AI Competitions and Benchmarks

The Science Behind the Contests

Edited by Adrien Pavão, Isabelle Guyon and Evelyne Viegas

Written by Jacob Albrecht, Gaia Andreoletti, Prasanna Balaprakash, Xavier Baró, Kristin P. Bennett, Julie Bletz, Yuna Blum, Paul Boutros, Harald Carlens, Albert Clapés, James C. Costello, Phil Culliton, Romain Egele, Hugo Jair Escalante, Sergio Escalera, Simon Frieder, Justin Guinney, Isabelle Guyon, Addison Howard, Julio C. S. Jacques Junior, Aleksandra Kruchinina, Antoine Marot, Thomas Moeslund, Luis Oala, Adrien Pavão, Walter Reade, Anka Reuel, Magali Richard, David Rousseau, Julio Saez-Rodriguez, Gustavo Stolovitsky, Sébastien Tréguer, Wei-Wei Tu, Andrey Ustyuzhanin, Jan N. Van Rijn, Joaquin Vanschoren, Evelyne Viegas, Jun Wan, Zhen Xu and Mouadh Yagoubi

2025 **ISBN:** XXX-XX-XXXX-XXXX-X

Contents

Foreword

In the rapidly evolving landscape of artificial intelligence (AI), the significance of competitions and benchmarks cannot be overstated. This book provides a comprehensive exploration of the role, design, and impact of AI challenges and benchmarks across academic, industrial, and educational domains. From historical perspectives and design principles to hands-on tutorials, the book offers an invaluable analysis of the organization and execution of AI competitions.

This book compiles insights from experienced challenge organizers, providing guidelines for the effective design of data-driven scientific competitions. The authors represent various institutions from academia, industry, and non-profit.

The book offers critical insights for researchers, engineers, and organizers to develop high impact competitions, through an exploration of dataset development, evaluation metrics, competition platforms, incentives, execution, and practical aspects. By addressing both theoretical and real-world considerations, this book serves as an essential guide for anyone looking to understand, participate in, or organize AI challenges and benchmarks.

Over the last 15 years, challenges in machine learning, data science, and artificial intelligence have proven to be effective and cost-efficient methods for rapidly bringing research solutions into industry. They have also emerged as a means to direct academic research, advance the state-of-the-art, and explore entirely new domains. Additionally, these challenges, with their engaging and playful nature, naturally attract students, making them an excellent educational resource. Finally, challenges act as a catalyst for community engagement by offering a structured and stimulating environment for individuals to collectively work towards a common goal.

This book addresses the gap in the literature on the theoretical foundations and optimization of challenge protocols, which has persisted despite the remarkable successes and progress achieved in challenge organization. It assembles leading experts in challenge organization to provide insights and directions for future research. It also provides a deeper understanding of challenge design, and introduces new methods and application domains for designing and implementing high-impact challenges that advance the frontiers of innovation.

Acknowledgments and Disclosure of Funding

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

Broader Impact Statement

This book offers educational benefits, emphasizing the pedagogic value of challenges in AI. The content fosters community engagement, showcasing how challenges can drive collective innovation. Moreover, it reinforces the bridge between academia and industry, highlighting the transformative role of challenges in transitioning research to real-world applications. The book encourages the community to design competitions and benchmarks well aligned with real-world needs and ethical codes.

The life cycle of challenges and benchmarks

Gustavo Stolovitzky GUSTAVO.STOLO@GMAIL.COM
DREAM Challenges, New York, New York, USA

Julio Saez-Rodriguez SAEZ@EBI.AC.UK
European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridgeshire, U.K. and Heidelberg University, Faculty of Medicine, and Heidelberg University Hospital, Institute for Computational Biomedicine, Heidelberg, Germany

Julie Bletz JULIE.BLETZ@ARPA-H.GOV
Blu Omega, Ashburn, VA, USA

Jacob Albrecht JAKE.ALBRECHT@SAGEBASE.ORG
Sage Bionetworks, Seattle, WA, USA

Gaia Andreoletti GAIA.ANDREOLETTI@SAGEBASE.ORG
Sage Bionetworks, Seattle, WA, USA

James C. Costello JAMES.COSTELLO@CUANSCHUTZ.EDU
Department of Pharmacology, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Paul Boutros PBOUTROS@MEDNET.UCLA.EDU
Department of Urology and Human Genetics, University of California, Los Angeles, CA, USA

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

> ... it is known with certainty that there is scarcely anything which more greatly excites noble and ingenious spirits to labors which lead to the increase of knowledge than to propose difficult and at the same time useful problems through the solution of which, as by no other means, they may attain to fame and build for themselves eternal monuments among posterity.
>
> *Johann Bernoulli, June 1696*
> Bernoulli

Abstract

Data Science research is undergoing a revolution fueled by the transformative power of technology, the Internet, and an ever-increasing computational capacity. The rate at which sophisticated algorithms can be developed is unprecedented, yet they remain outpaced by the massive amounts of data that are increasingly available to researchers. Here we argue for the need to creatively leverage the scientific research and algorithm development community as an axis of robust innovation. En-

gaging these communities in the scientific discovery enterprise by critical assessments, community experiments, and/or crowdsourcing will multiply opportunities to develop new data-driven, reproducible and well-benchmarked algorithmic solutions to fundamental and applied problems of current interest. Coordinated community engagement in the analysis of highly complex and massive data has emerged as one approach to find robust methodologies that best address these challenges. When community engagement is done in the form of challenges — by which we mean a skill-based scientific contest, with a limited time duration, ending by a total ranking of participants according to a pre-defined scoring metric, and the selection of winners — the validation of the analytical methodology is inherently addressed, establishing performance benchmarks. Finally, challenges foster open innovation across multiple disciplines to create communities that collaborate directly or indirectly to address significant scientific gaps. Together, participants can solve important problems as varied as health research, climate change, and social equity. Ultimately, challenges can catalyze and accelerate the synthesis of complex data into knowledge or actionable information, and should be viewed a powerful tool to make lasting social and research contributions.

Keywords: crowdsourcing, benchmarking, contest, competition, challenge

1 Brief history of crowdsourcing

The idea of leveraging a community of experts and non-experts to solve a scientific problem has been around for hundreds of years. One early example is the 1714 British Board of Longitude Prize, which was to be awarded to the person who could solve arguably the most important technological problem of the time: to determine a ship's longitude at sea Sobel (2005). After eluding many established scientists of the time, the prize was awarded to John Harrison for his invention of the marine chronometer. There are two important take home messages from the Longitude Prize example. One is the fact that the winner of the prize was John Harrison, an unknown carpenter and clock-maker, and not a more recognized scientist of that era. The second key idea is that the problem was posed as an open participation contest (defined here as an event created by organizers, governed by rules, and directed to a group of participants, offering the opportunity to win an award or a prize), what we refer to today as crowdsourcing. When it comes to analyzing large or novel datasets, it is very likely that the analytical methods and breakthroughs that get the most useful signal may reside with groups other than the data generators or the most established and best published groups in a field.

Coined by Jeff Howe in an article in Wired Magazine Howe (2006), crowdsourcing mixes the bottom-up creative intelligence of the community that volunteers solutions with the top-down management of the organization that poses the problem. Crowdsourcing has been used in many contexts such as business (the design of consumer products) Boudreau and Lakhani (2013), journalism (the collection of information), and peer-review (in the evaluation of patent applications). Here, we are interested in the application of crowdsourcing to computational problems in science and technology.

2 Types of crowdsourcing

Different types of crowdsourcing exist. Generally speaking, crowdsourcing can refer to efforts in which the crowd provides or generates data (e.g. patients provide their medical information) to be mined by others, or alternatively competitions –a skill-based scientific contest, with a limited time duration, involving the submission of proposals, project propositions, project outcomes, and/or prototypes that are evaluated by a panel of judges– or challenges -a skill-based scientific contest, with a limited time duration, ending by a total ranking of participants according to a pre-defined scoring metric, and the selection of winners– in which the crowd actively works on solving a problem from established data Good and Su (2013).

With respect to active crowdsourcing, there are several types. There is labor-focused crowdsourcing, where the job is made open and any individual willing to work can take up the job Boudreau and Lakhani (2013). A well-known example of labor-focused crowdsourcing is the 'Mechanical Turk' or MTurk service run by Amazon and named after an 18th century chess automaton with a hidden human inside. The MTurk approach provides an online workforce that allows people to complete work, or "Human Intelligence Tasks", in exchange for a small amount of money Goodman et al. (2013).

A problem can also be crowdsourced in a manner that divides it up into a set of separate smaller tasks to solve. Crowdsourcing of data annotation and curation in Bioinformatics can be handled well with this approach. This scheme has also been applied to provide pathway resources Ansari et al. (2013); Kutmon et al. (2016), reconstruct the human metabolic network Thiele et al. (2013), annotate molecular interactions in Mycobacterium tuberculosis Vashisht et al. (2012), and identify critical errors in ontologies Mortensen et al. (2015).

In contrast to labor-focused forms of crowdsourcing, where people are paid for their efforts, there are forms of crowdsourcing where individuals volunteer their effort because of their interest in the project or cause. An example of this is the crowdsourced approach taken to the development of Wikipedia, the Internet's largest and most popular general reference source. In some instances, such as Wikipedia and the protein structure game Foldit Cooper et al. (2010), participants contribute their time and intellectual capacity, while in other examples, such as the Folding@home Larson et al. (2009) and rosetta@home Das et al. (2007) projects, participants provide computational power from their personal equipment to help solve the problem. Interestingly, Foldit evolved from rosetta@home, when participants realized that, in some cases, they could intuitively see better structures than those predicted computationally. NASA has also leveraged the power of crowdsourcing with its Citizen Science Projects NAS in which NASA scientists and interested members of the public (the citizen scientists) work together in dozens of astronomy problems such as finding new features in Jupiter's atmosphere and searching for exoplanets, to name a few.

In some instances, crowdsourcing can be implemented in the form of a game Good and Su (2011) in order to maximize the number of solvers who work on the problem and increase the likelihood that they will stay engaged. For example, in the Foldit project, the problem of determining protein structure is transformed into an entertaining game. Such "gamification", where one applies game-design elements to allow an enjoyable experience, has proven a spectacular approach to raise participant numbers and interest. It also leads to impact: Foldit, with its hundreds of thousand registered players, provided useful results that matched or outperformed algorithmically computed solutions Cooper et al. (2010). Foldit was followed by a similarly popular project, EteRNA Treuille and Das (2014), where more than 250,000 registered participants, provided an RNA sequence that fits in a given shape. The best designs, as chosen by the community, were then tested experimentally Lee et al. (2014); Cooper et al. (2010). More recently, the EteRNA platform was used to task the community of RNA enthusiasts to design RNA molecular sensors Andreasson et al. (2022). The thousands of designed RNA molecules were synthesized and experimentally tested using high-throughput biochemical assays as part of the "game", in an iterative process that closed the loop between the real world and the online world. A diversity of tasks can be embedded in existing games. For example, a mini-game in the EVE Online game engaged millions of registered gamers to provide tens of million classifications. Combining these with deep learning led to largely improve image-classification Sullivan et al. (2018). In another example, the Borderlands Science mini-game in Borderlands 3 Waldispühl et al. (2020) engaged over a million participants to solve 50 million tasks within three months. The results of the mini-game are aggregated to improve a reference alignment of millions of ribosomal sequences for gut flora.

Crowdsourcing projects are also effective for collecting new ideas or directions that may be needed to solve a tough problem. These are referred to as "ideation" competitions, and the Board of Longitude Prize mentioned above falls within this category. More recently the Longitude Prize 2014 Rees (2014) built on the success of its predecessor to address the problem of antibiotic resis-

tance through the creation of point-of-care test kits for bacterial infections. Amongst many other ideation prizes, the XPRIZE Qualcomm Tricoder prize Chandler (2014) encouraged participants to develop a handheld wireless device that monitors and diagnoses health conditions.

When combined with crowdsourcing, benchmarking –that is, the evaluation of methods or models in well-defined conditions for the purpose of making standardized comparison– becomes a powerful approach to rapidly develop solutions that exceed the state of the art. At the most basic level, a performance benchmark requires a task, a metric, and a means of authenticating the result. In this modality of crowdsourcing, data is provided to participants along with the particular question to be addressed. Often, the organizers of such challenges will have "ground truth" or "Gold Standard" data that is known only to them and allows them to objectively score the methods that participants develop. Participants submit their solutions so that the organizers evaluate against the Gold Standard data. In this way, it is possible to find the best available method to solve the problem posed, and the participants can get an objective assessment of their methods. The organization of these efforts requires clear scoring metrics for evaluation of the solutions and availability of non-public datasets for use as Gold Standards. With the addition of a time constraint, participants are motivated to work concurrently, thus increasing the opportunities for collaboration and idea exchange.

Perhaps the first benchmarking challenge, posed to the many, but for which the solution is known only to a few, was used by Johann Bernoulli in the 17th century Herrera (1994) to compare methods to solve a mathematical problem. Johann Bernoulli had solved what today is known as the problem of the brachistochrone, consisting of determining the shape of the non vertical curve that would make the time taken by a bead sliding on it under the effect of gravity, reach the other extreme in the shortest time. This challenge was advertised in an article Bernoulli published in June 1696 using the words cited in the epigraph of this chapter. The incentive was then, as it is now, to make or break reputations, as indicated by Bernoulli when he states that through the solution of these challenges, solvers *"may [...] build for themselves eternal monuments among posterity."*. In the end, five solutions were submitted, including an anonymous one bearing an English postmark which was likely Newton's submission. So, who won this challenge? In some ways it was a tie, as all 5 solutions were correct. But Jakob Bernoulli's solution distinguishes itself as using methods that would eventually pave the way to the development of the calculus of variations.

Benchmarking challenges have played a key role in the evolution of many areas of AI and predictive modeling. The standard ingredients of benchmarking competitions include a stated problem, enrolled competitors, a public **dataset** on which competitors train their models, and a scoring methodology which assesses the accuracy of the predictions on a private, hidden test set. This methodology is known in some quarters as the Common Task Framework Liberman (2010); Donoho (2017)). In 1986 DARPA adopted the the Common Task Framework as a new paradigm to advance machine translation research. Over time, incremental improvements have accumulated to yield much of the technology that is taken for granted today, such as optical character recognition, instant automatic translation, dictation, and commanding computers by voice. The combination of this objective approach to benchmarking with predictive modeling culture has been described as the "secret sauce" of machine learning Donoho (2017).

The field of "protein folding" provides a sterling example of how benchmarking can nucleate communities and be a powerful driving force that lead to the solution of fundamental problems in science. Proteins are large, complex molecules that are essential to all life forms. Proteins are polymers composed of a linear chain of sub-units called aminoacids which interact with each other and with the surrounding water. These interactions make the polymer fold in 3D space until it reaches a stable minimum energy configuration. The function of proteins depend to a large degree on the 3D configuration it adopts. Figuring out the 3D shape of proteins is known as the "protein folding problem", a grand challenge in biology whose solution defied scientists for the past 50 years. CASP (Critical Assessment of protein Structure Prediction), a benchmarking challenge to assess the accuracy of protein structure prediction methods, has played a pivotal role in advancing

the field Kryshtafovych et al. (2021). Since its first edition in 1994, CASP has provided a biennial forum to benchmark the performance of structure prediction methods by comparing the predicted fold of proteins against their actual 3D structures known only by the organizers. In this way CASP can establish the state of the art in this field of science. Recently, CASP was at the center of a major scientific advance, by enabling the demonstration that the AI system called AlphaFold achieved an unprecedented accuracy in solving the protein structures proposed by the CASP organizers Jumper et al. (2021a). This breakthrough demonstrates the impact that benchmarking challenges can have in motivating the community and taking the pulse of the progress towards the solution of fundamental scientific problems.

The previous examples aim to illustrate how benchmarking challenges can 1) encourage the community to focus their efforts to solve important problems in science and technology by providing enticing incentives for participation, 2) provide a fair and objective comparison of the submitted solutions and help set the state of the art of a field, and 3) accelerate the pace of research by crowdsourcing a problem to a community of experts. These types of challenges, in which a competition is a means to a collaborative effort, are sometimes called collaborative competitions, critical assessments, or community experiments, and are the focus of this Chapter.

3 Collaborative competitions

A collaborative competition - often called a challenge or **coopetition**- is a specific form of crowdsourcing that has grown in popularity amongst research scientists during the past few decades. This kind of challenge leverages the use of leaderboards that allow participants to monitor their performance ranking with respect to others. Leaderboards also provide real-time feedback throughout a challenge to assess how their performance changes as their model evolves. Collaborative competitions also offer incentives, such as monetary prizes and/or the opportunity to co-author scientific publications that result from the challenge (See Chapter 13). General factors that lead to successful challenges include low entry barriers, continuous stimulation from the organizers, small intermediate milestones/rewards, an important problem with high economical and/or societal impact, and opportunities to get recognition or potentially a job offer.

It has been observed in many contexts such as those described above that there is a wisdom in the crowd Surowiecki (2005), a sort of emerging intelligence in which an aggregation of solutions proposed by teams, as long as they are working relatively independently, is often more accurate than any of the single solutions. This community wisdom gives real meaning to the notion of collaboration by competition.

Many collaborative competitions, covered in Chapters 7, 8, and 9, have emerged over the last few decades. These cover multiple areas of science and technology, ranging from fundamental to applied questions in machine learning (Kaggle Goldbloom and Hamner (2010), Codalab Pavao et al. (2022)), structural biology such as protein structure prediction (CASP Kryshtafovych et al. (2021) – Critical Assessment of protein Structure Prediction, CAPRI33), to asses algorithms developed to interpret human genetic variation (CAGI Hoskins et al. (2017) - the Critical Assessment of Genome Interpretation), to the assessment of data coming from a specific experimental technology (e.g. BioCreative for natural language processing Liu et al. (2019) or MICCAI Schnabel et al. (2019) for Medical Image Analysis). Collaborative competitions also provide a framework to evaluate software pipelines to process different data types, such as RGASP (RNA-seq Genome Annotation Assessment Project) that runs a challenge to evaluate software to align partial transcript reads to a reference genome sequence, a key step in RNA-seq data processing Engström et al. (2013). Other initiatives started with a narrow focus and then broadened their spectrum. For example the DREAM Challenges originally addressed the inference of gene regulatory networks from experimental data Stolovitzky et al. (2007), and hence the name DREAM: Dialogue for Reverse Engineering Assessment and Methods. However, over the years DREAM has evolved to address challenges ranging from regulatory genomics to translational medicine Saez-Rodriguez et al. (2016).

Recently, NeurIPS, the premiere conference in machine learning, has started challenges (NeurIPS Competitions) on a broad range of topics. These initiatives are often driven by academic efforts, although companies or other institutions and disease foundations (e.g. Prize4Life, Alzheimer's Research UK and the Arthritis Foundation) also run or support them.

