sungwoong Kim, and Sungbin Lim. Autoclint: The winning method in autocv challenge 2019. *arXiv*, 2020.

Shuang Bai and Shan An. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304, 2018. doi: 10.1016/j.neucom.2018.05.080. URL `https://doi.org/10.1016/j.neucom.2018.05.080`.

Xavier Baró, Jordi Gonzàlez, Junior Fabian, Miguel Ángel Bautista, Marc Oliu, Hugo Jair Escalante, Isabelle Guyon, and Sergio Escalera. Chalearn looking at people 2015 challenges: Action spotting and cultural event recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2015, Boston, MA, USA, June 7-12, 2015*, pages 1–9. IEEE Computer Society, 2015. doi: 10.1109/CVPRW.2015.7301329. URL `https://doi.org/10.1109/CVPRW.2015.7301329`.

Alan Black and Maxine Eskenazi. The spoken dialogue challenge. In *Proceedings of the SIGDIAL 2009 Conference*, pages 337–340, London, UK, September 2009. Association for Computational Linguistics. URL https://aclanthology.org/W09-3950.

Martin Braschler. Clef 2000 — overview of results. In Carol Peters, editor, *Cross-Language Information Retrieval and Evaluation*, pages 89–101, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-44645-3.

Harald Carlens. State of competitive machine learning in 2022. *ML Contests Research*, 2023. https://mlcontests.com/state-of-competitive-machine-learning-2022.

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

Harald Carlens. State of machine learning competitions in 2024. *ML Contests Research*, 2025. https://mlcontests.com/state-of-machine-learning-competitions-2024.

Dustin Carrión-Ojeda, Hong Chen, Adrian El Baz, Sergio Escalera, Chaoyu Guan, Isabelle Guyon, Ihsan Ullah, Xin Wang, and Wenwu Zhu. Neurips'22 cross-domain metadl competition: Design and baseline results, 2022. URL https://arxiv.org/abs/2208.14686.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021. URL https://arxiv.org/abs/2107.03374.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Marco Ciccone, Gustavo Stolovitzky, and Jacob Albrecht. Neurips 2022 competition track revised selected papers. In Marco Ciccone, Gustavo Stolovitzky, and Jacob Albrecht, editors, *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages i–i. PMLR, 28 Nov–09 Dec 2023. URL https://proceedings.mlr.press/v220/ciccone23a.html.

Albert Clapés, Júlio C. S. Jacques Júnior, Carla Morral, and Sergio Escalera. Chalearn LAP 2020 challenge on identity-preserved human detection: Dataset and results. In

15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020, Buenos Aires, Argentina, November 16-20, 2020, pages 801–808. IEEE, 2020. doi: 10.1109/FG47880.2020.00135. URL https://doi.org/10.1109/FG47880.2020.00135.

Paul D. Clough, Henning Müller, and Mark Sanderson. Seven years of image retrieval evaluation. In Henning Müller, Paul D. Clough, Thomas Deselaers, and Barbara Caputo, editors, *ImageCLEF, Experimental Evaluation in Visual Information Retrieval*, pages 3–18. Springer, 2010. doi: 10.1007/978-3-642-15181-1_1. URL https://doi.org/10.1007/978-3-642-15181-1_1.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web, 2023.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, Léo Boisvert, Megh Thakkar, Quentin Cappart, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. Workarena: How capable are web agents at solving common knowledge work tasks?, 2024.

Christian Eichenberger, Moritz Neun, Henry Martin, Pedro Herruzo, Markus Spanring, Yichao Lu, Sungbin Choi, Vsevolod Konyakhin, Nina Lukashina, Aleksei Shpilman, Nina Wiedemann, Martin Raubal, Bo Wang, Hai L. Vu, Reza Mohajerpoor, Chen Cai, Inhi Kim, Luca Hermes, Andrew Melnik, Riza Velioglu, Markus Vieth, Malte Schilling, Alabi Bojesomo, Hasan Al Marzouqi, Panos Liatsis, Jay Santokhi, Dylan Hillier, Yiming Yang, Joned Sarwar, Anna Jordan, Emil Hewage, David Jonietz, Fei Tang, Aleksandra Gruca, Michael Kopp, David Kreil, and Sepp Hochreiter. Traffic4cast at neurips 2021 - temporal and spatial few-shot transfer learning in gridded geo-spatial processes. In Douwe Kiela, Marco Ciccone, and Barbara Caputo, editors, *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pages 97–112. PMLR, 06–14 Dec 2022. URL https://proceedings.mlr.press/v176/eichenberger22a.html.

Adrian El Baz, Isabelle Guyon, Zhengying Liu, Jan N. van Rijn, Sebastien Treguer, and Joaquin Vanschoren. Metadl challenge design and baseline results. In *AAAI Workshop on Meta-Learning and MetaDL Challenge*, volume 140 of *Proceedings of Machine Learning Research*, pages 1–16. PMLR, 2021a.