4 The ingredients for a collaborative competition

Central to any crowdsourced challenge is the scientific question that the challenge intends to address. Not all questions are amenable to the challenge framework and not all questions have the potential to spark a community to action. The question must be fundamental to a field of research. Often times challenges are incentivized with academic publications (see Chapter 5), cash prizes and travel grants (see Chapter 13), as well as the interest of researchers to help to advance their field. Good challenge questions are conceptually straightforward and attract researchers from many fields of study who apply their specific principles and methods to address the question. For example, in the DREAM Network Inference Challenge Marbach et al. (2012), the task was to infer and benchmark transcription factor-to-target gene relationships. Because networks and graphic modeling are fundamental to many areas, researchers from Computational Biology, Engineering, Computer Science and Physics, to name a few, applied methods developed in their fields to address this challenge. Another example includes the Higgs Boson Machine Learning Challenge [1], which was timely and attracted the largest number of Kaggle competitors at that time. Participants were eager to learn how physics experiments to discover new particles were conducted and were motivated to contribute. All top ten ranking participants were not particle physicists and novel methods from oceanography and other fields were contributed. This trend, namely the convergence of experts in different disciplines to try to solve a timely problem, has been observed in other challenges. Inspiring new approaches to a fundamental question is central to crowdsourced challenges, thus relating science and technology questions to principles shared across many disciplines helps drive innovation.

From a practical perspective, a challenge must have enough data already generated to address the scientific question. Importantly, collaborative competitions multiply the impact of the crowdsourced **datasets** since dozens and sometimes hundreds of researchers will look at the data to extract insights. While much data is being generated in different areas of research and business, there still remain many interesting questions for which appropriate datasets are lacking. The underlying data must be not only sufficiently large but also of sufficient quality, diversity and complexity so that researchers can extract revealing patterns from the data. Additionally, the datasets must be of sufficient size so that robust statistical evaluation can be performed.

Crowdsourcing needs crowds: the more participation in a challenge, the higher the probability of solving it. However, and perhaps perversely, if hundreds of solutions are submitted, the statistical significance of the potential solutions degrades because of multiple hypothesis testing.

Fundamental to crowdsourced collaborative competitions is the unbiased, rigorous benchmarking of methods. This requires not only a sufficient amount and quality of data, but also a gold standard set of data to evaluate method performance. This gold standard must be inaccessible to challenge participants, which is an essential component of rigor in challenges. Also central to benchmarking is the metric used for scoring and evaluation. There are often many ways to evaluate method performance with each metric capturing a different aspect of performance. Selecting the proper metric must be considered in the context of the question being addressed. While the gold standard data and evaluation metrics establish rigor, unbiased evaluation is unique to crowdsourced challenges. In the typical research process, researchers develop methods and evaluate them on metrics and data that they select. Even unconsciously, this form of self-evaluation presents a biased method assessment, which can lead to what has been termed the "self-assessment trap." Norel et al. (2011) Crowdsourced challenges remove this bias in favor of a proper benchmarking of methods;

1. https://www.kaggle.com/c/higgs-boson

however, the results of a challenge must be interpreted in the context of the data, question, and scoring metric.

The rest of the book is devoted to understanding important aspects of contests, competitions, challenges and benchmarks. Chapter 2 provides practical advice on challenge design and Chapter 3 is devoted to details on dataset preparation. In Chapter 4 details on scoring metrics and challenge judging are discussed. Additionally, see Chapter 11 for a tutorial on how to create your own challenge or benchmark and Chapter 12 for challenges that have special designs, such as where teams submit an agent, instead of a discrete model, and the agent interacts with a simulator.

5 Challenge Organization

While it may seem simple at first sight, running challenges poses significant operational problems that require a coordinated effort from the organizers. Figure 1 shows the typical tasks involved in a challenge which require four layers of expertise: scientific, technical, legal and ethical. The genesis of a challenge could be the existence of a dataset with complex data whose analysis could benefit from the crowdsourcing paradigm Green et al. (2015); Abdallah et al. (2015); Atassi et al. (2014). More often challenges arise from scientific questions for which the answer requires new method development and validation Bansal et al. (2014), or from the need to benchmark algorithms that yield divergent results and for which an objective evaluation is needed Boutros et al. (2014).

The starting point for challenge organization is the definition of the scientific question for the challenge (see Chapter 2). This question must be of an applied or basic research nature, and to be impactful its importance has to be apparent. Addressing an impactful question is important to secure funding and to motivate challenge participation (see Chapter 13). Many challenges are hosted by organizations that focus on specific types of problems, such as DREAM for biomedicine, CASP for protein structure prediction, or ChaLearn Cha for Machine Learning. Typically these organizations decide through internal vetting process what challenges are organized. Other organizations that foster the organization of challenges, such as NeurIPS, makes yearly open calls for competitions Neu, and send the competition proposals for peer review to ensure that the challenge objectives are scientifically significant and of societal value. In all cases, the challenge question should be formulated in a way that it can be addressed in a collaborative competition setting, typically in the form of a predictive or classification algorithms. In some cases the question may simply be one of improving the state of the art in the performance of algorithms to solve a problem (e.g., protein structure prediction). In others the question could be a novel question for which no reference exists in a specific domain. In the latter case, competition organizers may involve coordination with a steering committee of experts in the domain area (technologists, physicians, economists, researchers, etc.). Additional considerations should also be had into account if the data are from sensitive sources or the results from the challenge could be used in an unethical manner.

The second step is assembling a team of data scientists, data governance and IT engineers to manage the data use agreements, data analysis tasks and IT infrastructure, or set up the challenge in the competition platform of choice (Chapter 10). It should be noted that data governance and establishing robust and reasonable data use agreements can be a time and resource consuming task, but it is essential for protection of sensitive data and to ensure its proper and ethical use.

Procurement and processing of the data is critical (see Chapter 3). It is essential that part of the data be unpublished so that it can be used as a gold standard dataset against which to score challenge submissions. This step usually requires Big Data processing such as normalization, compression, etc. If synthetic data is used for the challenge, appropriate tests are required to ensure it maintains privacy while maintaining fidelity to the original data. Having the data organized and packaged in easy-to-use datasets is necessary to reduce the barriers for participation. Adequate data governance must be in place, such as agreements with data producers, study participants, and Institutional Review Board (IRB) approvals for human data. Restrictions on the use of the data outside of the challenge should be clearly stated in the challenge rules.

Figure 1: The steps and tasks in the organization of a typical challenge.

Conducting an internal "dry-run" amongst the challenge organizing team or "mock participants" to test if the posed challenge question can be addressed with the data at hand is critical to determine if the challenge has sufficient data to solve the problem, but also that the challenge cannot be solved with trivial approaches. The dry run outcomes are (a) Dataset splits into balanced **training data**, **validation data** and **test data** sets, (b) Selection of an **evaluation metric**, (c) An estimate of the difficulty in answering the challenge questions with the data at hand (if a challenge seems impossible, then it may be better not to do it!), and (e) A definition of a baseline solution that the participants should improve upon, ideally using the present state-of-the-art approach. Further

documenting the solution with background information (e.g. a challenge overview and selected literature) and step-by-step guide to submissions as part of a "starter kit" is also extremely useful in reducing the participation barrier and increasing engagement.

A challenge needs an information technology infrastructure and web content. Important ingredients of such infrastructure are a registration system (with the requirement that participants agree to the challenge terms and conditions) Dyke et al. (2018), a challenge website containing a detailed challenge description, dataset storage and download and submission uploads, leaderboards for real-time feedback of performance, and a discussion forum where participants can communicate with organizers and other participants. The DREAM Challenges, for example, use Sage Bionetworks' Synapse platform `https://www.synapse.org`. Finally, in our era of Big Data, challenge platforms have started to use cloud services, for example Kaggle Notebooks `https://www.kaggle.com/code` where models written as code notebooks are developed and can directly access both public and private data. See Chapter 10 for a review of challenge platforms.

The next step is the definition of incentives to recruit as many participants as possible to solve the challenge. Incentives can include an invitation to the best performers to help draft and submit a challenge overview paper, a speaking invitation to conferences, monetary awards, and in some challenges a job offer. Often times, individual teams are encouraged to publish their methodology as a separate publication. In the academic community the most popular incentive is the prospect of co-authoring manuscripts, but many participants are motivated to participate in a collaborative effort where they can work on interesting and unpublished data to address an impactful problem. Chapter 5 discusses how to boost the impact of a challenge.

Before launching the challenge, it is desirable that a robust marketing and outreach campaign should be in place. This step is highly necessary as the success of the challenge depends on the degree of participation. Successful marketing approaches include the use of press releases, mailing lists, pre-challenge commentaries in top tier journals, interactive "warm up" or "kickoff" sessions, and direct outreach to researchers in the domain most directly connected to the challenge in question. See Chapter 13 for an in depth discussion on practical considerations in challenges, such as sponsors, grants, prizes, dissemination, and publicity.

After much preparation, the day arrives when the challenge is launched. The data is crowdsourced and solutions to the scientific problems posed in the challenge are solicited in the "open phase," also referred to as the "feedback phase" or "**development phase**." The open phase is characterized by the improvement of algorithms and methods using leaderboards to monitor progress, an open discussion of ideas and data features, using a Discussion Forum, and a deadline to submit the solutions. To submit their solutions participants can simply send a file with their list of predictions of the gold standard target values, or executables such as docker containers that will be run on withheld data for training and/or inference. By bringing the model to the data, the latter code-submission competitions overcome the difficulties of accessing private datasets, and increase the reproducibility of the results Ellrott et al. (2019). In the open phase, limits in the number of allowed submissions are needed to prevent learning the gold standard **test data** by trial and error. The number of submissions per day or total number of submissions may be limited. Submission limits may be circumvented by coordination between teams or creating multiple registrations per participant. To avoid this, such team coordination or creating fake teams with same participants but different team names must be explicitly prohibited in the challenge rules. Depending on the organization, there are different restrictions about the copyright and intellectual property rights associated with the submissions. Organizations that foster open and collaborative science typically require that participants submit open source code and an explanation of the methods used in the predictions to be accessed publicly.

When the open phase of the challenge finishes, the final assessment phase starts, in which submissions are evaluated to determine the best performers. The best submission from the open phase or a single final submission can be used to determine the final leaderboard metrics. Scoring

consists of comparing the submitted predictions of outcomes against the true outcomes using a held out gold standard, which is known to organizers but not to the participants. In order for a final score to be meaningful, it has to be accompanied with a statistical criterion of how difficult reaching that degree of performance is, typically under a null hypothesis. Several scoring metrics can be used to assess different aspects of the predictions Cokelaer et al. (2015).

After the final scoring phase ends, the opportunity to learn from the aggregate of submissions starts. In the DREAM Challenges, for example, organizers and interested challenge participants work together to analyze the results of the challenge and write a paper that describes it. Many challenge platforms organize a conference to discuss take home lessons, and present results. There are a few venues that focus on reporting challenges, such as the NeurIPS Competition Track and the Challenges in Machine Learning (ChaLearn) book series `http://www.chalearn.org/books.html`. See Chapter 5 for tips on writing a post-challenge paper.

Besides papers reporting on the overall challenge, the legacy of a challenge can be curating a database that archives and makes available for future use in education, research and benchmarking the concrete outcomes of the challenge. This database typically includes software code and documentation wikis of the participants and teams who provided a final submission. Example repositories include Sage Bionetworks Synapse `https://www.synapse.org`, UCI machine learning repository `https://archive.ics.uci.edu/ml/index.php`,Kaggle datasets `https://www.kaggle.com/datasets`, and Papers with Code `https://paperswithcode.com/`.

In this section we explored general considerations regarding the various steps and tasks necessary for organizing an effective challenge. Many of the subsequent Chapters in this book will delve deeper into the specific of these steps.

6 Challenge platforms

The success of the crowdsourcing paradigm has spurred a proliferation of challenge initiatives and platforms. Wikipedia lists close to 150 crowdsourcing projects in very diverse areas such as design and technology innovation. Figure 2 shows some of the most popular researcher driven challenge initiatives as well as the most popular commercial challenge platforms. The list is not intended to be comprehensive, as it omits some significant initiatives. However, it aims to highlight consistent and well-established challenge initiatives. Chapter 10 describes other popular challenge and competition platforms.

Among the researcher driven challenges, biomedicine is an exemplar of using challenges to support the significant changes in research needs. With the rapid increase in the amount of data, especially molecular information on genetic sequences, signaling pathways, protein structures, and medical imaging, public challenges are a strategy to develop and share novel methods. The topics that have profited the most from these types of efforts are structural biology (CAPRI, CASP), genomics (Sequence Squeeze, Assemblathon, CAMDA), systems biology (sbv-IMPROVER), text mining (BioCreative, CACAO, TREC Crowd), medicine (CLARITY), medical imaging (MICCAI), and emerging technologies in search of benchmarking and new analytical tools (RGASP, FlowCAP). Some challenge initiatives straddle more than one domain areas such as CAFA (Genetics/Genomics, Structural Biology), CAGI (genetics/Genomics, Systems Biology) and DREAM (Genetics/Genomics, Systems Biology, Emerging Technologies and Medicine). Initiatives such as DREAM, FlowCAP, CAGI and sbv-IMPROVER organize several challenges per year, and only the generic initiatives and not specific challenges are shown in Figure 2. As mentioned earlier, the case of structural biology and CASP is a paradigmatic example of how challenges have nurtured a field. For the last three decades, scientists tackled the key problems in this field in biannual challenges, followed by conferences where results were discussed. The field steadily progressed until in 2020, a deep learning model, AlphaFold, provided spectacular results Callaway (2020); Jumper et al. (2021b). While there are many open questions in the field, this was a major leap that built on the work of the community over the years. Machine learning research has similarly experienced

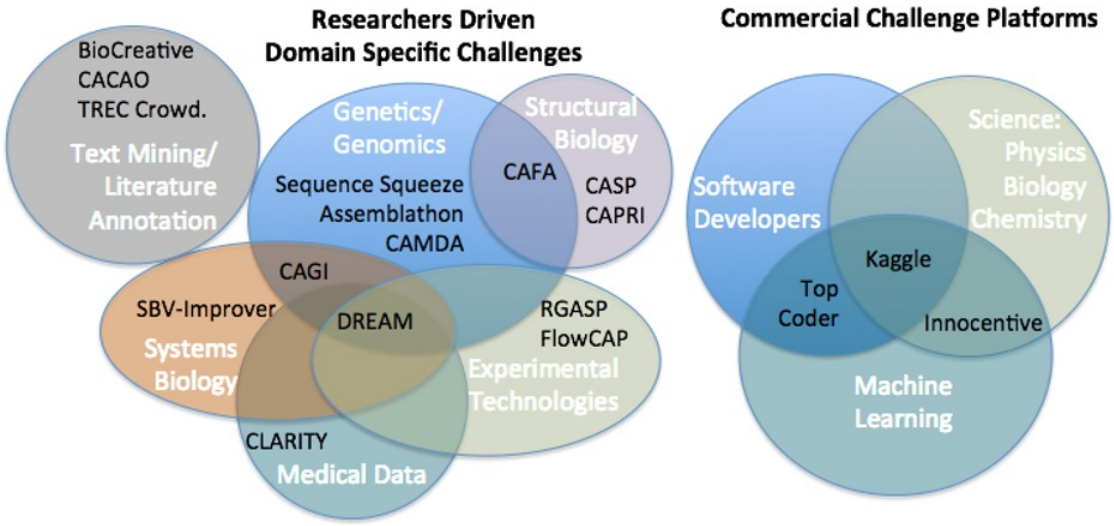


Figure 2: Non-exhaustive subset of challenge platforms and organizations in the biomedicine space (left), and some platforms that organize challenges in a variety of areas of science and technology (right).

a rapid expansion in method research, to support research in this area, the Conference on Neural Information Processing Systems (NeurIPS) has created a dedicated competition track since 2017 to highlight challenging problems and innovative solutions.

Crowdsourcing can be a profitable business. The business consists of organizing challenges as a fee for service for other companies or organizations that may not have the expertise necessary to give solutions to a specialized task. In such cases crowds can fill that expertise gap. Amongst the most popular and successful commercial challenge platforms we can mention Innocentive (`https://www.innocentive.com/`), which crowdsources challenges in science and technology (Social, Physics, Biology, Chemistry); TopCoder (`https://www.topcoder.com/`), which serves the software developer community; and Kaggle (`https://www.kaggle.com/`), that administers challenges to machine learning and computer savvy experts, addressing predictive analytics problems in a wide range of disciplines. Other open source competition platforms include Codalab Pavao et al. (2022) (`https://codalab.lisn.upsaclay.fr/`), EvalAI (`https://eval.ai/`) and Grand Challenge (`https://grand-challenge.org/`), and are discussed in Chapter 10.

7 Conclusions/perspective

Crowdsourced challenges offer a different way of doing science or solving problems collaboratively. This is not to say better than traditional approaches, but an alternative way to engage solvers and make valuable data available to the community. As discussed earlier in this Chapter, crowdsourcing is not a new idea, though when applied to science, it does cut against the grain of traditional, silo-ed academic research. Ultimately, we are at a point where the generation of data is outpacing our ability to make sense of this data. Team science has grown concomitantly with the

generation of these data because modern science and technology questions are often complex and require a multi-disciplinary approach. As the nature of data continues to change, both in richness and volume, analysis of these data is a continual difficulty we face. To advance science we must explore different modes of research to expand the frontiers of knowledge.

There are many opportunities for crowdsourced challenges to accelerate many aspects of academic education and research. In education, challenges in different fields can be used as learning modules to introduce computational methodologies and their rigorous evaluation. Students from high school to graduate level could develop their skills in ongoing challenges, while learning to collaborate with others to solve pressing problems in biomedicine, sustainability, environmental sciences, etc. Chapter 8 discusses the educational value of challenges and benchmarks in more detail. In research, the sheer amount of work that can be focused on one question in a short period of time is unmatched. As an illustration, a typical challenge that runs for a period of 5 months with 150 participants. Assuming that each researcher on average worked 100 hours on the challenge,represents about 15,000 hours of research effort dedicated to addressing one question; there are only 3,600 hours in 5 months. Even if an individual were to dedicate this amount of time to address a single question, it is unlikely that this individual would have the cross-disciplinary knowledge of 150 participants, thus a much smaller sampling of methods would be explored. One can imagine community efforts that both produce data and run a challenge to address a question in a time frame shorter than even the best funded research institutions can attain. If harnessed, this energy could potentially impart an extraordinary increase in the velocity and depth with which important problems are attacked. Chapters 6, 7 and 8 will provide overviews and special considerations for academic, industrial and educational challenges respectively.

Crowdsourced challenges produce rigorous and, if well run, unbiased benchmarked methods, with results representing the state-of-the-art in the field. (Chapter 9 further discusses benchmarks in machine learning.) Best performing methods produced in any given challenge have undergone a vetting process that cannot be done by any individual research group. This vetting process can be used to aid in academic publications, or a *challenge-assisted peer review. Successfully applied by the journal Science Translational Medicine in the Sage-BCC DREAM Challenge Margolin et al. (2013), this form of peer review is more rigorous than any individual reviewer can do.*

Data is being generated at an unprecedented rate and the number of potential crowdsourced challenges is increasing. This poses a hurdle in and of itself. The selection of challenges will become increasingly difficult with many worthy challenges competing for the attention of the research community and the research community growing more fatigued of these challenges. It remains an open question how to most effectively select questions for the community to address, with a potential solution being that the community itself crowdsources the question, that is, let the community decide which challenge to address.

Without participation, crowdsourced challenges would not exist, So as challenge organizers, we thank all the curious and ambitious solvers that have contributed to these community efforts. If the reader in interested in organizing a challenge or benchmark, Chapter 11 provides an excellent hands-on tutorial to get started. .

References

Nasa citizen science projects, http://www.chalearn.org. URL `http://www.chalearn.org`.

Nasa citizen science projects, https://science.nasa.gov/citizenscience. URL `https://science.nasa.gov/citizenscience`.

Neurips competition track, https://blog.neurips.cc/tag/competitions/. URL `https://blog.neurips.cc/tag/competitions/`.

Kald Abdallah, Charles Hugh-Jones, Thea Norman, Stephen Friend, and Gustavo Stolovitzky. The prostate cancer dream challenge: a community-wide effort to use open clinical trial data for the quantitative prediction of outcomes in metastatic prostate cancer. The oncologist*, 20(5):459, 2015.*

Johan OL Andreasson, Michael R Gotrik, Michelle J Wu, Hannah K Wayment-Steele, Wipapat Kladwang, Fernando Portela, Roger Wellington-Oguri, Eterna Participants, Rhiju Das, and William J Greenleaf. Crowdsourced rna design discovers diverse, reversible, efficient, self-contained molecular switches. Proceedings of the National Academy of Sciences*, 119(18):e2112979119, 2022.*

Sam Ansari, Jean Binder, Stephanie Boue, Anselmo Di Fabio, William Hayes, Julia Hoeng, Anita Iskandar, Robin Kleiman, Raquel Norel, Bruce O'neel, et al. On crowd-verification of biological networks. Bioinformatics and biology insights*, 7:BBI–S12932, 2013.*

Nazem Atassi, James Berry, Amy Shui, Neta Zach, Alexander Sherman, Ervin Sinani, Jason Walker, Igor Katsovskiy, David Schoenfeld, Merit Cudkowicz, et al. The pro-act database: design, initial analyses, and predictive features. Neurology*, 83(19):1719–1725, 2014.*

Mukesh Bansal, Jichen Yang, Charles Karan, Michael P Menden, James C Costello, Hao Tang, Guanghua Xiao, Yajuan Li, Jeffrey Allen, Rui Zhong, et al. A community computational challenge to predict the activity of pairs of compounds. Nature biotechnology*, 32(12):1213–1222, 2014.*

J Bernoulli. A new problem to whose solution mathematicians are invited. Acta Eruditorum*, 18:269.*

Kevin J Boudreau and Karim R Lakhani. Using the crowd as an innovation partner. Harvard business review*, 91(4):60–9, 2013.*

Paul C Boutros, Adam D Ewing, Kyle Ellrott, Thea C Norman, Kristen K Dang, Yin Hu, Michael R Kellen, Christine Suver, J Christopher Bare, Lincoln D Stein, et al. Global optimization of somatic variant identification in cancer genomes with a global community challenge. Nature genetics*, 46(4):318–319, 2014.*

Ewen Callaway. 'it will change everything': DeepMind's AI makes gigantic leap in solving protein structures. Nature*, 588(7837):203–204, November 2020. doi: 10.1038/d41586-020-03348-4. URL* https://doi.org/10.1038/d41586-020-03348-4*.*

David L Chandler. A doctor in the palm of your hand: how the qualcomm tricorder x-prize could help to revolutionize medical diagnosis. IEEE pulse*, 5(2):50–54, 2014.*

Thomas Cokelaer, Mukesh Bansal, Christopher Bare, Erhan Bilal, Brian M Bot, Elias Chaibub Neto, Federica Eduati, Alberto de la Fuente, Mehmet Gönen, Steven M Hill, et al. Dreamtools: a python package for scoring collaborative challenges. F1000Research*, 4, 2015.*

Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, et al. Predicting protein structures with a multiplayer online game. Nature*, 466(7307):756–760, 2010.*

Rhiju Das, Bin Qian, Srivatsan Raman, Robert Vernon, James Thompson, Philip Bradley, Sagar Khare, Michael D Tyka, Divya Bhat, Dylan Chivian, et al. Structure prediction for casp7 targets using extensive all-atom refinement with rosetta@ home. Proteins: Structure, Function, and Bioinformatics*, 69(S8):118–128, 2007.*

David Donoho. 50 years of data science. Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017. doi: 10.1080/10618600.2017.1384734. URL* https://doi.org/10.1080/10618600.2017.1384734*.*

SOM Dyke, M Linden, I Lappalainen, JR De Argila, K Carey, D Lloyd, JD Spalding, MN Cabili, G Kerry, J Foreman, T Cutts, M Shabani, LL Rodriguez, M Haeussler, B Walsh, X Jiang, S Wang, D Perrett, T Boughtwood, A Matern, AJ Brookes, M Cupak, M Fiume, R Pandya, I Tulchinsky, S Scollen, J Törnroos, S Das, AC Evans, BA Malin, S Beck, SE Brenner, T Nyrönen, N Blomberg, HV Firth, M Hurles, AA Philippakis, G Rätsch, M Brudno, KM Boycott, HL Rehm, M Baudis, ST Sherry, K Kato, BM Knoppers, D Baker, and P Flicek. Registered access: authorizing data access. Eur J Hum Genet.*, 26:1721–1731, 2018.*

Kyle Ellrott, Alex Buchanan, Allison Creason, Michael Mason, Thomas Schaffter, Bruce Hoff, James Eddy, John M Chilton, Thomas Yu, Joshua M Stuart, et al. Reproducible biomedical benchmarking in the cloud: lessons from crowd-sourced data challenges. Genome biology*, 20(1):1–9, 2019.*

Pär G Engström, Tamara Steijger, Botond Sipos, Gregory R Grant, André Kahles, Gunnar Rätsch, Nick Goldman, Tim J Hubbard, Jennifer Harrow, Roderic Guigó, et al. Systematic evaluation of spliced alignment programs for rna-seq data. Nature methods*, 10(12):1185–1191, 2013.*

Anthony Goldbloom and Ben Hamner. Kaggle. 2010. URL `https://www.kaggle.com/`*.*

Benjamin M Good and Andrew I Su. Games with a scientific purpose. Genome biology*, 12(12):1–3, 2011.*

Benjamin M Good and Andrew I Su. Crowdsourcing for bioinformatics. Bioinformatics*, 29(16):1925–1933, 2013.*

Joseph K Goodman, Cynthia E Cryder, and Amar Cheema. Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. Journal of Behavioral Decision Making*, 26(3):213–224, 2013.*

Angela K Green, Katherine E Reeder-Hayes, Robert W Corty, Ethan Basch, Mathew I Milowsky, Stacie B Dusetzina, Antonia V Bennett, and William A Wood. The project data sphere initiative: accelerating cancer research by sharing data. The oncologist*, 20(5):464, 2015.*

M Herrera. Galileo, bernoulli, leibniz and newton around the brachistochrone problem. Rev Mexicana Fis*, 40(3):459–475, 1994.*

Roger A Hoskins, Susanna Repo, Daniel Barsky, Gaia Andreoletti, John Moult, and Steven E. Brenner. Reports from cagi: The critical assessment of genome interpretation. Human Mutation*, 38(9):1039–1041, 2017. doi: https://doi.org/10.1002/humu.23290. URL* `https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.23290`*.*

Jeff Howe. The rise of crowdsourcing. Wired magazine*, 14(6):1–4, 2006.*

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Applying and improving alphafold at casp14. Proteins: Structure, Function, and Bioinformatics*, 89(12):1711–1721, 2021a.*

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. Nature*, 596(7873):583–589, 2021b.*

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

Martina Kutmon, Anders Riutta, Nuno Nunes, Kristina Hanspers, Egon L Willighagen, Anwesha Bohler, Jonathan Mélius, Andra Waagmeester, Sravanthi R Sinha, Ryan Miller, et al. Wikipathways: capturing the full diversity of pathway knowledge. Nucleic acids research*, 44(D1):D488–D494, 2016.*

Stefan M Larson, Christopher D Snow, Michael Shirts, and Vijay S Pande. Folding@ home and genome@ home: Using distributed computing to tackle previously intractable problems in computational biology. arXiv preprint arXiv:0901.0866*, 2009.*

Jeehyung Lee, Wipapat Kladwang, Minjae Lee, Daniel Cantu, Martin Azizyan, Hanjoo Kim, Alex Limpaecher, Snehal Gaikwad, Sungroh Yoon, Adrien Treuille, et al. Rna design rules from a massive open laboratory. Proceedings of the National Academy of Sciences*, 111(6):2122–2127, 2014.*

Mark Liberman. Fred Jelinek. Computational Linguistics*, 36(4):595–599, 12 2010. ISSN 0891-2017. doi: 10.1162/coli_a_00032. URL* https://doi.org/10.1162/coli_a_00032*.*

Sijia Liu, Yanshan Wang, and Hongfang Liu. Selected articles from the biocreative/ohnlp challenge 2018, 2019.

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. Nature methods*, 9(8):796–804, 2012.*

Adam A Margolin, Erhan Bilal, Erich Huang, Thea C Norman, Lars Ottestad, Brigham H Mecham, Ben Sauerwine, Michael R Kellen, Lara M Mangravite, Matthew D Furia, et al. Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. Science translational medicine*, 5(181):181re1–181re1, 2013.*

Jonathan M Mortensen, Evan P Minty, Michael Januszyk, Timothy E Sweeney, Alan L Rector, Natalya F Noy, and Mark A Musen. Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of snomed ct. Journal of the American Medical Informatics Association*, 22(3):640–648, 2015.*

Raquel Norel, John Jeremy Rice, and Gustavo Stolovitzky. The self-assessment trap: can we all be better than average? Molecular systems biology*, 7(1):537, 2011.*

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. CodaLab Competitions: An open source platform to organize scientific challenges. Technical report, Université Paris-Saclay, FRA., April 2022. URL https://hal.inria.fr/hal-03629462*.*

Martin Rees. A longitude prize for the twenty-first century. Nature News*, 509(7501):401, 2014.*

Julio Saez-Rodriguez, James C Costello, Stephen H Friend, Michael R Kellen, Lara Mangravite, Pablo Meyer, Thea Norman, and Gustavo Stolovitzky. Crowdsourcing biomedical research: leveraging communities as innovation engines. Nature Reviews Genetics*, 17(8):470–486, 2016.*

Julia A. Schnabel, Christos Davatzikos, Gabor Fichtinger, Alejandro F. Frangi, and Carlos Alberola-López. Special issue on miccai 2018. Medical Image Analysis*, 58:101560, 2019. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.101560. URL* https://www.sciencedirect.com/science/article/pii/S1361841519301021*.*

Dava Sobel. Longitude: The true story of a lone genius who solved the greatest scientific problem of his time*. Macmillan, 2005.*

Gustavo Stolovitzky, DON Monroe, and Andrea Califano. Dialogue on reverse-engineering assessment and methods: the dream of high-throughput pathway inference. Annals of the New York Academy of Sciences*, 1115(1):1–22, 2007.*

Devin P. Sullivan, Casper F. Winsnes, Lovisa Åkesson, Martin Hjelmare, Mikaela Wiking, Rutger Schutten, Linzi Campbell, Hjalti Leifsson, Scott Rhodes, Andie Nordgren, Kevin Smith, Bernard Revaz, Bergur Finnbogason, Attila Szantner, and Emma Lundberg. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. Nature Biotechnology*, 36(9):820–828, Oct 2018. ISSN 1546-1696. doi: 10.1038/nbt.4225. URL* https://doi.org/10.1038/nbt.4225*.*

James Surowiecki. The wisdom of crowds*. Anchor, 2005.*

Ines Thiele, Neil Swainston, Ronan MT Fleming, Andreas Hoppe, Swagatika Sahoo, Maike K Aurich, Hulda Haraldsdottir, Monica L Mo, Ottar Rolfsson, Miranda D Stobbe, et al. A community-driven global reconstruction of human metabolism. Nature biotechnology*, 31(5):419–425, 2013.*

Adrien Treuille and Rhiju Das. Scientific rigor through videogames. Trends in biochemical sciences*, 39(11): 507–509, 2014.*

Rohit Vashisht, Anupam Kumar Mondal, Akanksha Jain, Anup Shah, Priti Vishnoi, Priyanka Priyadarshini, Kausik Bhattacharyya, Harsha Rohira, Ashwini G Bhat, Anurag Passi, et al. Crowd sourcing a new paradigm for interactome driven drug target identification in mycobacterium tuberculosis. PloS one*, 7(7): e39808, 2012.*

Jérôme Waldispühl, Attila Szantner, Rob Knight, Sébastien Caisse, and Randy Pitchford. Leveling up citizen science. Nature Biotechnology*, 38(10):1124–1126, 2020.*

Journal of Data-centric Machine Learning Research (2025) Submitted 5/24; Revised 1/25; Published 5/25

Challenge design roadmap

Hugo Jair Escalante HUGO.JAIR@INAOEP.MX
Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico

Isabelle Guyon GUYON@CHALEARN.ORG
University Paris-Saclay, France, ChaLearn, USA, and Google, USA

Addison Howard ADDISONHOWARD@GOOGLE.COM
KAGGLE, USA

Walter Reade INVERSION@GOOGLE.COM
KAGGLE, USA

Sébastien Treguer STREGUER@GMAIL.COM
INRIA, France

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

Editor: Sebastian Schelter

Abstract

This document serves as a comprehensive guide for designing and organizing effective challenges, particularly within the domains of machine learning and artificial intelligence. It provides detailed guidelines on every phase of the process, from conception and execution to post-challenge analysis. Challenges function as motivational mechanisms that drive participants to address significant tasks. Consequently, organizers must establish rules that fulfill objectives beyond mere participant engagement. These objectives include solving real-world problems, advancing scientific or technical fields, facilitating discoveries, educating the public, providing platforms for skill development, and recruiting new talent. The creation of a challenge is analogous to product development; it requires enthusiasm, rigorous testing, and aims to attract participants. The process commences with a comprehensive plan, such as a challenge proposal submitted for peer review at an international conference. This document presents guidelines for developing such a robust challenge plan, ensuring it is both engaging and impactful.

Keywords: Challenge design, organizer guidelines, challenge proposal

1 Before you start

This section delineates the essential inquiries that challenge organizers must address prior to initiating the process of challenge organization. Early consideration of these questions assists organizers in accurately estimating the resources required to achieve their objectives and in enhancing the preparation of their proposals. This document primarily focuses on data-driven challenges evaluated using quantitative objective metrics, with participants ranked on a leaderboard. Nevertheless, the methodology is also broadly applicable to jury-evaluated competitions and to benchmarks

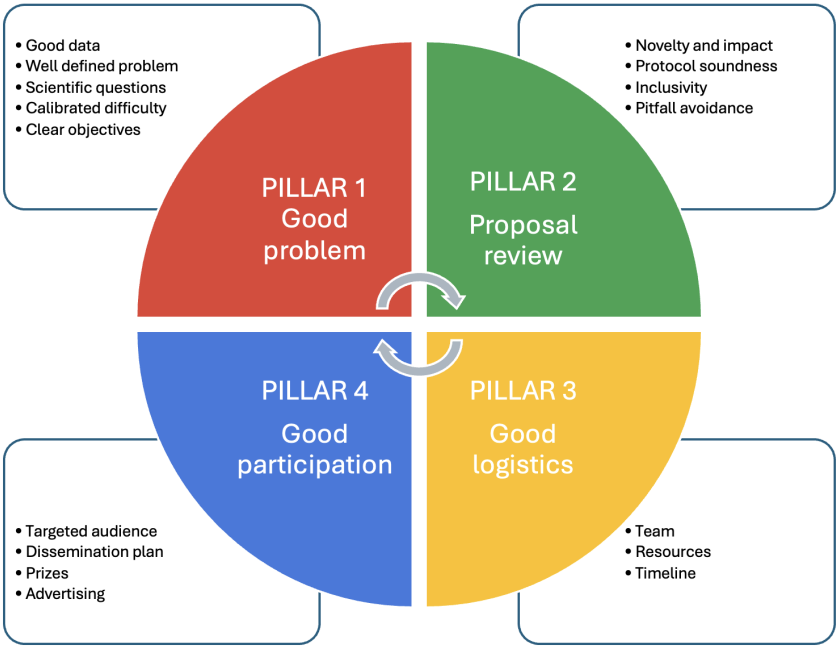


Figure 1: Challenge design principal pillars.

The material presented in the subsequent sections draws[1] on the preparation guidelines from organizations such as Kaggle[2], ChaLearn[3] and Tailor[4], as well as the NeurIPS proposal template, to which some of the authors contributed. This document constitutes Chapter 2 of the book "AI competitions and benchmarks: the science behind the contests[5]" (under preparation). The essential components of a successful challenge are summarized in Figure 1. Although the organization of this document does not strictly adhere to the structure of the figure, we will repeatedly refer to these essential ingredients.

PILLAR 1: Good problem

This section highlights key prerequisites prospective challenge organizers must address. First of all, do you have a well defined challenge problem to be solved? Maybe not yet! You may just be interested in becoming a challenge organizer. In that case, we recommend that you **partner** with researchers, industrial or non-profit organizations who have data and an interesting problem to address. Even with a strong partner, this section may alert you to critical issues requiring attention.

1. Please note that while we have tried to provide context and justify every statement and recommendation of this document, part of the material presented in this document is based on the experience of authors who have dedicated a considerable amount of time to all aspects of challenge organization.
2. `http://www.kaggle.com/`
3. `http://www.chalearn.org/`
4. `https://tailor-network.eu/`
5. `https://sites.google.com/chalearn.org/book/home`

Do you have (enough) good data?

Data-driven challenges rely heavily on the availability of good data. Popular datasets can be beneficial for benchmarking purpose as they allow researchers to compare their results against a substantial body of prior work, providing context and establishing a standard for future research (Donoho, 2017). A dataset that has been used by many researchers in the past can be re-purposed to answer specific research questions. In Table 1 we list some **data sources** that can inspire you. However, using publicly available widely-used datasets carries a risk: competitors or the base models they use (such as pre-trained deep-net backbones) may have prior exposure to such data, which could bias evaluation or offer an unfair advantage.

Therefore, it is advisable to source datasets that have not yet been extensively explored by machine learning researchers. For further guidance on selecting, collecting, and preparing data, refer to document 3 of this book for a detailed description of the dataset development cycle (Egele et al., 2024). Additionally, Appendix C elaborates on data leakage, a major issue with data driven competitions.

The NeurIPS datasets and benchmarks track is also a growing resource of well-reviewed datasets. Using this resource type is particularly convenient prior to its publication. Otherwise, the dataset will have the public exposure, and means for data obfuscation or detecting prior exposure of models would be necessary. Please note that a considerable amount of datasets in this track are intended to become benchmarks (i.e., authors release everything you need to evaluate and compare models). Hence, this type of resource may not align well with certain challenge protocols, see Table 2.

When assessing potential data, organizers must consider numerous facets of **quality data**. Does the dataset contain biases (Ntoutsi et al., 2020) that would lead to an unacceptably biased model? Are there data leakages[6] that would spoil the objective of the challenge (Kaufman et al., 2012)? If the plan is to recycled or re-purpose data, do you have detailed information about the original intent (Koch et al., 2021)? Can you guarantee that the **test data is "fresh"**, i.e., that none of the participants had prior access to these data? If public re-purposed data is used for a challenge, is it **obfuscated enough** so that they are not recognisable to the participants, and/or hidden to the participants at test time? In addition to these considerations, it is critical that organizers ensure they have the **right to use the data and/or code** used in the challenge!