Adrian El Baz, Ihsan Ullah, Edesio Alcobaça, André C. P. L. F. Carvalho, Hong Chen, Fabio Ferreira, Henry Gouk, Chaoyu Guan, Isabelle Guyon, Timothy Hospedales, Shell Hu, Mike Huisman, Frank Hutter, Zhengying Liu, Felix Mohr, Ekrem Öztürk, Jan N van Rijn, Haozhe Sun, Xin Wang, and Wenwu Zhu. Lessons learned from the NeurIPS 2021 MetaDL challenge: Backbone fine-tuning without episodic meta-learning dominates for few-shot learning image classification. In *NeurIPS 2021 Competition and Demonstration Track*, On-line, United States, December 2021b. URL https://hal.archives-ouvertes.fr/hal-03688638.

Hugo Jair Escalante and Raia Hadsell. Neurips 2019 competition and demonstration track revised selected papers. In Hugo Jair Escalante and Raia Hadsell, editors, *NeurIPS 2019 Competition and Demonstration Track, 8-14 December 2019, Vancouver, Canada. Revised selected papers*, volume 123 of *Proceedings of Machine Learning Research*, pages 1–12. PMLR, 2019. URL http://proceedings.mlr.press/v123/escalante20a.html.

Hugo Jair Escalante and Katja Hofmann. Neurips 2020 competition and demonstration track: Revised selected papers. In Hugo Jair Escalante and Katja Hofmann, editors, *NeurIPS 2020 Competition and Demonstration Track, 6-12 December 2020, Virtual Event / Vancouver, BC, Canada*, volume 133 of *Proceedings of Machine Learning Research*, pages 1–2. PMLR, 2020. URL http://proceedings.mlr.press/v133/escalante21a.html.

Hugo Jair Escalante, Carlos A. Hernández, Jesús A. González, Aurelio López-López, Manuel Montes-y-Gómez, Eduardo F. Morales, Luis Enrique Sucar, Luis Villaseñor Pineda, and Michael Grubinger. The segmented and annotated IAPR TC-12 benchmark. *Comput. Vis. Image Underst.*, 114(4):419–428, 2010. doi: 10.1016/j.cviu.2009.03.008. URL https://doi.org/10.1016/j.cviu.2009.03.008.

Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Júlio C. S. Jacques Júnior, Meysam Madadi, Xavier Baró, Stéphane Ayache, Evelyne Viegas, Yagmur Güçlütürk, Umut Güçlü, Marcel A. J. van Gerven, and Rob van Lier. Design of an explainable machine learning challenge for video interviews. In *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, AK, USA, May 14-19, 2017*, pages 3688–3695. IEEE, 2017. doi: 10.1109/IJCNN.2017.7966320. URL https://doi.org/10.1109/IJCNN.2017.7966320.

Hugo Jair Escalante, Wei-Wei Tu, Isabelle Guyon, Daniel L. Silver, Evelyne Viegas, Yuqiang Chen, Wenyuan Dai, and Qiang Yang. Automl @ neurips 2018 challenge: Design and results. *CoRR*, abs/1903.05263, 2019. URL http://arxiv.org/abs/1903.05263.

Hugo Jair Escalante, Heysem Kaya, Albert Ali Salah, Sergio Escalera, Yagmur Güçlütürk, Umut Güçlü, Xavier Baró, Isabelle Guyon, Júlio C. S. Jacques Júnior, Meysam Madadi, Stéphane Ayache, Evelyne Viegas, Furkan Gürpinar, Achmadnoer Sukma Wicaksana, Cynthia C. S. Liem, Marcel A. J. van Gerven, and Rob van Lier. Modeling, recognizing, and explaining apparent personality from videos. *IEEE Trans. Affect. Comput.*, 13(2):894–911, 2022. doi: 10.1109/TAFFC.2020.2973984. URL https://doi.org/10.1109/TAFFC.2020.2973984.

Sergio Escalera, Jordi Gonzàlez, Xavier Baró, Miguel Reyes, Oscar Lopes, Isabelle Guyon, Vassilis Athitsos, and Hugo Jair Escalante. Multi-modal gesture recognition challenge 2013: dataset and results. In Julien Epps, Fang Chen, Sharon L. Oviatt, Kenji Mase, Andrew Sears, Kristiina Jokinen, and Björn W. Schuller, editors, *2013 International Conference on Multimodal Interaction, ICMI '13, Sydney, NSW, Australia, December 9-13, 2013*, pages 445–452. ACM, 2013. doi: 10.1145/2522848.2532595. URL https://doi.org/10.1145/2522848.2532595.