Do you have a problem lending itself to a challenge?

Having data is necessary, but not sufficient to organize a challenge. Do you have a **good definition of your problem** and have you tried yourself to solve it with some **simple baseline method**? Do you have a sense of how hard it is (it should **neither be trivial nor too hard to solve**)? If not, it is premature to organize a challenge with that problem, you need first to get familiar with this problem and be able to define criteria of success (called "metrics"), and have some preliminary ideas on how to optimize them. Make sure you understand how to **cast your problem into AI tasks**, which may range from machine learning tasks (binary classification, multi-class or multi-labels classification, regression, recommendation, policy optimization, etc.) (Burkov, 2019), optimization tasks (continuous, combinatorial or mixed; single- or multi-objective, etc.) (Aggarwal, 2020 - 2020), reasoning (logic or probabilistic), planning, constraint programming, ... (Russell and Norvig, 2010), or a combination of several types of tasks. Having participated yourself to a challenge may be helpful. In

6. https://www.kaggle.com/docs/challenges#leakage

Domain	Data source	Data type
Machine Learning	Kaggle datasets	The largest general-purpose ML dataset repository with >170K datasets in various formats, but generally coming with illustrative Jupyter notebooks.
CodaLab	CodaLab	A large repository with data from hundreds of challenges, mostly academic (Pavao et al., 2022).
Machine Learning	Hugging Face	More than 1000 datasets with well defined format/metadata and data loaders. Moatly for Audio, CV, and NLP modalities.
Machine Learning	OpenML	More than 5000 datasets, mostly in tabular format.
Machine Learning	UCI ML Repository	Historcal repository created in 1987 by D. Aha and his students. Hundreds of datasets, mostly small and in tabular format.
Reinforcement Learning	Farama gymnasium	New version of OpenAI gym. It includes a variety of environments, such as classic control problems and Atari games.
Miscellaneous	Data.gov	Over 200,000 datasets from the US government. The datasets cover a wide range of topics, from climate to crime.
Sensor Data	NOAA OpenSignal Array of Things)	Sensor data from a variety of sources, such as IoT devices, wearables, and environmental sensors, can provide rich information about the physical world. This data can be used to develop models for a wide range of applications, such as health monitoring, environmental sensing, and predictive maintenance.
Audio data	SpRUce Million Song VoxCeleb	Audio data, such as speech and music, is a rich source of information that can be used for a variety of applications, such as speech recognition, speaker identification, and music recommendation.
Textual data	Common Crawl Reuters News Amazon Reviews	Textual data, such as news articles, social media posts, and customer reviews, is a rich source of information that can be used for a variety of applications, such as sentiment analysis, topic modeling, and natural language understanding.
Satellite imagery	Planet Labs European Space Agency NASA	Satellite imagery can provide high-resolution images of the Earth's surface, which can be used for applications such as land use classification, urban planning, and disaster response.
Financial data	Quandl Yahoo Finance FRED	Financial data, such as stock prices, market trends, and economic indicators, can be used to develop models for predicting stock prices, identifying trading opportunities, and understanding economic trends.
Neuroscience Data	Open Neuro	A free and open platform for validating and sharing BIDS-compliant MRI PET MEG EEG iEEG

Table 1: **Sources of data.** There is no good challenge without good data. An important aspect is to find "fresh data" to reduce the risk that the participants have been exposed to the challenge data previously, and can have an unfair advantage. Datasets commonly used in ML research (such as those colored in cyan) can be used as illustrative examples in the starting kit. Public datasets (such as those colored in yellow) can be used for the development phase (public leaderboard), but it is preferable to use novel fresh data for the final phase (private leaderboard).

that respect, the Kaggle book provides a gentle introduction to data challenges (Banachewicz and Massaron, 2022).

Additionally, not all problems lend themselves to a challenge. Importantly, challenges are "games of skills", NOT "games of chance". Can you devise a quantitative way of evaluating participants (using your metrics) in which **chance plays no role** (this is a legal requirement to organize scientific contest in most countries, to avoid that they fall under gambling regulations)? This may be particularly difficult. If the evaluation of participants entries relies on a statistic computed from data (typically some test set performance score using your metric), do you have **enough data** to obtain small error bars or a stable ranking of participants? In (Guyon et al., 1998), the authors provide a rule-of-thumb for classification problems. Usually, tens or thousands of examples must be reserved for test data, if you want good error bars; alternatively, you can use multiple datasets. Remember that, if the participants are allowed to make multiple submissions to the challenge and get feed-back on their performance on the test set on a leaderboard, they will soon overfit the leaderboard. In statistics, this is known as the problem of multiple testing. Leaderboard overfitting can be alleviated, to some extent, with a technique called "The ladder" (Blum and Hardt, 2015), which essentially boils down to quantizing the test set performance. However, to really prevent leaderboard overfitting, it is advisable to run the challenge in **multiple phases**: during a development phase, the participants are allowed to make multiple submissions and view their performances on a "public leaderboard"; during the final phase, the winners are chosen on the basis of a single submission, evaluated on a separate fresh test set. The performances are kept secret on a so-called "private leaderboard", invisible to the participants until the challenge is over. Since challenges are run in this way, leaderboard overfitting seems to have largely disappeared (Roelofs et al., 2019).

Be mindful that the **metric really reflects the objective you want to optimize**: a common mistake, for instance, is to use the error rate for classification problems, not distinguishing between the cost for false positive and false negative errors (e.g., it may be more detrimental to send a sick patient back home than to ask a sound patient to do one more medical exam). Also, would you rather your model provide you with a definitive prediction or a range of confidences? Finally, will you be declaring ties if performances between two participants are too close? See document 4 of this book for a detailed treatment on judging competitions, comprising critical aspects like the matching of metrics with objectives, statistical analysis evaluation, and fusion of multiple scores among other relevant topics.

What are your scientific questions?

What is the main problem[7] we want to address and would like to be solved? Asking the good questions is key to get results inline with the initial goals. What are the objectives of the challenge? Is our priority to address scientific questions, and which ones precisely, or to get as outcomes models easy to transfer to a production system with all its constraints in terms of robustness, explainability, performance monitoring, maintenance? Is the only objective, the final accuracy at the end of training without constraints on resources: compute, memory and/or time? Or should the participants also

7. Defining a problem and formulating scientific questions are closely related topics. However, they differ in their objectives and focus. When defining a problem one makes emphasis on finding well-defined tasks that can be solved by using machine learning, with clear goals and evaluation metrics. In contrast, scientific questions are broader and aim to advance the fundamental understanding of machine learning concepts, theories, and methodologies. They prioritize exploration, analysis, and theoretical contributions. While challenges focus on application and measurable outcomes, scientific questions emphasize knowledge generation and methodological innovation.

take into account limits in training time, compute power, memory size, and more, with the goal to find the sweet-spots for good trade-offs?

The definition of each task to achieve must help to solve a specific question raised by the challenge, but must also carefully take into account all constraints and reflections previously mentioned. Then what are the constraints in terms of data: volume, balance or unbalance of classes, fairness, privacy, external *vs.* internal, etc.? These questions are related to the tasks that are themselves related to the initial questions to be addressed. For each scientific question, you will then need to define some metrics allowing you to measure how well each participant answers the question: more details in Section 2 (*What metric for what purpose?*) of Document 4 of this book.

In general, AI challenges should have very specific objectives. There is a natural human tendency, when expending significant time and resources collecting and preparing data for a challenge, to want to answer as many questions as possible from the challenge. This is almost always counterproductive. While there may be considerable secondary information that can be gleaned at the conclusion of a challenge, challenges should be designed to have a very specific primary question to be addressed. This primary question is commonly in the form of the maximum predictive performance a model can extract from a given dataset.

Machine learning challenges often aim to address a multitude of scientific questions. These questions can be categorized using a taxonomy based on their overarching 5W themes: what, why, how, whether, and what for. Here's a potential breakdown:

1. **What (Discovery):** What patterns can be discovered in data? What features are more significant or relevant to the target variable? What groups or segments naturally form in the dataset? What are the characteristics of these clusters? An example of a discovery challenge would be the Higgs Boson challenge, that aimed at discovering a new high energy particle (Adam-Bourdarios et al., 2015).

2. **Why (Causality):** Why did a specific event or outcome occur? Are there variables that directly influence this outcome? Why does a certain data point deviate from the norm? For example, ChaLearn organized several causality challenges, including the Causation and Prediction challenge (Guyon et al., 2008) and the Cause-effect pair challenge (Guyon et al., 2019a).

3. **How (Prescriptive):** How can we allocate resources efficiently to achieve a goal? How can an agent take actions in an environment to maximize some notion of cumulative reward? For example the Black box optimization challenge asked participants to figure out how to optimize hyperparameters of models in a black box setting. The Learning to Run a Power Network (L2RPN) challenge series is asking how an agent can control the power grid to transport electricity efficiently, while maintaining equipment safe (Marot et al., 2020a,b, 2021).

4. **Whether (Comparative):** Whether Algorithm A is better than Algorithm B for a specific task? Whether a given preprocessing or hyperparameter setting X improves the model's performance over technique Y? Whether there is a trade-off e.g., between performance metrics (like precision and recall). Whether a model trained on dataset A performs better on dataset B compared to a model directly trained on dataset B (a transfer learning problem)? Whether a certain RL method performs better in environment condition X compared to condition Y?" For example, the Agnostic Learning vs. Prior Knowlege (AlvsPK) challenge answers the

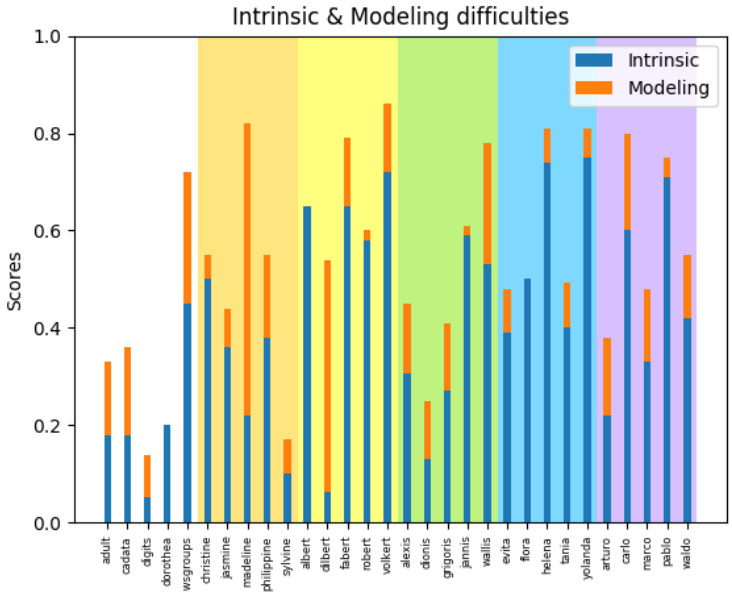


Figure 2: Example of intrinsic and modeling difficulty of datasets.

question whether prior knowledge is useful to devise a good preprocessing or whether using agnostic features is enough (Guyon et al., 2007).

5. **What For (Purpose-driven):** For whom might this model be unfair or biased? For what populations does this model perform sub-optimally? For what new tasks or domains can knowledge from one domain be useful? For example, the Jigsaw Unintended Bias in Toxicity Classification asked predicting and understanding toxicity in comments while considering potential biases against certain populations (Cjadams et al., 2019).

This taxonomy not only provides a structured way to think about scientific questions in machine learning but also helps in deciding the kind of algorithms, data processing techniques, and validation strategies that might be best suited to address them.

Have you calibrated the difficulty of the challenge?

Challenges that are overly difficult or too simple fail to advance the field. A critical aspect of challenge design is calibrating its difficulty, which may involve data engineering, metric refinement, and benchmarking tasks against established baseline methods.

Figure 2, taken from the AutoML challenge (Guyon et al., 2019b), shows how the difficulty of datasets might be calibrated. Test set scores are represented (normalized between 0 and 1, 0 is the level of random guessing). The height of the blue bar represents the score of the best model (our best estimated of the irreducible error of "intrinsic difficulty"). The height of the orange bar represents the range of scores between the best and the worst models, which we use to evaluate the "modeling difficulty". The datasets that are most suitable to separate methods well are those with small "intrinsic difficulty" and large "modeling difficulty". During beta-testing, you may want

to set up an "inverted challenge" among organizers, in which datasets are submitted against established baseline methods, then select those datasets with the largest ratio of modeling difficulty over intrinsic difficulty. If after adjusting the task difficulty, there is still little participation during the challenge, be ready to **lower the barrier of entry**, but providing more clues to get started, code snippets, notebooks, and/or tutorials.

Do you have clear objectives?

Are you more interested in finding a champion, benchmarking algorithms to evaluate or (incrementally) push the state of the art, or discovering brand new methods? Or are you simply interested in making your company/institution visible? While these three goals may be not mutually exclusive, your challenge design should take them into account.

In **recruiting challenges**, your goal is to find a **champion**. You may want to select a representative problem of what it is like to work at your company or institution to find top talents to employ. You will NOT need to put in effort in preprocessing data, designing a good API, etc.: the participants will be expected to do all the dirty work and show they **excel at solving all aspects of the problem**. Make part of the challenge deliverable that top ranking participants must deliver a technical report on their work to best evaluate them.

In **Research and Development challenges**, your goal is to benchmark algorithms. You may want to carefully design your API and have participants supply an object or a function, which addresses the **specific sub-task you have identified as a bottleneck of your problem**. Make sure to sort out licensing conditions of the winner's solutions. One simple way is to ask winners to open-source their code as a condition of being eligible for prizes.

In **Academic challenges**, your goal is to discover new methods for a problem that you largely do not know how to solve. You may want to have both quantitative and qualitative evaluations, e.g., in the form of a **best paper award**. This is because it is not obvious when you invent a new method to optimize it and get it to outperform others quantitatively, this may involve tuning and engineering.

In **Public Relation challenges**, your goal is to make your company or institution known, e.g., to attract new customers or students, and to expose your specific data or problem of interest to this public. It is essential to keep the challenge as simple as possible (at least in its problem statement) and as didactic as possible, and build around it great communication with the public, including using mass media. You may want to have intermediate goals and prizes and organize interviews of participants making good progress, to boost participation.

In **Branding challenges**, your goal is to put your name in front of a large community of data science practitioners by releasing a new technology or dataset (e.g., incentivizing the use of Tensorflow, or releasing a SOTA image classification set like ImageNet).

In **Hackathons** your goal is to provide innovative solutions to specific problems in a very short period of time. Participants are expected to work intensively and collaboratively around a theme, commonly aiming at open-ended innovation. You may want to dedicate extra effort in planning logistics (e.g., in terms of technical infrastructure and ensuring comfort for participants, providing mentorship and support, and even ensuring an enjoyable experience for participants).

PILLAR 2: Proposal review

Even experienced challenge organizers can make mistakes. Conferences with competition programs offer valuable opportunities to recruit participants and obtain feedback on your proposal.

We strongly recommend submitting your proposal to such venues. Other review instances may be required, such as ethical review boards, if human subjects are involved. Section 2 provides a proposal template, while this section highlights key questions to address for a strong and successful submission.

Is your problem novel and impactful?

Organizing a challenge entails significant responsibilities, as it engages the time and resources of both organizers and participants and may involve ethical considerations. Above all, reviewers will focus on the challenge's novelty and its potential impact (both positive and negative).

Ask yourself:

- whether other challenges or benchmarks have addressed the same problem before and whether the results or your new challenge will be **incremental or ground breaking**;
- whether you can illustrate your problem in one or several **domains of interest to the public**, which may include medicine, environmental science, economy, sociology, transport, arts, education;
- whether the outcome of the challenge will have a **practical societal or economical impact**;
- whether the approached task (or any aspect associated to it, e.g., application, domain, scenario) represents a potential hazard to users' rights, including **ethical issues**, e.g., linked to experimenting with human subjects, privacy concerns, etc.;
- how you can lower the barrier of entry to increase participation; this may mean e.g., using a uniform data format, providing data readers, providing sample code.

Have you a well thought of protocol?

After defining a challenge problem with robust data and metrics, numerous protocol choices remain. Will you use single or multiple metrics? Will these metrics span single or multiple phases or tracks? Will participants submit results or code? How much information will they have? What resources will organizers provide? Will participants face resource or time constraints?

Have you chosen your type of challenge protocol? Table 2 illustrates a hierarchy of challenge protocols for supervised learning tasks (Liu, 2021). As one progresses from one level to the next, participants have access to less information, whereas their submissions receive more. Level λ is associated with challenges requiring result submissions, whereas the higher levels pertain to code submissions, from a class named ALGO. Level α is based on the premise of submitting pre-trained models. Level β depicts a scenario of thorough code blind testing, where both training and testing occur on the platform. If multiple datasets are accessible, meta-testing can be conducted, given that participants are provided with examples of analogous tasks for meta-training. In the final stage, the γ level, participants aren't provided with the meta-training set. Instead, they receive a basic task description. Consequently, from level λ to γ, the submissions evolve to be increasingly autonomous, aligning more closely with genuine automated machine learning (AutoML). However, this also necessitates increased computational resources to implement the challenge.

Have you managed to reduce your challenge to a single clear objective that will give it focus? It is tempting, but generally unwise, to try to combine different metrics into a single objective.

Table 2: Hierarchy of challenge protocols.

Level	Information available to participants	Information available to the algorithm only	Type of submission
λ	Everything, except test labels	Nothing	`RESULTS` of test set predictions
α	Labeled TRAINING set	Unlabeled TEST set	`ALGO.predict()`
β	META-TRAINING set	META-TEST set (each test set wo test labels)	`ALGO.fit()`, `ALGO.predict()`
γ	Nothing, except starting kit and sample data	META-TRAINING set & META-TEST set (each test set wo test labels)	`ALGO.meta-fit()`, `ALGO.fit()`, `ALGO.predict()`

When disparate scientific questions are force-fit into a single evaluation metric, the unintended consequences may be that the setup favors optimization of one question over the others, or that all the questions sub-optimized to optimize the combined metric. However, there are scenarios where multiple objectives genuinely arise, necessitating consideration of several metrics. For example, the metrics might include both accuracy and speed or memory footprint. In such cases, it is essential to strike a balance, ensuring that all objectives are addressed without overly compromising on any single aspect. Properly defined and carefully weighted metrics can help ensure that all objectives are optimized without undue sacrifices.

When facing multiple primary questions, it can be beneficial to introduce separate challenge tracks. For instance, the M5 forecasting challenge featured a track for precise point forecasts[8] and another for estimating the uncertainty distribution of the values[9]. Despite using the same dataset, the former adopted the weighted root mean squared scaled error, while the latter employed the weighted scaled pinball loss as their respective metrics. By creating distinct tracks, the challenge enabled researchers to advance the state-of-the-art in each area, rather than settling for methods that performed adequately across both but didn't excel in either. The choice boils down to whether the aim is to nurture specialist or generalist algorithms. To motivate participants to engage in multiple tracks while still promoting comprehensive methods, an incentivized prize structure can be adopted. For example, winners might receive a reward of x for one track, double that amount for two, and exponentially more, as in $2^n x$, for triumphing in n tracks.

There is a challenge format that does allow for more open-ended discovery of a dataset, which we'll refer to as an analytics challenge. The idea here is to provide data, give general guidance on the objective, and allow the competitors to analyze the data for new insights. An example of this was the NFL Punt Analytics challenge[10], where the goal was to analyze NFL game data and suggest rules to improve player safety during punt plays. While the scientific question was specific ("what rule changes would improve safety?"), the format allowed for a much broader exploration of the data and provided insights that wouldn't have surfaced by optimizing a single metric. While this can be beneficial, Analytics challenges tend to have much lower participation than predictive challenges, and require a significant amount of work after the challenge deadline to review and manually score (using a predefined rubric) each of the submissions from the teams.

Useful guidelines have been provided in Hutter (2019): "*The typical setup in machine learning challenges is to provide one or more datasets and a performance metric, leaving it entirely up to*

8. `https://www.kaggle.com/competitions/m5-forecasting-accuracy`
9. `https://www.kaggle.com/competitions/m5-forecasting-uncertainty`
10. `https://www.kaggle.com/competitions/NFL-Punt-Analytics-challenge`

participants which approach to use, how to engineer better features, whether and how to pretrain models on related data, how to tune hyperparameters, how to combine multiple models in an ensemble, etc. The fact that work on each of these components often leads to substantial improvements has several consequences: (1) amongst several skilled teams, the one with the most manpower and engineering drive often wins; (2) it is often unclear **why** *one entry performs better than another one; and (3) scientific insights remain limited. Based on my experience in both participating in several challenges and also organizing some, I will propose a new challenge design that instead emphasizes scientific insight by dividing the various ways in which teams could improve performance into (largely orthogonal) modular components, each of which defines its own challenge. E.g., one could run a challenge focusing only on effective hyperparameter tuning of a given pipeline (across private datasets). With the same code base and datasets, one could likewise run a challenge focusing only on finding better neural architectures, or only better preprocessing methods, or only a better training pipeline, or only better pre-training methods, etc. One could also run multiple of these challenges in parallel, hot-swapping better components found in one challenge into the other challenges. I will argue that the result would likely be substantially more valuable in terms of scientific insights than traditional challenges and may even lead to better final performance.*"

Will your challenge be inclusive and favor open science?

Inclusivity and open science are often overlooked but should be prioritized early in challenge design. In fact, competition tracks in several conferences require that organizers justify the potential impact (open science, economical, societal, humanitarian, etc.) of challenges, see Section 2. It is particularly important to consider the following aspects:

- **Encourage Open-Source Collaboration:** Highlight strategies to engage the open-source community, such as making challenge datasets and code repositories publicly available, fostering transparency, and encouraging contributions that improve the challenge infrastructure.
- **Support Underrepresented Researchers:** Offer specific guidelines or incentives to increase accessibility for underrepresented groups, such as cloud computing time, reduced registration fees in associate events, mentorship programs, or targeted outreach to underrepresented communities.
- **Promote Inclusive Participation:** Explore methods to create a more inclusive environment, like remote participation options, multilingual resources, and tailored support mechanisms to ensure broader and equitable engagement.

Have you reviewed common pitfalls?

When everything appears ready, review this final checklist to ensure soundness.

1. **Lack of clarity.** Perhaps the most common mistake, is to offer a challenge that has a lack of clarity in the problem definition and the goals to be reached, a **too complex metric** defying intuition, or a **lack of focus** by addressing too many problems at once. When designing an AI challenge, it is important to understand that there is no way to optimize for all of the questions you might want to answer. It is better to put in the hard work up front to decide what the specific primary question should be and how to measure success with a single simple

metric. If there are secondary questions that you would like addressed, these should only be considered if they can be answered without jeopardizing the primary question. Clarity in the **rules** is also important. If you need a long list of rules for legal reasons, summarize them. Do not forget to have the participants accept the rules at the time of registration. See ChaLean contest rules, for inspiration. Add a FAQ page answering most frequently asked questions.

2. **Fatal flaws.** Another pitfall is to discourage serious competitors to enter because of **obvious flaws** (in data or in challenge rules). **Beta-test** your challenge thoroughly using volunteers (who will then not be eligible to enter the challenge) or solicit people you know well to enter the challenge as soon as it opens and report possible flaws. If possible, make a **"dry run"** or organize first a scaled-down version of your challenge to test its protocol, before you launch the "big one". Such trials will also allow you to **calibrate the difficulty of the task** to the target audience.

3. **Needless constraints.** Another way of discouraging participation is to put too many constraints or prerequisites to enter the challenge: having attended a previous challenge or event, registering with a nominative account, attending a conference, open-sourcing code, etc. While more constraints can be placed on the winners (or the entrants to the final phase), it is advisable to facilitate as much as possible entering the feed-back phase.

4. **Inconclusive results.** The success of the challenge also rests on having **quality data**, unknown yet to the public. We mentioned the problem of **bias in data** or **data leakage** in the introduction. Appendix C provides guidance on how to avoid falling into the most common traps when preparing data. In addition to having quality data, you must also have **sufficiently large datasets**. A common mistake is to reserve a fixed fraction of the dataset for training and testing (typically 10% of the data for testing), without anticipating the error bars. A simple rule-of-thumb to obtain at least one significant digit in a 1-sigma error bar, for classification problems, is to reserve at least $N = 100/E$ test examples, where E is the anticipated error rate of the best classifier (the challenge winner) (Guyon et al., 1998). So, if you anticipate $E = 10\%$, $N = 1000$, if you anticipate $E = 1\%$, $N = 10000$.

PILLAR 3: Good logistics

Organizing a challenge is akin to launching a product, requiring **managerial skills** in addition to technical expertise.

Are you sufficiently qualified to organize your challenge?

There are many difficult aspects to tackle in the organization of a challenge and this can be daunting. It is rare that a single person is qualified to address them all. Think of **partnering with other experts** if you do not know how to answer any of the questions of the previous sections (or the following ones).

For instance, depending on the data types or modalities (tabular, univariate or multivariate time-series, image, video, text, speech, graph) and application-dependent considerations, **appropriate evaluation metrics** should be chosen to assess performance of the submissions. It may make a lot of difference if one chooses accuracy rather than balanced accuracy or AUC if classes are imbalanced, for instance. You may know that and know the difference between MSE and MAE for regression,

but do you know what SSIM, SHIFT, SURF are for image similarity? Do you know what F1 score is and what the difference is between micro-averaging and macro-averaging? Have you thought about whether you rather evaluate best objective value or time to reach given precision, for optimization tasks? Success or failure, time-to-solution for reasoning tasks? or do you need qualitative metrics possibly implying human evaluation (i.e., by some expert committee)? For details, see document 4 of this book.

Also, regarding data modalities, the choice of data and the evaluation of its quality require a lot of expertise. We have mentioned already the problem of data leakage, which leads to inadvertent disclosure of information about the "solution" of the challenge. Document 3 of this book reviews many more aspects of data that require attention, including legal aspects of ownership and attribution, privacy, fairness, and other legal aspects Egele et al. (2024). Each aspect may be better handled by an appropriate expert.

Another aspect requiring expertise will be the **preparation of baseline methods**. Make sure to include in your organizing team members who are knowledgeable of state-of-the-art methods and capable of implementing or running them on your tasks, to obtain baseline results. The code for baseline methods could be provided to the participants as part of a "starting kit". One motivating factor for the participant is "upskilling" themselves. Make sure you document well the baseline methods and provide good tutorial material, adapted to your audience. This will be much appreciated!

The adage "a rising tide lifts all boats" aptly fits this context. It conveys that when there's a general advancement or progress in a particular scenario (here referring to publicly accessible notebooks), all individuals involved reap the benefits, irrespective of their initial conditions. For instance, on Kaggle, you can kickstart a project using a pre-existing notebook created by someone else, paving the way for collective growth and assistance for all participants.

Do you have enough resources to organize a good challenge?

Challenges with **code submission** thought of being preferable to those with **result submission** in academia. This allows organizers to compare methods in a controlled environment and fosters fairer evaluations by providing to participants equal computational resources. However, with the advent of large foundational models (Chang et al., 2024), training on the challenge platform has become infeasible, for computational reasons. In that case, one can resort to letting the participants train their model on their own premises and submit the code of trained model, to be tested on the platform. Also, while code challenges are often better to raise equity across participants without similar access to resources, computational constraints can hamper participants from using the processing pipelines they are used to, which would not lead to establishing state-of-the-art performance. Other elements of fairness may include not requiring entry fees or the purchase of material (e.g., robots), not to favor entrants who are economically advantaged.

As a reminder, depending on the type of challenge protocol (Table 2), from level λ to γ are associated increasing computational resources to implement the challenge.

Do you have a **budget** to cover these costs and others (like preparing data)? Here is a non-exhaustive list of possible costs. See document 13 of this book for more details (Richard et al., 2024):

- Data collection, labeling, cleaning.

- Data preprocessing and formatting.
- Compensation of engineers preparing baseline methods or impementing the challenge.
- Computational resources to run the challenge.
- Prizes.
- Advertising.
- Organization of a workshop.
- Travel awards to attend the workshop.
- Fees for a challenge hosting service.

Did you budget enough preparation and execution time?

Depending on the novelty and difficulty of the challenge, it can take anywhere from a few weeks to a full year to prepare well a challenge. The participants also need sufficient time to familiarize themselves with the material and execute the tasks. In our experience, a minimum of 40 days is usually needed, but some challenges may require a few months. Finally, post-challenge analyses also require time and effort. Budgeting a few weeks in often necessary.

Did you budget enough time and account for possible delays in getting data, necessary protocol reviews or approvals?

PILLAR 4: Good participation

Few or no participants after months of organizing a challenge can be highly frustrating. While predicting a challenge's success is challenging, it is essential to take all possible measures to ensure strong and quality participation.

Do you know your target audience?

It is important to define the target audience, in order to design a challenge which is attractive enough and adapt the level of difficulty with a barrier to enter that is not too high.

Do you know the population of targeted participants? In order to adapt the difficulty level of a challenge, adapt the content and materials to enter in the challenge, it is recommended to define your target participants and ask yourself what are their backgrounds, skills, strength, weakness? Are they young students, experience professionals, research scientists, etc, with backgrounds in which fields?

If the target audience is a mix of beginners and more experienced practitioners in Artificial Intelligence, a crucial issue is to find a sweet spot, to set the barrier low enough to allow for beginners to enter without too much headache, while keeping the challenge challenging enough for experienced practitioners. Lowering the barrier to enter can be achieved by providing good documentation along with a simplified tutorial in a starting kit, providing compute resources to make it accessible to anyone, not only people with own access to farms of GPU or TPU. And at the same time, keeping the problem to solve interesting enough for experienced practitioners might require several levels of difficulty, several phases of the challenge.

Choose the start date, time length, and time investment required, according to the targeted audience. If you target researchers, make sure that being successful doesn't involve too much engineering time efforts with respect to the scientific contribution, and coordinate with other conferences, workshops and other challenges in the field.

Select a subject that aligns with the interests of your target audience at the time of the challenge. The suitability of a problem for a successful challenge is not solely determined by technical considerations but also by its potential to generate sufficient interest or avoid controversy. It is important to recognize that prizes constitute only a minor component of the incentive for participation, as most participants do not win, and the monetary rewards often pale in comparison to the time invested. Participants are, in effect, **donating their time**! driven by intellectual curiosity, professional growth, or community engagement.

When introducing the topic of your challenge, it is essential to craft a compelling "hook" that piques interest. Aim to attract a diverse range of participants. The beauty of this approach is that a machine learning expert, even without domain-specific knowledge, might clinch victory in a very specialized challenge. Conversely, an industry engineer looking to enhance their skills could very well triumph in a machine learning challenge.

Do you have a plan to harvest the results of your challenge?

Providing participants with opportunities to disseminate their work is both an important motivation to them to enter the challenge and a means of harvesting results. You may want to target one or several conferences (NeurIPS, KDD, WCCI, IDCAR have challenge programs, others welcome challenges organized in conjunction with workshops).

Inform participants of the challenge's start date well in advance to generate anticipation and allow adequate preparation. Ensure that sufficient time is allocated to finalize all necessary preparations before the launch. It is advisable to avoid scheduling challenge deadlines to coincide with major academic events, such as conference submission dates or student examination periods, as these may limit participation. Recurring challenges can foster a cumulative effect, gradually building a dedicated community that contributes to advancing the state-of-the-art over successive iterations.

Conferences do not usually have proceedings for their workshops, so you may have to make your own arrangements for proceedings. One venue that has been welcoming challenge proceedings in PMLR, the Proceedings of Machine Learning Research. The recently founded DMLR journal has also been welcoming challenge papers.

At the very least, top ranking participants should be asked to fill out fact sheets (see and example in Appendix B). Fact sheets can be a mix of textual descriptions providing a brief report in human readable format, and survey answers to a few questions, which can easily be aggregated as statistics. It is best to ask the participants to fill out the fact sheets before revealing the final leaderboard results, because otherwise non-winners have little incentive to put in the effort. Also, it is best to put as a condition to winning prizes to open-source the code and fill out the fact sheets.

Do not under-estimate the duration of challenge preparation, which, depending on the data readiness, the complexity of implementation of the challenge protocol and of establishing results may vary from a few days to over a year. Refer to document 3 of this book for recommendations on how to prepare data (Egele et al., 2024).

Do you have monetary prizes?

Publication venue is an essential motivation for participants of academic challenges. But, a recent analysis has determined that prizes are the greatest factor in boosting participation for Kaggle challenges. While overall participation is mostly driven by the approachability (a challenge with a tabular dataset will have more participation than one with 3D images), all else equal, the prize is 75 percent more important than any other factor. However, substantial prizes can attract participants more interested in monetary gain than scientific advancement, potentially leading to rule violations or the exploitation of challenge loopholes without disclosure. Another form of compensation is reimbursing travel expenses for attending a workshop where challenge results are discussed, which also facilitates result dissemination.

1.1 Do you have an advertising plan?

Last but not least, do not forget to **communicate well with your participants**. This starts with announcing your challenge ahead of time and advertising aggressively a few days into the challenge (once you are confident everything is running smoothly). Use all possible means available: mailing lists, social media, personal contact. Monitor the level of participation, get feed-back and stimulate participation, if needed, by adding bootcamps, webinars, and tutorial sessions. Make use of a forum and stimulate discussions between organizers and participants and between participants.

2 The proposal

In the Section, we provide a template of a proposal and provide a few tips about how to write a good proposal.

ABSTRACT AND KEYWORDS

Briefly describe your challenge. Follow the following template (2 sentences maximum each topic):

- Background and motivation (stress impact).
- Tasks proposed and data used.
- Novelty (compared to previous challenges and benchmarks).
- Baseline methods and results (positioning the state of the art).
- Scientific outcomes expected (list questions asked).

Indicate whether this is a “regular challenge” running over a few months, a “hackathon” taking place over a day or two, and whether this will include a “live challenge” in the form of a demonstration requiring on-site presence. Also, provide up to five keywords, from generic to specific.

challenge description

BACKGROUND AND IMPACT

Provide some background on the problem approached by the challenge and fields of research involved. Describe the scope and indicate the anticipated impact of the challenge prepared (eco-

nomical, humanitarian, societal, etc.). Some venues privilege tasks of humanitarian and/or positive societal impact.

Justify the relevance of the problem to the targeted community and indicate whether it is of interest to a large audience or limited to a small number of domain experts (estimate the number of participants). A good consequence for a challenge is to learn something new by answering a scientific question or make a significant technical advance.

Describe typical real life scenarios and/or delivery vehicles for the challenge. This is particularly important for live challenges, but may also be relevant to regular challenges. For instance: what is the application setting, will you use a virtual or a game environment, what situation(s)/context(s) will participants/players/agents be facing?

Put special emphasis on relating the, necessarily simplified, task of the challenge to a real problem faced in industry or academia. If the task cannot be cast in those terms, provide a detailed hypothetical scenario and focus on relevance to the target audience.

Consider adding in a “hook" as an opening description, to attract those who are unfamiliar with the subject.

Novelty

Have you heard about similar challenges in the past? If yes, describe the key differences. Indicate whether this is a completely new challenge, a challenge part of a series, eventually re-using old data.

Data

If the challenge uses an evaluation based on the analysis of data, provide detailed information about the available data and their annotations. Document your dataset thoroughly, using guidelines such as those provided in (Gebru et al., 2018). The data and their documentation should be ready prior to the official launch of the challenge.

Quantity and quality of data: Justify that: (1) you have access to large enough datasets to make the challenge interesting and draw conclusive results; (2) the data will be made freely available after the contest; (3) the ground truth has been kept confidential.

Legal and ethical issues: Verify and document permissions or licenses to use the chosen data. If new data are collected or generated, provide details on the procedure, including permissions to collect such data obtained by an ethics committee, if human subjects are involved. Minimize exposing personally identifiable information in datasets without informed consent and seek explicit consent when using real data from real people, explaining any inability to do so. If the data are recycled, verify that your dataset is not “deprecated”. The authors of the original data may have recalled the dataset for some good reasons, e.g., data are biased in some way. The conference you are targeting may supply a list of deprecated datasets. Otherwise search on the Internet with your dataset name and “deprecated”. For instance the search for “tiny images deprecated” yields this results: “The deprecation notice for Tiny Images was posted in direct response to a critique by external researchers, who showed that the dataset contained racist and misogynist slurs and other offensive terms, including labels such as *rape suspect* and *child molester*)”.

See document 3 of this book for more details on how to prepare a good dataset (Egele et al., 2024).

TASKS AND APPLICATION SCENARIOS

Describe the tasks of the challenge and explain to which specific real-world scenario(s) they correspond to. If the challenge does not lend itself to real-world scenarios, provide a justification. Justify that the problem posed are scientifically or technically challenging but not impossible to solve. If data are used, think of illustrating the same scientific problem using several datasets from various application domains.

METRICS AND EVALUATION METHODS

For quantitative evaluations, select a scoring metric and justify that it effectively assesses the efficacy of solving the problem at hand. It should be possible to evaluate the results objectively. If no metrics are used, explain how the evaluation will be carried out. Explain how error bars will be computed and/or how the significance in performance difference between participants will be evaluated.

You can include subjective measures provided by human judges (particularly for live / demonstration challenges). In that case, describe the judging criteria, which must be as orthogonal as possible, sensible, and specific. Provide details on the judging protocol, especially how to break ties between judges. Explain how judges will be recruited and, if possible, give a tentative list of judges, justifying their qualifications. See document 4 of this book for help on evaluating a challenge.

BASELINES, CODE, AND MATERIAL PROVIDED

Describe baseline methods that can solve the problems posed in your challenge. Beta-test your challenge with such baseline methods and report the results. This is important to demonstrate that the challenge is not too easy nor too hard. You should have a range of baseline methods, from simple to sophisticated (state of the art methods). The results should show a large difference between unsophisticated and sophisticated methods.

Make the baseline methods part of the participants' "starting kit", which you should make publicly available together with sample data. The starting kit should allow participants to develop their solution and test it in conditions identical to those in which it will be tested on the challenge platform.

For certain challenges, material provided may include a hardware platform. Ideally the participants who cannot afford buying special hardware or do not have access to large computing resources should not be discriminated against. Find a way to make enough resources freely available to deserving participants in need (e.g. participant having demonstrated sufficient motivation by going through a screening test).

TUTORIAL AND DOCUMENTATION

Provide a reference to a white paper you wrote describing the problem and/or explain what tutorial material you will provide. This may include FAQs, Jupyter notebooks, videos, webinars, bootcamps.