Sergio Escalera, Xavier Baró, Jordi Gonzàlez, Miguel Ángel Bautista, Meysam Madadi, Miguel Reyes, Víctor Ponce-López, Hugo Jair Escalante, Jamie Shotton, and Isabelle Guyon. Chalearn looking at people challenge 2014: Dataset and results. In Lourdes Agapito, Michael M. Bronstein, and Carsten Rother, editors, *Computer Vision - ECCV 2014 Workshops - Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I*, volume 8925 of *Lecture Notes in Computer Science*, pages 459–473. Springer, 2014. doi: 10.1007/978-3-319-16178-5_32. URL https://doi.org/10.1007/978-3-319-16178-5_32.

Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzàlez, Hugo Jair Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *2015 IEEE International Conference on Computer Vision Workshop, ICCV Workshops 2015, Santiago, Chile, December 7-13, 2015*, pages 243–251. IEEE Computer Society, 2015. doi: 10.1109/ICCVW.2015.40. URL https://doi.org/10.1109/ICCVW.2015.40.

Sergio Escalera, Mercedes Torres, Brais Martínez, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Georgios Tzimiropoulos, Ciprian A. Corneanu, Marc Oliu, Mohammad Ali Bagheri, and Michel F. Valstar. Chalearn looking at people and faces of the world: Face analysisworkshop and challenge 2016. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2016, Las Vegas, NV, USA, June 26 - July 1, 2016*, pages 706–713. IEEE Computer Society, 2016. doi: 10.1109/CVPRW.2016.93. URL https://doi.org/10.1109/CVPRW.2016.93.

Sergio Escalera, Xavier Baró, Hugo Jair Escalante, and Isabelle Guyon. Chalearn looking at people: A review of events and resources. In *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, AK, USA, May 14-19, 2017*, pages 1594–1601. IEEE, 2017a. doi: 10.1109/IJCNN.2017.7966041. URL https://doi.org/10.1109/IJCNN.2017.7966041.

Sergio Escalera, Isabelle Guyon, and Vassilis Athitsos, editors. *Gesture Recognition*. Springer, 2017b. ISBN 978-3-319-57020-4. doi: 10.1007/978-3-319-57021-1. URL https://doi.org/10.1007/978-3-319-57021-1.

Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Sandy Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, David Jones, David Silver, Koray Kavukcuoglu, Demis Hassabis, and Andrew Senior. De novo structure prediction with deep-learning based scoring, 12 2018.

M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, January 2015.

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. *Auto-sklearn: Efficient and Robust Automated Machine*

Learning, pages 113–134. Springer International Publishing, Cham, 2019. ISBN 978-3-030-05318-5. doi: 10.1007/978-3-030-05318-5_6. URL https://doi.org/10.1007/978-3-030-05318-5_6.

Ryan W. Gardner, Corey Lowman, Casey Richardson, Ashley J. Llorens, Jared Markowitz, Nathan Drenkow, Andrew Newman, Gregory Clark, Gino Perrotta, Robert Perrotta, Timothy Highley, Vlad Shcherbina, William Bernadoni, Mark Jordan, and Asen Asenov. The first international competition in machine reconnaissance blind chess. In Hugo Jair Escalante and Raia Hadsell, editors, *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, volume 123 of *Proceedings of Machine Learning Research*, pages 121–130. PMLR, 08–14 Dec 2020. URL https://proceedings.mlr.press/v123/gardner20a.html.

John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees. The trec spoken document retrieval track: A success story. In *Content-Based Multimedia Information Access - Volume 1*, RIAO '00, page 1–20, Paris, FRA, 2000. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

Michael Garris, J Blue, Gerald Candela, Patrick Grother, Stanley Janet, and Charles Wilson. Nist form-based handprint recognition system, 1997-01-01 1997.

Patrick Grother. Nist special database 19 handprinted forms and characters database, 1995.

Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S. Ren, and Radu Timofte. Ntire 2022 challenge on perceptual image quality assessment, 2022. URL https://arxiv.org/abs/2206.11695.

William Hebgen Guss, Stephanie Milani, Nicholay Topin, Brandon Houghton, Sharada Mohanty, Andrew Melnik, Augustin Harter, Benoit Buschmaas, Bjarne Jaster, Christoph Berganski, Dennis Heitkamp, Marko Henning, Helge Ritter, Chengjie Wu, Xiaotian Hao, Yiming Lu, Hangyu Mao, Yihuan Mao, Chao Wang, Michal Opanowicz, Anssi Kanervisto, Yanick Schraner, Christian Scheller, Xiren Zhou, Lu Liu, Daichi Nishio, Toi Tsuneda, Karolis Ramanauskas, and Gabija Juceviciute. Towards robust and domain agnostic reinforcement learning competitions: Minerl 2020. In Hugo Jair Escalante and Katja Hofmann, editors, *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 233–252. PMLR, 06–12 Dec 2021. URL https://proceedings.mlr.press/v133/guss21a.html.

Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the nips 2003 feature selection challenge. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper/2004/file/5e751896e527c862bf67251a474b3819-Paper.pdf.

Isabelle Guyon, Amir Reza Saffari Azar Alamdari, Gideon Dror, and Joachim M. Buhmann. Performanceprediction challenge. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2006, part of the IEEE World Congress on Computational Intelligence, WCCI 2006, Vancouver, BC, Canada, 16-21 July 2006*, pages 1649–1656.

IEEE, 2006. doi: 10.1109/IJCNN.2006.246632. URL https://doi.org/10.1109/IJCNN.2006.246632.

Isabelle Guyon, Amir Saffari, Gideon Dror, and Gavin C. Cawley. Analysis of the IJCNN 2007 agnostic learning vs. prior knowledge challenge. *Neural Networks*, 21(2-3):544–550, 2008. doi: 10.1016/j.neunet.2007.12.024. URL https://doi.org/10.1016/j.neunet.2007.12.024.

Isabelle Guyon, Lisheng Sun-Hosoya, Marc Boullé, Hugo Jair Escalante, Sergio Escalera, Zhengying Liu, Damir Jajetic, Bisakha Ray, Mehreen Saeed, Michéle Sebag, Alexander R. Statnikov, Wei-Wei Tu, and Evelyne Viegas. Analysis of the automl challenge series 2015-2018. In Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren, editors, *Automated Machine Learning - Methods, Systems, Challenges*, The Springer Series on Challenges in Machine Learning, pages 177–219. Springer, 2019. doi: 10.1007/978-3-030-05318-5_10. URL https://doi.org/10.1007/978-3-030-05318-5_10.

Donna Harman. Overview of the first trec conference. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, page 36–47, New York, NY, USA, 1993. Association for Computing Machinery. ISBN 0897916050. doi: 10.1145/160688.160692. URL https://doi.org/10.1145/160688.160692.

Donna Harman. Overview of the fourth text retrieval conference (TREC-4). In Donna K. Harman, editor, *Proceedings of The Fourth Text REtrieval Conference, TREC 1995, Gaithersburg, Maryland, USA, November 1-3, 1995*, volume 500-236 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 1995. URL http://trec.nist.gov/pubs/trec4/overview.ps.gz.

Donna Harman. The Text REtrieval Conferences (TRECs) and the Cross-Language Track. In *First International Conference on Language Resources & Evaluation, Granada, Spain*, pages 517–522, 1998.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Lorne J Hofseth. Getting rigorous with scientific rigor. *Carcinogenesis*, 39(1):21–25, 08 2017. ISSN 0143-3334. doi: 10.1093/carcin/bgx085. URL https://doi.org/10.1093/carcin/bgx085.

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *AutoML: Methods, Systems, Challenges*. Springer Series in Challenges in Machine Learning. Springer, 2018.

Moult J, Fidelis K, Kryshtafovych A, Schwede T, and Tramontano. Critical assessment of methods of protein structure prediction (casp) round X. *Proteins*, 2(2):1–6, 2014.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues?, 2024. URL https://arxiv.org/abs/2310.06770.

Bo Ju, Wei Yang, Jinrang Jia, Xiaoqing Ye, Qu Chen, Xiao Tan, Hao Sun, Yifeng Shi, and Errui Ding. Danet: Dimension apart network for radar object detection. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, ICMR '21, page 533–539, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384636. doi: 10.1145/3460426.3463656. URL https://doi.org/10.1145/3460426.3463656.

John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David A. Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021.

Shachar Kaufman, Saharon Rosset, and Claudia Perlich. Leakage in data mining: Formulation, detection, and avoidance. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 6, pages 556–563, 01 2011. doi: 10.1145/2020408.2020496.

Lukasz Kidzinski, Sharada Prasanna Mohanty, Carmichael F. Ong, Zhewei Huang, Shuchang Zhou, Anton Pechenko, Adam Stelmaszczyk, Piotr Jarosik, Mikhail Pavlov, Sergey Kolesnikov, Sergey M. Plis, Zhibo Chen, Zhizheng Zhang, Jiale Chen, Jun Shi, Zhuobin Zheng, Chun Yuan, Zhihui Lin, Henryk Michalewski, Piotr Milos, Blazej Osinski, Andrew Melnik, Malte Schilling, Helge J. Ritter, Sean F. Carroll, Jennifer L. Hicks, Sergey Levine, Marcel Salathé, and Scott L. Delp. Learning to run challenge solutions: Adapting reinforcement learning methods for neuromusculoskeletal environments. *CoRR*, abs/1804.00361, 2018. URL http://arxiv.org/abs/1804.00361.