Organizational aspects

PROTOCOL

Explain the procedure of the challenge:

- what the participants will have to do, what will be submitted (results or code), and the evaluation procedure;
- whether there will there be several phases;
- whether you will you use a challenge platform with online submissions and a leaderboard;
- what you will do for cheating detection and prevention;
- what you will do for beta-testing.

Code submission challenges can be resource-intensive but offer a plethora of benefits including:

- A controlled environment.
- Confidentiality of data.
- Equal time allocation for participants.
- Implementation of intricate protocols.
- Reduced chances of cheating.
- Accumulation of code for subsequent analysis.

RULES

In this section, provide:

1. A verbatim copy of (a draft of) the contest rules given to the contestants.
2. A discussion of those rules and how they lead to the desired outcome of your challenge.
3. A discussion about cheating prevention. Choose inclusive rules, which allow the broadest possible participation from the target audience.

It is imperative to clearly delineate the rules right from the outset and ensure they remain unchanged throughout. Maintaining transparency is key in fostering trust and participation engagement. Serious competitors prefer well-defined winning conditions. Evaluation procedures must be robust and tested prior to challenge launch to prevent issues. Although maintaining consistent rules is vital, organizers should retain the prerogative to amend rules or data if it's deemed essential. Such modifications, however infrequent, may be necessary to avert nullifying the entire challenge. Any alterations made early in the challenge are typically more acceptable to participants. Late-stage changes can cause discontent as participants might've dedicated significant time and resources, and such amendments might nullify their efforts. Organizers must balance the advantages of a change against its repercussions on the participants. For instance, last-minute minor data corrections might not merit the potential turmoil they could incite amongst competitors.

Organizers face numerous choices like:

- The option between single or multiple accounts.
- Anonymity regulations.

- Setting limits on submission counts.
- Deciding between result or code submissions.
- Instituting rebuttal or review mechanisms for results by fellow participants.

It's beneficial to have an adjudicating body or an uppermost appellate authority. The winners' codes should be subjected to internal result releases and peer review to ensure authenticity and merit.

We provide a concrete example of rules, corresponding to the challenge whose proposal is found in Appendix A.

SCHEDULE AND READINESS

Provide a timeline for challenge preparation and for running the challenge itself. Propose a reasonable schedule leaving enough time for the organizers to prepare the event (a few months), enough time for the participants to develop their methods (e.g. 90 days), enough time for the organizers to review the entries, analyze and publish the results.

For live/demonstration challenges, indicate how much overall time you will need (we do not guarantee all challenges will get the time they request). Also provide a detailed schedule for the on-site contest. This schedule should at least include times for introduction talks/video presentations, demos by the contestants, and an award ceremony.

Will the participants need to prepare their contribution in advance (e.g. prepare a demonstration) and bring ready-made software and hardware to the challenge site? Or, on the contrary, can will they be provided with everything they need to enter the challenge on the day of the challenge? Do they need to register in advance? What can they expect to be available to them on the premises of the live challenge (tables, outlets, hardware, software and network connectivity)? What do they need to bring (multiple connectors, extension cords, etc.)?

Indicate what, at the time of writing this proposal, is already ready.

CHALLENGE PROMOTION

Describe the plan that organizers have to promote participation in the challenge (e.g., mailing lists in which the call will be distributed, invited talks, etc.).

Also describe your plan for attracting participants of groups under-represented in challenge programs.

Resources

ORGANIZING TEAM

Provide a short biography of all team members, stressing their competence for their assignments in the challenge organization. Please note that diversity in the organizing team is encouraged, please elaborate on this aspect as well. Make sure to include: coordinators, data providers, platform administrators, baseline method providers, beta testers, and evaluators.

RESOURCES PROVIDED BY ORGANIZERS, INCLUDING PRIZES

Describe your resources (computers, support staff, equipment, sponsors, and available prizes and travel awards).

For live/demonstration challenges, explain how much will be provided by the organizers (demo framework, software, hardware) and what the participants will need to contribute (laptop, phone, other hardware or software).

SUPPORT REQUESTED

Indicate the kind of support you need from the conference.

For live/demonstration challenges, indicate what you will need in order to run the live challenge.

3 A sample successful proposal

To exemplify the previous guidelines, we provide an example of successful NeurIPS proposal in Appendix A.

4 Conclusion

In this document, we have covered the fundamentals of organizing challenges. For a more comprehensive understanding of all aspects of challenge organization, please refer to the subsequent documents of this book.

The success of any challenge hinges predominantly on a strong team and a well-structured plan. A few key suggestions: don't underestimate the effort needed to execute your organization plan and bring in additional volunteers as necessary. Allow ample time to beta-test your challenge. If you're new to this, joining a seasoned organizing team to gain hands-on experience is likely the best approach.

Acknowledgements:

This work was supported in part by ANR Chair of Artificial Intelligence HUMANIA ANR-19-CHIA-0022 and TAILOR EU Horizon 2020 grant 952215.

Appendix A: Example of challenge proposal

This appendix provides an example of challenge proposal, the Cross-Domain Meta-Learning Challenge.

Cross-Domain Meta-Learning NeurIPS 2022 Competition Proposal

Dustin Carrión* **Ihsan Ullah*** **Sergio Escalera** **Isabelle Guyon** **Felix Mohr** **Manh Hung Nguyen**

metadl@chalearn.org

Abstract

Meta-learning aims to leverage the experience from previous tasks to solve new tasks using only little training data, train faster and/or get better performance. The proposed challenge focuses on "cross-domain meta-learning" for few-shot image classification using a novel "any-way" and "any-shot" setting. The goal is to meta-learn a good model that can quickly learn tasks from a variety of domains, with any number of classes also called "ways" (within the range 2-20) and any number of training examples per class also called "shots" (within the range 1-20). We carve such tasks from various "mother datasets" selected from diverse domains, such as healthcare, ecology, biology, manufacturing, and others. By using mother datasets from these practical domains, we aim to maximize the humanitarian and societal impact. The competition is with code submission, fully blind-tested on the CodaLab challenge platform. A single (final) submission will be evaluated during the final phase, using ten datasets previously unused by the meta-learning community. After the competition is over, it will remain active to be used as a long-lasting benchmark resource for research in this field. The scientific and technical motivations of this challenge include scalability, robustness to domain changes, and generalization ability to tasks (a.k.a. episodes) in different regimes (any-way any-shot).

Keywords

Deep Learning, AutoML, Few Shot Learning, Meta-Learning, Cross-Domain Meta-Learning.

1 Competition description

1.1 Background and impact

Traditionally, image classification has been tackled using deep learning methods whose performance relies on the availability of large amounts of data [1, 2]. Recent efforts in meta-learning [3, 4, 5] have contributed to making a lot of progress in few-shot learning for image classification problems. Tasks or "episodes" are made of a certain number of classes or "ways" and number of examples per class or "shots". Depending on the regime (number of shots and ways) various techniques have been proposed [6, 7, 8]. Despite progress made, allowing the community to reach accuracies in the high 90% in the last ChaLearn meta-learning challenge [9], evaluation protocols have a common drawback: Even when evaluated on multiple domains (e.g., insect classification, texture classification, satellite images, etc.), models meta-trained on a given domain are meta-tested on the same domain, usually simply assimilated to a large multi-class dataset from which tasks are carved. Furthermore, the number of ways and shots is usually fixed.

*The two first authors are co-lead organizers. Other authors are in alphabetical order of last name.

Preprint. Under review.

In contract, in this proposed challenge, domains vary, number of shots vary, and number of ways vary in every task (a.k.a. episode). For the purpose of this challenge, “domain” is not defined as a single large dataset, but as a collection of datasets from a similar application domain, with same type of images (object or texture) and same scale (microscopic, humans scale, or macroscopic). We call “mother dataset” a multi-class dataset from a particular domain from which we carve out tasks. Our mother datasets have each at least 20 classes and 40 examples per class (but usually many more). We have selected well differentiated domains, spanning object and texture recognition problems, at different scales. We formatted 30 mother datasets from 10 domains. Meta-training and meta-testing is performed on different mother datasets, one from each domain in each phase. For meta-testing, N-way k-shot tasks are drawn from one of the 10 mother datasets, with N in 2-20 and k in 1-20.

As documented in the literature, single domain meta-learning approaches have poor generalization ability to unrelated domains [10, 11, 12]. Nevertheless, this kind of generalization is crucial since there are scenarios where only one or two examples per class are available (e.g., rare birds or plants), and there is no close domain with enough information that can serve as source [13]. Therefore, addressing domain variations has become a research area of great interest.

In this sense, with the proposed challenge, we aim to provide a benchmark instrument that any person interested in the problem of cross-domain meta-learning for image recognition can use. Due to the rapidly increasing interest in meta-learning, we expect a large audience to be actively involved in this event. Additionally, the data we have collected for this challenge maximizes the societal impact by assembling datasets from various practical domains directly relevant to “AI for good", including medicine, ecology, biology, and others. We plan to reach out to a diverse community of participants with the organization of a bootcamp and prizes in several leagues (NewInML, Women, and participants of rarely represented countries). Moreover, the outcomes of this competition will help in the democratization of AI since the code of the winners will be open-sourced. On one hand, trained meta-learners can be used to create on-demand few-shot image classifiers for users having no particular knowledge of machine learning. On the other hand, meta-learners can be used to create few-shot model trainers, in other areas than image classification.

1.2 Novelty

Cross-domain meta-learning competition is part of the MetaDL competition series (see Table 1). It is built as an advancement of the NeurIPS 2021 MetaDL competition to target the problem of cross-domain meta-learning. Although the proposed competition is the third one around few-shot learning, the focus of the previous competitions (NeurIPS 2021 MetaDL and AAAI 2020 MetaDL-mini) was on single domain meta-learning, i.e., to build algorithms that perform well on unseen data, which is similar, but not same, to the seen data during training. On the other hand, the proposed Cross-domain meta-learning competition aims to tackle a problem more related to the real world scenario where the data can come from different domains. The idea is to build algorithms capable of generalizing the knowledge acquired during meta-training to any domain using only a few data. Thus, the models should “learn to learn" features that are domain independent to facilitate their adaptation to unseen tasks.

Table 1: ChaLearn Competition Series.

Conference	**Challenge**	**Description**
ICML & NeurIPS 2016-18	AutoML	Automating complete ML pipeline
WAIC 2019	AutoNLP	Natural Language Processing
ECML PKDD 2019	AutoCV	Computer Vision
ACML 2019	AutoWeakly	Weakly Supervised Learning
WSDM 2019	AutoSeries	Time Series
NeurIPS 2019	AutoDL	Misc. domains
KDD cup 2020	AutoGraph	Classification of Graph Data
InterSpeech 2020	AutoSpeech	Speech Recognition
AAAI 2021	2020 MetaDL-mini	Few shot learning, trial run
NeurIPS 2021	2021 MetaDL	Few shot learning, meta-learning

In the NeurIPS 2021 MetaDL competition (last in the series), algorithms were evaluated in each phase on tasks drawn from a single dataset of a given domain, experiments were repeated on multiple

datasets from various domains, and performances were fused for the final ranking. In contrast, in the proposed NeurIPS 2022 competition, the evaluation is carried out by pooling domains: tasks are drawn from any domain, both during meta-training and meta-testing. We have also increased the number of domains from 5 to 10 and the number of datasets from 15 to 30, see next section.

1.3 Data

Our competition will use a meta-dataset called Mata-Album[2], prepared in parallel with the competition, to be released after the competition ends. Several team members of the competition are also co-authors of this meta-dataset. It consists of 10 domains with three image "mother datasets" per domain. Although these datasets are publicly available, we selected for test purposes dataset that are not part of past meta-learning benchmarks. We preprocessed data in a standard format[3] suitable for few-shot learning. The preprocessing includes image resizing with anti-aliasing filters into a uniform shape of 128x128x3 pixels. The complete preprocessing pipeline is published[4].

Similar to our previous challenges, this competition has 3 phases: (1) **public phase** (release of starting kit and 10 public datasets), (2) **feedback phase** (participants submit their code to get instant feedback on 10 hidden datasets), and (3) **final phase** (the last submission of each participant from feedback phase evaluated on 10 new hidden datasets).

The datasets used for each phase are selected based on their novelty to the community: the most novel ones are used for meta-testing in the final phase. All the phases have one dataset per domain. Table 2 shows the datasets that will be used in each phase from each domain and their associated classification task. Except for one dataset, which might be replaced, we have obtained licenses for all the datasets. They will be released on OpenML [14] after the competition ends.

1.4 Tasks and application scenarios

In this challenge, we aim at pushing the full automation of few-shot learning by demanding participants to design learning agents capable of producing a trained classifier in the cross-domain few-shot setting. We will use the traditional N-way k-shot classification setting illustrated in Figure 1. This setting consists of three phases–meta-training, meta-validation, and meta-testing–which are used for meta-learning, hyperparameter tuning, and evaluation, respectively.

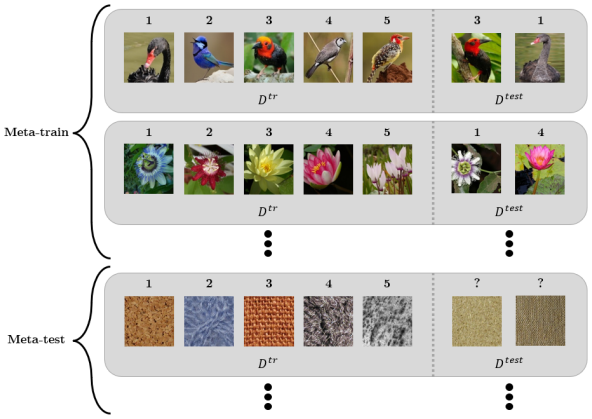


Figure 1: Illustration of 5-way 1-shot classification. This means that we have 5 classes and only one example of each class for learning. The test set includes a number of "query" examples, which are labeled in the meta-training set, but unlabeled in the meta-test set. Meta-validation tasks are not displayed. Figure adapted from [15].

Each phase is composed of multiple *episodes* which are small tasks $\mathcal{T}_j = \left(\mathcal{D}_{\mathcal{T}_j}^{train}, \mathcal{D}_{\mathcal{T}_j}^{test}\right)$ where $\mathcal{D}_{\mathcal{T}_j}^{train}$ and $\mathcal{D}_{\mathcal{T}_j}^{test}$ are known as *support set* and *query set*, respectively. Since the proposed setting assumes a cross-domain scenario, each task can be generated from any of the available mother datasets $\mathcal{D}$ in each phase. Moreover, the N-way k-shot classification setting states that every support

[2] Meta-Album: https://github.com/ihsaan-ullah/meta-album

[3] Data Format: https://github.com/ihsaan-ullah/meta-album/tree/master/DataFormat

[4] Preprocessing pipeline: https://github.com/ihsaan-ullah/meta-album/tree/master/PreProcessing

Table 2: **Datasets** to be used in the proposed competition. The 10 first are the "freshest" and will be used for the final test phase; the 10 in the middle will be used in the feedback phase; the last 10 datasets will be released in the public phase. The dataset written in red means that we have not obtained its license.

Domain	Dataset	Meta-Album ID	Classification Task
FINAL TEST PHASE			
1. Large Animals	Animal with Attributes	$LR\_AM.AWA$	Mammals
2. Small Animals	Insects	$SM\_AM.INS$	Insects
3. Plants	Fungi	$PLT.FNG$	Fungi
4. Plant Diseases	Plant Doc	$PLT\_DIS.PLT\_DOC$	Sick and healthy leaves
5. Microscopy	Kimia 24	$MCR.KIMIA\_24$	Human tissues
6. Remote Sensing	RSD	$REM\_SEN.RSD$	Satellite images
7. Vehicles	Boats	$VCL.BTS$	Boat types
8. Manufacturing	Textures ALOT	$MNF.TEX\_ALOT$	Textures
9. Human Actions	MPII Human Pose	$HUM\_ACT.ACT\_410$	Human pose
10. OCR	OmniPrint-MD-6	$OCR.MD\_6$	Digital characters
FEEDBACK PHASE			
1. Large Animals	Stanford Dogs	$LR\_AM.DOG$	Dog breeds
2. Small Animals	Insects	$SM\_AM.INS\_2$	Insects
3. Plants	PlantNet	$PLT.PLT\_NET$	Plant types
4. Plant Diseases	Medicinal Leaf	$PLT\_DIS.MED\_LF$	Medicinal plants
5. Microscopy	PanNuke	$MCR.PNU$	Nuclei instance
6. Remote Sensing	RSICB	$REM\_SEN.RSICB$	Aerial images
7. Vehicles	Airplanes	$VCL.APL$	Airplane types
8. Manufacturing	Textures DTD	$MNF.TEX\_DTD$	Textures
9. Human Actions	Stanford 40 Actions	$HUM\_ACT.ACT\_40$	Human pose
10. OCR	OmniPrint-MD-5-bis	$OCR.MD\_5\_BIS$	Digital characters
PUBLIC PHASE			
1. Large Animals	Birds	$LR\_AM.BRD$	Bird species
2. Small Animals	Plankton	$SM\_AM.PLK$	Plankton types
3. Plants	Flowers	$PLT.FLW$	Flower categories
4. Plant Diseases	Plant Village	$PLT\_DIS.PLT\_VIL$	Plant leaves
5. Microscopy	DiBas	$MCR.BCT$	Bacterial colony
6. Remote Sensing	RESISC	$REM\_SEN.RESISC$	Aerial images
7. Vehicles	Cars	$VCL.CRS$	Car models
8. Manufacturing	Textures	$MNF.TEX$	Textures
9. Human Actions	73 sports	$HUM\_ACT.SPT$	Human sports pose
10. OCR	OmniPrint-MD-mix	$OCR.MD\_MIX$	Digital characters

set contains exactly N classes with k examples per class ($|\mathcal{D}_{\mathcal{T}_j}^{train}| = N \times k$). Furthermore, the classes in the query set $\mathcal{D}_{\mathcal{T}_j}^{test}$ must be present in the support set $\mathcal{D}_{\mathcal{T}_j}^{train}$ of a given task $\mathcal{T}_j$.

During the meta-training phase, the number of ways and shots for each task can be selected by each participant. However, during meta-validation and meta-testing, the number of ways will range from 2 to 20 with a number of shots ranging from 1 to 20, i.e., during meta-validation and meta-testing, the tasks will be any-way any-shot. To facilitate the creation of the learning agents, we will provide a large number of datasets formatted uniformly, amenable to meta-learning. Additionally, although different datasets are used in each phase, the domains remain the same.

The application scenarios are two-fold, corresponding to the 2 first prize leagues (see Section 2.4): (1) Few-shot image classification: a user seeks to create a classifier in a new domain, by providing a handful of examples of a number of classes. To that end, the meta-trained learning machine of the winners will be made readily available at the end of the challenge. (2) Meta-learning from limited amounts of meta-learning data: a user seeks to meta-train himself a learning machine in an application area other than image classification. To that end, the meta-learners of the winners will be open-sourced.

1.5 Metrics

Once the learning agents are meta-trained, they are presented with the meta-test dataset. The meta-test dataset consists of several episodes. For each episode, the agent is trained with the labeled support set $\mathcal{D}_{\mathcal{T}_j}^{train}$, and it is required to make predictions on the unlabeled query set $\mathcal{D}_{\mathcal{T}_j}^{test}$. The participants' performance on a domain will be the average classification accuracy of all tasks/episodes in the meta-test phase across all domains.