Douwe Kiela, Marco Ciccone, and Barbara Caputo. Neurips 2021 competition and demonstration track revised selected papers. In Douwe Kiela, Marco Ciccone, and Barbara Caputo, editors, *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pages i–ii. PMLR, 06–14 Dec 2022. URL https://proceedings.mlr.press/v176/kiela22a.html.

Hiroaki Kitano, Milind Tambe, Peter Stone, Manuela Veloso, Silvia Coradeschi, Eiichi Osawa, Hitoshi Matsubara, Itsuki Noda, and Minoru Asada. The robocup synthetic agent challenge 97. In Hiroaki Kitano, editor, *RoboCup-97: Robot Soccer World Cup I*, pages 62–73, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg. ISBN 978-3-540-69789-3.

Michael Kopp, David Kreil, Moritz Neun, David Jonietz, Henry Martin, Pedro Herruzo, Aleksandra Gruca, Ali Soleymani, Fanyou Wu, Yang Liu, Jingwei Xu, Jianjin Zhang, Jay Santokhi, Alabi Bojesomo, Hasan Al Marzouqi, Panos Liatsis, Pak Hay Kwok, Qi Qi, and Sepp Hochreiter. Traffic4cast at neurips 2020 - yet more on the unreasonable effectiveness of gridded geo-spatial processes. In Hugo Jair Escalante and Katja Hofmann, editors, *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 325–343. PMLR, 06–12 Dec 2021. URL https://proceedings.mlr.press/v133/kopp21a.html.

Y. Koren. The bellkor solution to the netflix grand prize. Netflix prize documentation 81, 1-10, 2009.

David P Kreil, Michael K Kopp, David Jonietz, Moritz Neun, Aleksandra Gruca, Pedro Herruzo, Henry Martin, Ali Soleymani, and Sepp Hochreiter. The surprising efficiency of framing geo-spatial time series forecasting as a video prediction task – insights from the iarai $\widehat{4}$c competition at neurips 2019. In Hugo Jair Escalante and Raia Hadsell, editors, *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, volume 123 of *Proceedings of Machine Learning Research*, pages 232–241. PMLR, 08–14 Dec 2020. URL https://proceedings.mlr.press/v123/kreil20a.html.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round XIV. *Proteins*, 89(12):1607–1617, 2021.

Joel Z. Leibo, Edgar Dué nez Guzmán, Alexander Sasha Vezhnevets, John P. Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charles Beattie, Igor Mordatch, and Thore Graepel. Scalable evaluation of multi-agent reinforcement learning with melting pot. In *International conference on machine learning*. PMLR, 2021. doi: 10.48550/arXiv.2107.06857. URL https://doi.org/10.48550/arXiv.2107.06857.

Xiaobai Li, Hu Han, Hao Lu, Xuesong Niu, Zitong Yu, Antitza Dantcheva, Guoying Zhao, and Shiguang Shan. The 1st challenge on remote physiological signal sensing (repss), 2020. URL https://arxiv.org/abs/2003.11756.

Xiaobai Li, Haomiao Sun, Zhaodong Sun, Hu Han, Antitza Dantcheva, Shiguang Shan, and Guoying Zhao. The 2nd challenge on remote physiological signal sensing (repss). In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2404–2413, 2021. doi: 10.1109/ICCVW54120.2021.00273.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=iO4LZibEqW.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer*

Science, pages 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zichang Tan, Qi Yuan, Kai Wang, Chi Lin, Guodong Guo, Isabelle Guyon, and Stan Z. Li. Multi-modal face anti-spoofing attack detection challenge at CVPR2019. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1601–1610. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPRW.2019.00202. URL http://openaccess.thecvf.com/content_CVPRW_2019/html/CFS/Liu_Multi-Modal_Face_Anti-Spoofing_Attack_Detection_Challenge_at_CVPR2019_CVPRW_2019_paper.html.

Ajian Liu, Xuan Li, Jun Wan, Yanyan Liang, Sergio Escalera, Hugo Jair Escalante, Meysam Madadi, Yi Jin, Zhuoyuan Wu, Xiaogang Yu, Zichang Tan, Qi Yuan, Ruikun Yang, Benjia Zhou, Guodong Guo, and Stan Z. Li. Cross-ethnicity face anti-spoofing recognition challenge: A review. *IET Biom.*, 10(1):24–43, 2021a. doi: 10.1049/bme2.12002. URL https://doi.org/10.1049/bme2.12002.

Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding?, 2024.