The error bars will be a 95% confidence interval of the mean classification accuracy computed as follows:

$$CI = \pm z^* \times \frac{\sigma}{\sqrt{n}}, \tag{1}$$

where z^* is the corresponding value of the Normal distribution based on the confidence level, since in this case it is 95%, $z^* = 1.96$; σ corresponds to the standard deviation of the accuracy obtained in all the episodes of the meta-test dataset, and n is the number of episodes in the meta-test dataset. If computationally feasible, n will be increased to obtain significant differences between top ranking participants at the 95% level. The stability of the ranking will also be evaluated by bootstrap resampling performances on the various tasks/episodes.

CI calculations and bootstrap experiments will only be indicative and not used to declare ties. Winners will be determined according to best rank in the final phase and ties broken according to first submission made.

1.6 Baselines, code, and material provided

The organizers will provide a "starting kit" that will be available to download directly from the challenge website as soon as the challenge starts. The starting kit will provide all the necessary code so the participants can make their local tests before submitting. Since 10 datasets will also be available, the starting kit will provide the corresponding data loader that will create the episodes as described in section 1.4. The competition is operationalized via a specific API to which participants must adhere, and which is documented in the starting kit.

The starting kit supplies five *baseline* agents that should be outperformed: (i) a random agent, which generates random predictions for each episode, (ii) a naïve approach, which accumulates all data from meta-train and trains a neural network on it and then applies it to the meta-test dataset, (iii) prototypical networks [16], (iv) MAML [17], and (v) prototypical networks with feature-wise transformations [13].

Table 3 shows the baseline results for the feedback phase that will be provided for the competition. This table includes the overall classification accuracy (i.e., the average classification accuracy of all meta-testing episodes) and the classification accuracy per dataset with their corresponding confidence intervals. The results of baseline (v) are not included in the table since it is under development.

1.7 Website, tutorial and documentation

We have set up a GitHub repository[5] for this competition which serves as the landing page and will be linked to the competition's CodaLab website. The GitHub repository covers the following points:

- Competition introduction and instructions for setting up the environment, including installing the required packages.
- Complete details of the evaluation process.
- Information about how to make a submission.
- Troubleshooting instructions for any possible issues and contact details for reporting issues.
- Link to CodaLab competition.
- Link to a dedicated forum on CodaLab platform for easy and efficient communication with participants.

In addition, a code tutorial is provided for the purpose of:

[5] GitHub repository: https://github.com/DustinCarrion/cd-metadl

Table 3: **Baseline results** for the feedback phase. TL and PN stand for Transfer Learning and Prototypical networks, respectively. The Overall row corresponds to the average classification accuracy of 1000 meta-testing episodes and its corresponding confidence interval. The remaining rows are the average classification accuracy and confidence interval of all the episodes carved out from each dataset. The name of the datasets corresponds to the Meta-Album IDs presented in Table 2.

	Random	**Naïve TL**	**PN**	**MAML**
Overall	0.14 ± 0.01	0.16 ± 0.01	0.38 ± 0.02	0.17 ± 0.01
DOG	0.15 ± 0.03	0.10 ± 0.02	0.29 ± 0.03	0.17 ± 0.02
INS_2	0.14 ± 0.03	0.16 ± 0.03	0.31 ± 0.04	0.16 ± 0.03
PLT_NET	0.15 ± 0.03	0.17 ± 0.03	0.38 ± 0.04	0.18 ± 0.03
MED_LF	0.16 ± 0.03	0.16 ± 0.03	0.71 ± 0.03	0.27 ± 0.04
PNU	0.12 ± 0.02	0.16 ± 0.03	0.20 ± 0.03	0.17 ± 0.02
RSICB	0.12 ± 0.02	0.24 ± 0.04	0.72 ± 0.03	0.16 ± 0.03
APL	0.14 ± 0.03	0.12 ± 0.02	0.47 ± 0.04	0.16 ± 0.03
TEXT_DTD	0.13 ± 0.03	0.19 ± 0.03	0.30 ± 0.03	0.14 ± 0.02
ACT_40	0.15 ± 0.03	0.16 ± 0.03	0.21 ± 0.03	0.14 ± 0.02
MD_5_BIS	0.14 ± 0.03	0.15 ± 0.03	0.17 ± 0.03	0.16 ± 0.03

- Loading and discovering properties of data.
- Explaining the coding structure and the expected functions to be implemented in the code submissions.
- Providing instructions and examples for running the baseline methods on the public datasets.

2 Organizational aspects

2.1 Protocol

The competition will be hosted and run on the CodaLab platform[6] to which participants submit their solutions and receive summaries on their submissions. The 10 public datasets and the starting kit, which contains the API description and example agents can be downloaded from the GitHub repository. With this material, submissions can be drafted and tested (on the public datasets); no registration is required up to this point. To be able to make submissions to the system and hence enter the competition, the participants must create an account on the CodaLab platform, and then they can register for the competition. Neither the creation of the CodaLab account nor the registration into the competition has a fee. Once the participants are registered, they can submit agent solutions to the CodaLab server, which will immediately execute them on the feedback phase datasets and automatically display the results on the leader-board as soon as the run is finished.

As soon as the competition starts, the participants have direct or indirect access to 20 datasets in total. The leaderboard shown on the CodaLab platform, which is of main interest to the participants, is based on the 10 datasets of the feedback phase, to which the participants have no immediate access. The main benefit of the leaderboard is to enable a fair and objective evaluation of the submissions: all the submissions will be restricted by 10 GPU-hours of execution, and the computational resources will be the same, i.e., the CodaLab server will execute all submissions. The same hardware will be used in the final phase. However, the provision of this order of resources implies the necessity of limitations: Each participant will be allowed to make only 5 submissions per day and a maximum of 100 submissions in the course of the challenge. To allow the participants to perform other experiments on their own hardware, they can make use of the 10 datasets of the public phase, which are directly available through a download.

The proposed protocol was already tested in the previous challenges (2020 MetaDL-mini and 2021 MetaDL), but it was also tested when running the provided baselines. Furthermore, since we have team members of the CodaLab platform as co-applicants (Isabelle Guyon, Sergio Escalera) in the competition, we will be able to address CodaLab bugs and issues efficiently.

[6] CodaLab platform: https://codalab.lisn.upsaclay.fr

2.2 Rules

Draft of the rules:

- **General Terms**: This challenge is governed by the General ChaLearn Contest Rule Terms, the CodaLab Terms and Conditions, and the specific rules set forth.
- **Announcements**: To receive announcements and be informed of any change in rules, the participants must provide a valid email.
- **Conditions of participation**: Participation requires complying with the rules of the challenge. Prize eligibility is restricted by US government export regulations, see the General ChaLearn Contest Rule Terms. The organizers, sponsors, their students, close family members (parents, sibling, spouse or children) and household members, as well as any person having had access to the truth values or to any information about the data or the challenge design giving him (or her) an unfair advantage, are excluded from participation. A disqualified person may submit one or several entries in the challenge and request to have them evaluated, provided that they notify the organizers of their conflict of interest. If a disqualified person submits an entry, this entry will not be part of the final ranking and does not qualify for prizes. The participants should be aware that ChaLearn and the organizers reserve the right to evaluate for scientific purposes any entry made in the challenge, whether or not it qualifies for prizes.
- **Dissemination**: The challenge is part of the official selection of the NeurIPS 2022 conference. There will be publication opportunities for competition reports co-authored by organizers and participants.
- **Registration**: The participants must register to CodaLab and provide a valid email address. Teams must register only once and provide a group email, which is forwarded to all team members. Teams or solo participants registering multiple times to gain an advantage in the competition may be disqualified.
- **Anonymity**: The participants who do not present their results at the conference can elect to remain anonymous by using a pseudonym. Their results will be published on the leaderboard under that pseudonym, and their real name will remain confidential. However, the participants must disclose their real identity to the organizers to claim any prize they might win. See our privacy policy for details.
- **Submission method**: The results must be submitted through this CodaLab competition site. The number of submissions per day and maximum total computational time are restrained and subject to change, according to the number of participants. Using multiple accounts to increase the number of submissions in NOT permitted. In case of problem, send email to `metalearningchallenge@googlegroups.com`. The entries must be formatted as specified on the Instructions page.
- **Reproducibility**: The participant should make efforts to guarantee the reproducibility of their method (for example by fixing all random seeds involved). In the Final Phase, all submissions will be run three times, and the worst performance will be used for final ranking.
- **Prizes**: The three top ranking participants in the Final phase blind testing may qualify for prizes. The last valid submission in Feedback Phase will be automatically submitted to the Final Phase for final evaluation. The participant must fill out a fact sheet briefly describing their methods. There is no other publication requirement. The winners will be required to make their code publicly available under an OSI-approved license such as, for instance, Apache 2.0, MIT or BSD-like license, if they accept their prize, within a week of the deadline for submitting the final results. Entries exceeding the time budget will not qualify for prizes. In case of a tie, the prize will go to the participant who submitted his/her entry first. Non winners or entrants who decline their prize retain all their rights on their entries and are not obliged to publicly release their code.

Discussion: The rules have been designed with the criteria of *inclusiveness for all participants* and *openness of results* in mind. We aim to achieve inclusiveness for all participants by allowing them to enter anonymously and providing them cycles of computation (for the feedback phase and final phase) on our compute resources. This way, participants that do not have ample computing resources will not be limited by this and have a fair chance to win the challenge. We aim to achieve openness of results by requiring all participants to upload their base code and, afterward, fill in a fact sheet about the used methods. This allows us to conduct post-challenge analyzes on the winners' methods.

Cheating prevention: We will execute the submissions on our own compute cluster to prevent participants from cheating, and the testing datasets will remain hidden in the CodaLab platform.

Peeking at the final evaluation datasets will be impossible since those datasets are not even installed on the server during the feedback phase. Using the data in an unintended way during the final phase will be prevented by not revealing the true test labels to the agents ever, but only showing them to the scoring program on the platform. Moreover, different mother datasets for each domain are used in each phase to avoid both domain-specific cheating and also overfitting. We will also monitor submissions and reach out to participants with suspicious submission patterns. Finally, the candidate winners will have to open-source their code to claim their prize. Their code will be individually scrutinized by all other participants before they earn their prize.

2.3 Schedule and readiness

The competition preparations started in November 2021. The running duration of the competition will be 4 months, from June 2022 to September 2022. This time includes all the challenge phases. We are currently finishing the preparation of the baselines and working simultaneously on setting up the competition on the CodaLab platform. The data and protocol preparation is already finished. More details about the competition schedule are given in Table 4.

Table 4: Envisioned competition schedule.

Date	Phase	Description
November 2021 - January 2022	Preparation	Data preparation
February 2022 - March 2022	Preparation	Protocol preparation
April 2022 - May 2022	Preparation	Baselines preparation and Setting up challenge environment
June 2022	Public Phase	Start of the public phase, publicity
July 2022 - August 2022	Feedbak Phase	Start of competition submissions
September 2022	Final Phase	Evaluating performance on hidden datasets
October 2022	Results	Notification of winners

2.4 Competition promotion and incentives

To promote the competition, we will use the following channels:

- Mailing list from hundreds of participants from past challenges we organized.
- Advertisement on CodaLab, MLNews and comp.ai.neural-nets groups.
- Advertisement on the front page of OpenML (120,000 unique visitors yearly).
- In-network advertisement, e.g. personal Twitter accounts and personal emails.
- Organization of a bootcamp, held in presence at Universidad de La Sabana, Colombia, with remote participation permitted.

The bootcamp, organized similarly as the bootcamp of our previous meta-learning competition (https://metalearning.chalearn.org/metadlneurips2021) will encourage participation of South American students and researchers. Additionally, to encourage a diversity of participants and types of submissions, we will provide prizes in 5 different leagues:

- **Free-style league:** Submit a solution obeying basic challenge rules (pre-trained models allowed).
- **Meta-learning league:** Submit a solution that meta-learns from scratch (no pre-training allowed).
- **New-in-ML league:** Be a participant who has less than 10 ML publications, none of which ever accepted to the main track of a major conference.
- **Women league:** Special league to encourage women, since they rarely enter challenges.
- **Participant of a rarely represented country:** Be a participant of a group that is not in the top 10 most represented countries of Kaggle challenge participants[7].

The same participant can compete in several leagues. ChaLearn http://chalearn.org will donate a prize pool of 3000 USD, of which 600 USD will be distributed in each league (1^{st} rank=300, 2^{nd} rank=200, 3^{rd} rank=100), and we will issue certificates to the winners. Furthermore, we will invite the winning participants to work on a post-challenge collaborative paper. We already have experience working on such a collaborative paper thanks to the analysis of the NeurIPS 2019 AutoDL challenge [18] and the NeurIPS 2021 MetaDL challenge [9].

[7] Kaggle ranking users: https://towardsdatascience.com/kaggle-around-the-world-ccea741b2de2

A Resources

This information does not count towards the 8 pages limit.

A.1 Organizing team

- **Dustin Carrión**
  Université Paris-Saclay and LISN, France - dustin.carrion@gmail.com
  He is a second year master's student in Artificial Intelligence at Université Paris-Saclay, France. He is working under the supervision of Professors Isabelle Guyon and Sergio Escalera on cross-domain meta-learning. His research interests include meta-learning, self-supervised learning, semi-supervised learning and continual learning for computer vision applications. **He is the primary organizer with Ihsan Ullah. He contributed to the data collection, will implement the competition protocol, and will implement baseline methods.**
- **Ihsan Ullah** (https://ihsaan-ullah.github.io/)
  Université Paris-Saclay, France - ihsan2131@gmail.com
  He is a second year masters student in Artificial Intelligence at Université Paris-Saclay, France. He is working under the supervision of Professor Isabelle Guyon on challenge organization, data preparation, machine learning and meta-learning. **He coordinated the data collection and will contribute to the implementation of the competition protocol and the baseline methods.**
- **Sergio Escalera** (https://sergioescalera.com/)
  Full Professor at the Department of Mathematics and Informatics, Universitat de Barcelona, where he is the head of the Informatics degree. He is ICREA Academia. He leads the Human Pose Recovery and Behavior Analysis Group. He is an adjunct professor at Universitat Oberta de Catalunya and Dalhousie University, and Distinguished Professor at Aalborg University. He has been visiting professor at TU Delft and Aalborg Universities. He is a member of the Visual and Computational Learning consolidated research group of Catalonia. He is also a member of the Computer Vision Center at UAB, Mathematics Institute of the Universitat de Barcelona, and the Barcelona Graduate School of Mathematics. He is series editor of The Springer Series on Challenges in Machine Learning. He is vice-president of ChaLearn Challenges in Machine Learning, leading ChaLearn Looking at People events. He is co-creator of CodaLab open source platform for challenges organization and co-founder of the NeurIPS competition and Datasets & Benchmarks tracks. He is also Fellow of the ELLIS European Laboratory for Learning and Intelligent Systems working within the Human-centric Machine Learning program, member of the AAAC Association for the Advancement of Affective Computing, AERFAI Spanish Association on Pattern Recognition, ACIA Catalan Association of Artificial Intelligence, AEPIA Artificial Intelligence Spanish Association, INNS International Neural Network Society, Senior IEEE member, and vice-Chair of IAPR TC-12: Multimedia and visual information systems. He has different patents and registered models. He participated in several international funded projects and received an Amazon Research Award. He has published more than 300 research papers and participated in the organization of scientific events. He received a CVPR best paper award nominee and a CVPR outstanding reviewer award. He has been guest editor at TPAMI, JMLR, PR, TAC and IJCV, among others. He has been General co-Chair of FG20, Area Chair at CVPR, ECCV, NeurIPS, ECMLPKDD, AAAI, ICCV, WACV, IJCAI, FG, ICIAP, and BMVC, and Competition and Demo Chair at FG, NeurIPS, and ECMLPKDD, among others. His research interests include inclusive, transparent, and fair analysis of humans from visual and multi-modal data. **He is co-advisor of Dustin Carrion. He will provide expert advice on computer vision aspects and oversee fairness aspects of the competition.**
- **Isabelle Guyon** (https://guyon.chalearn.org/)
  Université Paris-Saclay, France and Chalearn, USA - guyon@chalearn.org
  She is chaired professor in artificial intelligence at the Université Paris-Saclay, specialised in statistical data analysis, pattern recognition and machine learning. She is co-founder and president of ChaLearn, a non-profit organisation dedicated to challenges in machine learning. Her areas of expertise include computer vision and bioinformatics. Prior to joining Paris-Saclay she worked as an independent consultant and was a researcher at ATT

Bell Laboratories, where she pioneered applications of neural networks to pen computer interfaces (with collaborators including Yann LeCun and Yoshua Bengio) and co-invented with Bernhard Boser and Vladimir Vapnik Support Vector Machines (SVM), which became a textbook machine learning method. She worked on early applications of Convolutional Neural Networks (CNN) to handwriting recognition in the 1990s. She is also the primary inventor of SVM-RFE, a variable selection technique based on SVM. The SVM-RFE paper has thousands of citations and is often used as a reference method against which new feature selection methods are benchmarked. She also authored a seminal paper on feature selection that received thousands of citations. She organised many challenges in Machine Learning since 2003 supported by the EU network Pascal2, NSF, and DARPA, with prizes sponsored by Microsoft, Google, Facebook, Amazon, Disney Research, and Texas Instrument. She was a leader in the organisation of the AutoML and AutoDL challenge series https://autodl.chalearn.org/ and the meta-learning challenge series https://metalearning.chalearn.org/. **She is the advisor of the two primary organizers. She will oversee the good conduct of the competition, post-challenge analyses, and provide computational and platform resources.**

- **Felix Mohr** (https://github.com/fmohr)
  Associate Professor at Universidad de La Sabana, Chía, Colombia. He received his PhD in 2016 from Paderborn University, Germany, on the topic of automated service composition. He is an expert on Automated Machine Learning (AutoML) and is main author of several approaches and tools in this area including ML-Plan and Naive AutoML. His research interests include efficiency in stochastic optimization with a focus on the domain Automated Machine Learning. His latest interest is to use learning curves for meta-learning in order to increase the efficiency of model selection. **He will be in charge of the bootcamp organization. He will also provide expert advice on meta-learning and review protocols and rules of the competition.**

- **Manh Hung Nguyen** (https://mhnguyenn.github.io/)
  Chalearn, USA - hungnm.vnu@gmail.com
  He holds a Master's degree in Big Data Management and Analytics (BDMA) from the University of Paris-Saclay. He was an EMJM full-ride scholarship holder. His research interests include Automated Machine Learning, Meta-learning and Reinforcement Learning from both theoretical and practical points of view. He completed his research internship on Meta-learning at Inria under the supervision of Prof. Isabelle Guyon and Dr. Lisheng Sun-Hosoya. He was the lead organiser of the the IEEE WCCI 2022 Meta-learning from Learning Curves competition. **He will be in charge of communication and advertising. He will also contribute to data collection, implementation of the competition protocol and baseline methods.**

A.2 Resources provided by organizers

We are relying on the following resources:

- **Competition infrastructure:** We will use the public instance of CodaLab https://codalab.lisn.fr/ hosted by Université Paris-Saclay (the home institution of 4/6 organizers) as competition platform. To process participant submissions, we will supply 20 compute workers dedicated to the competition. Each compute worker will be equipped with one GPU NVIDIA RTX 2080Ti, 4 vCPUs and 16 GB DDR4 RAM. Our protocol was designed such that these computing resources should suffice to support the challenge.