Zhengying Liu, Adrien Pavao, Zhen Xu, Sergio Escalera, Fabio Ferreira, Isabelle Guyon, Sirui Hong, Frank Hutter, Rongrong Ji, Julio C. S. Jacques Junior, Ge Li, Marius Lindauer, Zhipeng Luo, Meysam Madadi, Thomas Nierhoff, Kangning Niu, Chunguang Pan, Danny Stoll, Sebastien Treguer, Jin Wang, Peng Wang, Chenglin Wu, Youcheng Xiong, Arbër Zela, and Yang Zhang. Winning solutions and post-challenge analyses of the chalearn autodl challenge 2019. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9):3108–3125, 2021b. doi: 10.1109/TPAMI.2021.3075372.

Simon M. Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, and Robert Young. ICDAR 2003 robust reading competitions. In *7th International Conference on Document Analysis and Recognition (ICDAR 2003), 2-Volume Set, 3-6 August 2003, Edinburgh, Scotland, UK*, pages 682–687. IEEE Computer Society, 2003. doi: 10.1109/ICDAR.2003.1227749. URL https://doi.org/10.1109/ICDAR.2003.1227749.

L. Marak, J. Cousty, L. Najman, and H. Talbot. 4d morphological segmentation and the miccai lv-segmentation grand challenge. http://hdl.handle.net/10380/3085, 07 2009.

Mitchell P. Marcus. Overview of the fifth DARPA speech and natural language workshop. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992. URL https://aclanthology.org/H92-1001.

T. Matsui, T. Noumi, I. Yamashita, T. Wakahara, and M. Yoshimuro. State of the art of handwritten numeral recognition in japan-the results of the first IPTP character recognition competition. In *2nd International Conference Document Analysis and Recognition, ICDAR '93, October 20-22, 1993, Tsukuba City, Japan*, pages 391–396. IEEE Computer Society, 1993. doi: 10.1109/ICDAR.1993.395709. URL https://doi.org/10.1109/ICDAR.1993.395709.

Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Smriti Raje, Max Bartolo, Sabri Eyuboglu, Amirata Ghorbani, Emmett Goodman, Oana Inel, Tariq Kane, Christine R. Kirkpatrick, Tzu-Sheng Kuo, Jonas Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Lilith Bat-Leah, Ce Zhang, James Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi. Dataperf: Benchmarks for data-centric ai development, 2023. URL https://arxiv.org/abs/2207.10062.

Pablo Meyer and Julio Saez-Rodriguez. Advances in systems biology modeling: 10 years of crowdsourcing dream challenges. *Cell Systems*, 12(6):636–653, 2021. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2021.05.015. URL https://www.sciencedirect.com/science/article/pii/S2405471221002015.

Stephanie Milani, Nicholay Topin, Brandon Houghton, William H. Guss, Sharada P. Mohanty, Keisuke Nakata, Oriol Vinyals, and Noboru Sean Kuno. Retrospective analysis of the 2019 minerl competition on sample efficient reinforcement learning. In Hugo Jair Escalante and Raia Hadsell, editors, *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, volume 123 of *Proceedings of Machine Learning Research*, pages 203–214. PMLR, 08–14 Dec 2020. URL https://proceedings.mlr.press/v123/milani20a.html.

Cristina Palmero, Germán Barquero, Júlio C. S. Jacques Júnior, Albert Clapés, Johnny Núñez, David Curto, Sorina Smeureanu, Javier Selva, Zejian Zhang, David Saeteros, David Gallardo-Pujol, Georgina Guilera, David Leiva, Feng Han, Xiaoxue Feng, Jennifer He, Wei-Wei Tu, Thomas B. Moeslund, Isabelle Guyon, and Sergio Escalera. Chalearn LAP challenges on self-reported personality recognition and non-verbal behavior forecasting during social dyadic interactions: Dataset, design, and results. In Cristina Palmero, Júlio C. S. Jacques Júnior, Albert Clapés, Isabelle Guyon, Wei-Wei Tu, Thomas B. Moeslund, and Sergio Escalera, editors, *ChaLearn LAP Challenge on Understanding Social Behavior in Dyadic and Small Group Interactions, DYAD 2021, held in conjunction with ICCV 2021, Virtual, October 16, 2021*, volume 173 of *Proceedings of Machine Learning Research*, pages 4–52. PMLR, 2021. URL https://proceedings.mlr.press/v173/palmero22b.html.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 48371–48392. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen 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.

Saravanabalagi Ramachandran, Ganesh Sistu, John B. McDonald, and Senthil Kumar Yogamani. Woodscape fisheye semantic segmentation for autonomous driving - CVPR 2021 omnicv workshop challenge. *CoRR*, abs/2107.08246, 2021. URL https://arxiv.org/abs/2107.08246.

Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.91. URL https://doi.org/10.1109/CVPR.2016.91.

Brent R. Rowe, Dallas W. Wood, Albert N. Link, and Diglio A. Simon. Economic impact assessment of nist's text retrieval conference (trec) program. NIST Final Report, 2010.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec 2015. ISSN 1573-1405. doi: 10.1007/s11263-015-0816-y. URL https://doi.org/10.1007/s11263-015-0816-y.

M. Scully, V. Magnotta, C. Gasparovic, P. Pelligrimo, D. Feis, and H. Bockholt. 3d segmentation in the clinic: A grand challenge ii at miccai 2008 - ms lesion segmentation. http://hdl.handle.net/10380/1449, 07 2008.

Rohin Shah, Steven H. Wang, Cody Wild, Stephanie Milani, Anssi Kanervisto, Vinicius G. Goecks, Nicholas Waytowich, David Watkins-Valls, Bharat Prakash, Edmund Mills, Divyansh Garg, Alexander Fries, Alexandra Souly, Chan Jun Shern, Daniel del Castillo, and Tom Lieberum. Retrospective on the 2021 basalt competition on learning from human feedback, 2022. URL https://arxiv.org/abs/2204.07123.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556.

Ozge Mercanoglu Sincan, Júlio C. S. Jacques Júnior, Sergio Escalera, and Hacer Yalim Keles. Chalearn LAP large scale signer independent isolated sign language recognition challenge: Design, results and future research. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 3472–3481. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPRW53098.2021.00386. URL https://openaccess.thecvf.com/content/CVPR2021W/ChaLearn/html/Sincan_ChaLearn_LAP_Large_Scale_Signer_Independent_Isolated_Sign_Language_Recognition_CVPRW_2021_paper.html.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso,

and et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=uyTL5Bvosj.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on image captioning. *CoRR*, abs/2107.06912, 2021. URL https://arxiv.org/abs/2107.06912.

Gustavo Stolovitzky, Robert J. Prill, and Andrea Califano. Lessons from the dream2 challenges. *Annals of the New York Academy of Sciences*, 1158(1):159–195, 2009. doi: https://doi.org/10.1111/j.1749-6632.2009.04497.x. URL https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.2009.04497.x.

Dongmei Sun, Thanh M. Nguyen, Robert J. Allaway, Jelai Wang, Verena Chung, Thomas V. Yu, Michael Mason, Isaac Dimitrovsky, Lars Ericson, Hongyang Li, Yuanfang Guan, Ariel Israel, Alex Olar, Balint Armin Pataki, Gustavo Stolovitzky, Justin Guinney, Percio S. Gulko, Mason B. Frazier, Jake Y. Chen, James C. Costello, Jr Bridges, S. Louis, and RA2-DREAM Challenge Community. A Crowdsourcing Approach to Develop Machine Learning Models to Quantify Radiographic Joint Damage in Rheumatoid Arthritis. *JAMA Network Open*, 5(8):e2227423–e2227423, 08 2022. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2022.27423. URL https://doi.org/10.1001/jamanetworkopen.2022.27423.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1–9. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298594. URL https://doi.org/10.1109/CVPR.2015.7298594.

Adi L. Tarca, Bálint Ármin Pataki, Roberto Romero, Marina Sirota, Yuanfang Guan, Rintu Kutum, Nardhy Gomez-Lopez, Bogdan Done, Gaurav Bhatti, Thomas Yu, Gaia Andreoletti, Tinnakorn Chaiworapongsa, Sonia S. Hassan, Chaur-Dong Hsu, Nima Aghaeepour, Gustavo Stolovitzky, Istvan Csabai, and James C. Costello. Crowdsourcing assessment of maternal blood multi-omics for predicting gestational age and preterm birth. *bioRxiv*, 2020. doi: 10.1101/2020.06.05.130971. URL https://www.biorxiv.org/content/early/2020/06/06/2020.06.05.130971.

Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, Xintao Wang, Yapeng Tian, Ke Yu, Yulun Zhang, Shixiang Wu, Chao Dong, Liang Lin, Yu Qiao, Chen Change Loy, Woong Bae, Jae Jun Yoo, Yoseob Han, Jong Chul Ye, Jae-Seok Choi, Munchurl Kim, Yuchen Fan, Jiahui Yu, Wei Han, Ding Liu, Haichao Yu, Zhangyang Wang, Honghui Shi, Xinchao Wang, Thomas S. Huang, Yunjin Chen, Kai Zhang, Wangmeng Zuo, Zhimin Tang, Linkai Luo, Shaohui Li, Min Fu, Lei Cao, Wen Heng, Giang Bui, Truc Le, Ye Duan, Dacheng Tao, Ruxin Wang, Xu Lin, Jianxin Pang, Jinchang Xu, Yu Zhao, Xiangyu Xu, Jin-shan Pan, Deqing Sun, Yujin Zhang, Xibin Song, Yuchao Dai, Xueying Qin,

Xuan-Phung Huynh, Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, Vishal Monga, Cristóvão Cruz, Karen O. Egiazarian, Vladimir Katkovnik, Rakesh Mehta, Arnav Kumar Jain, Abhinav Agarwalla, Ch V. Sai Praveen, Ruofan Zhou, Hongdiao Wen, Che Zhu, Zhiqiang Xia, Zhengtao Wang, and Qi Guo. NTIRE 2017 challenge on single image super-resolution: Methods and results. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1110–1121. IEEE Computer Society, 2017. doi: 10.1109/CVPRW.2017.149. URL https://doi.org/10.1109/CVPRW.2017.149.

Ryan Turner, David Eriksson, Michael McCourt, Juha Kiili, Eero Laaksonen, Zhen Xu, and Isabelle Guyon. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. In Hugo Jair Escalante and Katja Hofmann, editors, *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 3–26. PMLR, 06–12 Dec 2021. URL https://proceedings.mlr.press/v133/turner21a.html.

Christopher Urmson, Joshua Anhalt, J. Andrew (Drew) Bagnell, Christopher R. Baker, Robert E. Bittner, John M. Dolan, David Duggins, David Ferguson, Tugrul Galatali, Hartmut Geyer, Michele Gittleman, Sam Harbaugh, Martial Hebert, Thomas Howard, Alonzo Kelly, David Kohanbash, Maxim Likhachev, Nick Miller, Kevin Peterson, Raj Rajkumar, Paul Rybski, Bryan Salesky, Sebastian Scherer, Young-Woo Seo, Reid Simmons, Sanjiv Singh, Jarrod M. Snider, Anthony (Tony) Stentz, William (Red) L. Whittaker, and Jason Ziglar. Tartan racing: A multi-modal approach to the darpa urban challenge. Technical report, Carnegie Mellon University, Pittsburgh, PA, April 2007.

Alexander Sasha Vezhnevets, John P Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A Duéñez-Guzmán, William A Cunningham, Simon Osindero, Danny Karmon, and Joel Z Leibo. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. *arXiv preprint arXiv:2312.03664*, 2023.

Ubbo Visser. 20 years of robocup. *Künstliche Intell.*, 30(3-4):217–220, 2016. doi: 10.1007/s13218-016-0439-7. URL https://doi.org/10.1007/s13218-016-0439-7.

Ellen M. Voorhees. The TREC question answering track. *Nat. Lang. Eng.*, 7(4): 361–378, 2001. doi: 10.1017/S1351324901002789. URL https://doi.org/10.1017/S1351324901002789.

Ellen M. Voorhees and Donna Harman. The text retrieval conferences (TRECS). In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, MD, USA, October 13-15, 1998*, pages 241–273. Morgan Kaufmann, 1998. doi: 10.3115/1119089.1119127. URL https://aclanthology.org/X98-1031/.

Ellen M. Voorhees and Donna K. Harman, editors. *TREC - Experiment and Evaluation in Information Retrieval*. The MIT Press, Cambridge, 2005.

Jun Wan, Sergio Escalera, Gholamreza Anbarjafari, Hugo Jair Escalante, Xavier Baró, Isabelle Guyon, Meysam Madadi, Juri Allik, Jelena Gorbova, Chi Lin, and Yiliang

Xie. Results and analysis of chalearn LAP multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pages 3189–3197. IEEE Computer Society, 2017. doi: 10.1109/ICCVW.2017.377. URL https://doi.org/10.1109/ICCVW.2017.377.

Jun Wan, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z. Li. *Multi-Modal Face Presentation Attack Detection*. Synthesis Lectures on Computer Vision. Morgan & Claypool Publishers, 2020. doi: 10.2200/S01032ED1V01Y202007COV017. URL https://doi.org/10.2200/S01032ED1V01Y202007COV017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

Yizhou Wang, Jenq-Neng Hwang, Gaoang Wang, Hui Liu, Kwang-Ju Kim, Hung-Min Hsu, Jiarui Cai, Haotian Zhang, Zhongyu Jiang, and Renshu Gu. Rod2021 challenge: A summary for radar object detection challenge for autonomous driving applications. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 553–559, 2021a.

Yizhou Wang, Zhongyu Jiang, Yudong Li, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. Rodnet: A real-time radar object detection network cross-supervised by camera-radar fused object 3d localization. *IEEE Journal of Selected Topics in Signal Processing*, 15 (4):954–967, 2021b. doi: 10.1109/JSTSP.2021.3058895.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments, 2024.

Qiantong Xu, Fenglu Hong, Bo Li, Changran Hu, Zhengyu Chen, and Jian Zhang. On the tool manipulation capability of open-source large language models, 2023.

Chhavi Yadav and Léon Bottou. Cold case: The lost mnist digits. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.