- **Other computing Resources:** Through our sponsors, we have access to additional computing resources, which we used to prepare data and perform baseline experiments, and we will use in post-challenge experiments. We received a Google grant of 100,000 credits, equivalent to approximately 91575 GPU hours on a Tesla M60 GPU. Additionally, we have access to our institutional computation clusters.

- **Support Staff:** The CodaLab platform is administered by dedicated engineering staff at Université Paris-Saclay. During all the competition, the organizers will be available to support the participants through the forum of the challenge.

A.3 Support requested

We will take the main responsibility for the publicity of our competition, ensuring plenty of participants from the NeurIPS community. To support us in this publicity, we count on the NeurIPS organization for the following matters:

- Display of the competition on the NeurIPS 2022 website.
- Referring participants to the various competitions that are organized.
- A time slot during the program, among which we can announce the winner and discuss the setup.

References

[1] N. Sharma, V. Jain, and A. Mishra, "An Analysis Of Convolutional Neural Networks For Image Classification," Procedia Computer Science, vol. 132, pp. 377–384, 2018.

[2] N. Jmour, S. Zayen, and A. Abdelkrim, "Convolutional neural networks for image classification," in International Conference on Advanced Systems and Electric Technologies (IC_ASET), 2018, pp. 397–402.

[3] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-Transfer Learning for Few-Shot Learning," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 403–412.

[4] M. A. Jamal and G.-J. Qi, "Task Agnostic Meta-Learning for Few-Shot Learning," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11 711–11 719.

[5] M. Patacchiola, J. Turner, E. J. Crowley, M. O' Boyle, and A. J. Storkey, "Bayesian Meta-Learning for the Few-Shot Setting via Deep Kernels," in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 16 108–16 118.

[6] L. Zhang, T. Xiang, and S. Gong, "Learning a Deep Embedding Model for Zero-Shot Learning," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3010–3019.

[7] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese Neural Networks for One-shot Image Recognition," in 32nd International Conference on Machine Learning, 2015, pp. 1–8.

[8] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to Compare: Relation Network for Few-Shot Learning," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1199–1208.

[9] A. E. Baz, I. Ullah, and etal, "Lessons learned from the neurips 2021 metadl challenge: Backbone fine-tuning without episodic meta-learning dominates for few-shot learning image classification," PMLR, 2022, to appear.

[10] B. Kang and J. Feng, "Transferable Meta Learning Across Domains," in 34th Conference on Uncertainty in Artificial Intelligence (UAI), 2018, pp. 1–11.

[11] Y. Guo, N. C. Codella, L. Karlinsky, J. V. Codella, J. R. Smith, K. Saenko, T. Rosing, and R. Feris, "A Broader Study of Cross-Domain Few-Shot Learning," in Computer Vision – ECCV 2020, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Springer International Publishing, 2020, pp. 124–141.

[12] C. P. Phoo and B. Hariharan, "Self-training For Few-shot Transfer Across Extreme Task Differences," in 9th International Conference on Learning Representations (ICLR), 2021, pp. 1–19.

[13] H.-Y. Tseng, H.-Y. Lee, J.-B. Huang, and M.-H. Yang, "Cross-Domain Few-Shot Classification via Learned Feature-Wise Transformation," in 8th International Conference on Learning Representations (ICLR), 2020, pp. 1–16.

[14] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo, "Openml: networked science in machine learning," SIGKDD Explorations, vol. 15, no. 2, pp. 49–60, 2013. [Online]. Available: http://doi.acm.org/10.1145/2641190.264119

[15] S. Ravi and H. Larochelle, "Optimization as a Model for Few-Shot Learning," in 5th International Conference on Learning Representations (ICLR), 2017, pp. 1–11.

[16] J. Snell, K. Swersky, and R. Zemel, "Prototypical Networks for Few-shot Learning," in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, 2017, pp. 1–13.

[17] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," in 34th International Conference on Machine Learning, 2017, pp. 1126—1135.

[18] Z. Liu, A. Pavao, Z. Xu, S. Escalera, F. Ferreira, I. Guyon, S. Hong, F. Hutter, R. Ji, J. C. S. J. Junior, G. Li, M. Lindauer, Z. Luo, M. Madadi, T. Nierhoff, K. Niu, C. Pan, D. Stoll, S. Treguer, J. Wang, P. Wang, C. Wu, Y. Xiong, A. Zela, and Y. Zhang, "Winning solutions and post-challenge analyses of the chalearn autodl challenge 2019," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 9, pp. 3108–3125, 2021.

Appendix B: Example of fact sheet

This appendix provides a template of fact sheet, used in the Cross-Domain Meta-Learning Challenge. The filled out fact sheets are found on the website of the challenge [11].

11. https://metalearning.chalearn.org/

Cross-Domain MetaDL challenge fact sheet template

MetaDL organizing team

September 2022

This is the template for Cross-Domain MetaDL challenge[1] fact sheet. Please fill out the following sections carefully in a scientific writing style. Part of the filled fact sheet could be re-used as subsections in future publications.

Please edit this `.tex` file directly. Once finished, please zip the `.tex` file with all related files (e.g. generated PDF, `.bib` file) and send it to metalearningchallenge@googlegroups.com before **September 20, 2022**.

1 Team details

- **Team name *:**
- **Team website URL (if any):**
- **Team leader name *:**
- **Team affiliation *:**
- **Team leader address *:**
- **Team leader phone number *:**
- **Team leader email *:**
- **Name of other team members (if any)**
- **Username for Free-style league (if any):**
- **Username for Meta-learning league (if any):**
- **Publication list of each team member *:**
- **Gender of each team member *:**
- **Nationality of each team member *:**

[1] https://codalab.lisn.upsaclay.fr/competitions/3627

2 Contribution details

- **Title of the contribution *:**
- **Summary ***
  In a few sentences outline what makes you proud of your contribution. Less than 50 words.
- **Motivation**
  Describe what motivates your method design and justify the importance of the approach. We expect a comparison with previous work and a clear explanation of the advantages of this approach to related work. Figures and tables may be added to support the text if needed.
- **Contributions ***
  An itemized list of your main contributions and critical element of success. **Highlight key words in bold.**
  Suggestions: Contrast your proposed method with others e.g. in terms of computational or implementation complexity, parallelism, memory cost, theoretical grounding.
- **Detailed method description ***
  In this part, contributions must be expanded and explained in more details. The explanations must be self-contained and one must be able to reproduce the approach by reading this section. You can explain and justify the approach by any means, e.g. citations, equations, tables, algorithms, platforms and code libraries utilised, etc. We expect a detailed explanation of the architecture, preprocessing, loss function, training details, hyper-parameters, etc.
- **Representative image / workflow diagram of the method ***
  An image (or several images) to support better description of your method. You can refer to these images in the method description part above.
- **Code repository ***
  Link to a code repository with complete and detailed instructions so that the results obtained on Codalab can be reproduced locally. This is recommended for all participants and mandatory for winners to claim their prizes.

3 Technical details

In this sections, multiple questions are asked on the technical details of your method. Please fill out this Google Forms *: https://forms.gle/3dasBFMwKCXv4kjx5

REMEMBER: After you filled out the form, you will get a link for later modification. Please right click on "Edit your response" and copy the link address for later modification and put it below:

_____

References

Appendix C: Data leakage avoidance

Leakage

Leakage in machine learning refers to the inclusion of information during training that (a) is unavailable at prediction time for unseen data, and (b) artificially inflates model performance, thereby undermining generalizability. Leakage is often subtle, difficult to detect, and can severely compromise the integrity of machine learning competitions, sometimes rendering months of work unusable.

While competition organizers may sometimes be careless, leakage is often the result of its inherent complexity. Participants may also exploit leakage intentionally, driven by motivations such as monetary rewards or prestige, rather than advancing the field. Organizers must anticipate and address these issues proactively, as rules alone are insufficient deterrents. Preventing leakage requires careful preparation, including ensuring datasets are free from exploitable information and considering potential methods participants might use to gain an unfair advantage.

Leakage can occur in three main categories: access to ground truth, intrinsic properties of the data, and issues introduced during data processing. While these categories provide a framework for identifying leakage vectors, they are not exhaustive, emphasizing the importance of experience and diligence in designing robust competitions.

Access to Ground Truth

Ensuring the security of test set ground truth is critical for maintaining the integrity of machine learning competitions. Allowing participants to access these labels, even inadvertently, undermines the competition's validity. Obfuscating or making the data difficult to find is insufficient, as participants are often resourceful and capable of bypassing such measures.

Test data should be securely stored with access restricted to a limited, identifiable group of individuals who are excluded from participating in the competition. Additionally, care must be taken to prevent unintentional leakage, such as publishing graphs, descriptions, or prior content that reveals test labels. Credentials for accessing data should also not be publicly available, such as in repositories like GitHub. Proactive measures and thorough checks are essential to safeguard test data and prevent compromise.

Leakage Intrinsic to the Data

Leakage is a common issue in competition data, as it can arise in numerous subtle ways. For instance, in a challenge to classify images of cats and dogs, leakage might occur through timestamps if images of each category were collected at different times, or through metadata such as camera type or resolution, which may inadvertently predict the target. Even when metadata is stripped, features like image resolution or compression patterns can still reveal information, necessitating careful standardization of files.

A notable example of leakage involved a competition where a high score was achieved not by analyzing the data but by exploiting file attributes like size-on-disk and embedded timestamps. This underscores the importance of addressing both explicit and implicit leakage risks.

Leakage is especially challenging in fields like medical imaging, where variations in imaging equipment may introduce unavoidable biases. Organizers must carefully weigh the risks and benefits of including such data and decide which metadata to retain. Additionally, time series data poses

inherent leakage risks, as future predictions may inadvertently be influenced by temporal patterns in the training data.

Preliminary exploratory data analysis (EDA) and feature importance analysis can help identify potential leakage. However, organizers must ensure that any identified features are valid for use in future unseen predictions, as importance alone does not confirm leakage. Thorough consideration and proactive measures are essential to mitigate leakage risks effectively.

Leakage Introduced when Processing the Data

Data leakage can be inadvertently introduced during the data processing phase through various mechanisms. One common source is ordering; if files are saved in a sequence that reflects the underlying labels or temporal patterns (e.g., all cat images saved first, followed by dog images), participants may infer the target variable indirectly. To mitigate this, files should be randomized in a repeatable, deterministic manner using set seeds or random-state parameters to ensure consistent reproducibility while breaking any unintentional ordering patterns. Similarly, leakage can occur when saving files with metadata or filenames that embed information about the target variable, such as timestamps, file sizes, or other distinguishing attributes. Ensuring that such metadata is stripped and filenames are anonymized is critical.

Additionally, the generation of synthetic data introduces unique challenges; if synthetic data retains traces of its generation process, such as embedding distinct features or artifacts that correlate with the target labels, it may provide unintended predictive signals. Careful preprocessing, validation, and testing of synthetic data are essential to prevent these artifacts from compromising the integrity of the competition. Proactively addressing these processing-related risks ensures fair competition and more generalizable model outcomes.

References

Claire Adam-Bourdarios, Glen Cowan, Cécile Germain, Isabelle Guyon, Balázs Kégl, and David Rousseau. The higgs boson machine learning challenge. In *NIPS 2014 workshop on high-energy physics and machine learning*, pages 19–55. PMLR, 2015.

Charu C. Aggarwal. *Linear algebra and optimization for machine learning : a textbook / Charu C. Aggarwal.* Springer, Cham, 2020 - 2020.

K. Banachewicz and L. Massaron. *The Kaggle Book: Data Analysis and Machine Learning for Competitive Data Science*. Expert insight. Packt Publishing, 2022. ISBN 9781801817479. URL https://books.google.fr/books?id=Cy-nzgEACAAJ.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1006–1014, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/blum15.html.

A. Burkov. *The Hundred-Page Machine Learning Book*. Andriy Burkov, 2019. ISBN 9781999579517. URL https://books.google.com/books?id=0jbxwQEACAAJ.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), mar 2024. ISSN 2157-6904. doi: 10.1145/3641289. URL https://doi.org/10.1145/3641289.

Cjadams, Daniel Borkan, Inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. Jigsaw unintended bias in toxicity classification. https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification, 2019. Kaggle.

David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26 (4):745–766, 2017. doi: 10.1080/10618600.2017.1384734. URL https://doi.org/10.1080/10618600.2017.1384734.

Romain Egele, Julio C. S. Jacques Junior, Jan N. van Rijn, Isabelle Guyon, Xavier Baró, Albert Clapés, Prasanna Balaprakash, Sergio Escalera, Thomas Moeslund, and Jun Wan. Ai competitions and benchmarks: Dataset development, 2024. URL https://arxiv.org/abs/2404.09703.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018. URL http://arxiv.org/abs/1803.09010.

Isabelle Guyon, John Makhoul, Richard Schwartz, and Vladimir Vapnik. What size test set gives good error rate estimates? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (1):52–64, 1998.

Isabelle Guyon, Amir Saffari, Gideon Dror, and Gavin Cawley. Agnostic learning vs. prior knowledge challenge. In *2007 International Joint Conference on Neural Networks*, pages 829–834. IEEE, 2007.

Isabelle Guyon, Constantin Aliferis, Greg Cooper, André Elisseeff, Jean-Philippe Pellet, Peter Spirtes, and Alexander Statnikov. Design and analysis of the causation and prediction challenge. In *Causation and Prediction Challenge*, pages 1–33. PMLR, 2008.

Isabelle Guyon, Alexander Statnikov, and Berna Bakir Batu. *Cause effect pairs in machine learning*. Springer, 2019a.

Isabelle Guyon, Lisheng Sun-Hosoya, Marc Boullé, Hugo Jair Escalante, Sergio Escalera, Zhengying Liu, Damir Jajetic, Bisakha Ray, Mehreen Saeed, Michèle Sebag, et al. Analysis of the automl challenge series. *Automated Machine Learning*, 177, 2019b.

Frank Hutter. A Proposal for a New challenge Design Emphasizing Scientific Insights. Keynote presentation at NeurIPS Workshop on Challenges in Machine Learning, 2019.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data*, 6(4), dec 2012. ISSN 1556-4681. doi: 10.1145/2382577.2382579. URL https://doi.org/10.1145/2382577.2382579.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. Reduced, reused and recycled: The life of a dataset in machine learning research. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/3b8a614226a953a8cd9526fca6fe9ba5-Paper-round2.pdf.

Zhengying Liu. *Automated Deep Learning : Principles and Practice*. Theses, Université Paris-Saclay, November 2021. URL https://theses.hal.science/tel-03464519.

Antoine Marot, Benjamin Donnot, Camilo Romero, Balthazar Donon, Marvin Lerousseau, Luca Veyrin-Forrer, and Isabelle Guyon. Learning to run a power network challenge for training topology controllers. *Electric Power Systems Research*, 189:106635, 2020a.

Antoine Marot, Isabelle Guyon, Benjamin Donnot, Gabriel Dulac-Arnold, Patrick Panciatici, Mariette Awad, Aidan O'Sullivan, Adrian Kelly, and Zigfried Hampel-Arias. L2rpn: Learning to run a power network in a sustainable world neurips2020 challenge design. *Tutorial5/pdf/111. pdf*, 2020b.

Antoine Marot, Benjamin Donnot, Gabriel Dulac-Arnold, Adrian Kelly, Aidan O'Sullivan, Jan Viebahn, Mariette Awad, Isabelle Guyon, Patrick Panciatici, and Camilo Romero. Learning to run a power network challenge: a retrospective analysis. In *NeurIPS 2020 Competition and Demonstration Track*, pages 112–132. PMLR, 2021.

Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. Bias in data-driven artificial intelligence systems—an introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10(3):e1356, 2020. doi: https://doi.org/10.1002/widm.1356. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1356.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. CodaLab Competitions: An open source platform to organize scientific challenges. Technical report, Université Paris-Saclay, FRA., April 2022. URL https://inria.hal.science/hal-03629462.

Magali Richard, Yuna Blum, Justin Guinney, Gustavo Stolovitzky, and Adrien Pavão. Ai competitions and benchmarks, practical issues: Proposals, grant money, sponsors, prizes, dissemination, publicity, 2024. URL https://arxiv.org/abs/2401.04452.

Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/ee39e503b6bedf0c98c388b7e8589aca-Paper.pdf.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition, 2010.

Dataset Development

Romain Egele* ROMAINEGELE@GMAIL.COM
University Paris-Saclay, France, and Argonne National Laboratory, USA

Julio C. S. Jacques Junior† JULIO.SILVEIRA@UB.EDU
University of Barcelona and Computer Vision Center, Spain

Jan N. van Rijn J.N.VAN.RIJN@LIACS.LEIDENUNIV.NL
Leiden Institute of Advanced Computer Science (LIACS), Leiden University, the Netherlands

Isabelle Guyon GUYON@CHALEARN.ORG
University Paris-Saclay, France, ChaLearn, USA, and Google, USA

Xavier Baró XBARO@UB.EDU
University of Barcelona, Spain

Albert Clapés ACLAPES@UB.EDU
University of Barcelona and Computer Vision Center, Spain

Prasanna Balaprakash PBALAPRA@ORNL.GOV
Oak Ridge National Laboratory, USA

Sergio Escalera SESCALERA@UB.EDU
University of Barcelona and Computer Vision Center, Spain

Thomas Moeslund TBM@CREATE.AAU.DK
Aalborg University, Denmark

Jun Wan JUN.WAN@IA.AC.CN
MAIS, Institute of Automation, Chinese Academy of Sciences, China

Walter Reade INVERSION@GOOGLE.COM
Google, Kaggle, USA

Abstract

Machine learning is now used in many applications due to its ability to predict, generate, or discover patterns from large quantities of data. However, the process of collecting and transforming data for practical use is intricate. Even in today's digital era, where substantial data is generated daily, it is uncommon for it to be readily usable; most often, it necessitates meticulous manual data preparation. The haste in developing new models can frequently result in various shortcomings, potentially posing risks when deployed in real-world scenarios (e.g., social discrimination, critical failures), leading to the failure or substantial escalation of costs in AI-based projects. In this chapter, we propose a comprehensive framework for dataset development. The framework consists of several stages (i.e., requirements, design, implementation, evaluation, distribution, and maintenance), each consisting of a set of possible operators (e.g., data cleaning or data reduction). We describe the various operators in detail. Finally, we address practical considerations regarding dataset distribution and maintenance. While the framework is partially based on our experience, we aim to substantiate the steps with scientific references where possible.

Keywords: Data-centric machine learning, dataset development, data preparation

∗. These authors contributed equally to this work.

1 Introduction

In today's digital world, large amounts of data are generated daily in various domains. Machine learning methods can utilize this data to train AI models that address or automate various tasks. As machine learning is used widely in research and industry, following the wrong procedures in collecting and processing a dataset can lead to various downstream problems when models are being trained on this data, e.g., problems with privacy or fairness. In this chapter, we present a framework that aims to help develop a dataset in a more principled way and identify core actions to be performed for better management of such a project.

As mentioned by Hutchinson et al. (2021), dataset development is not a linear process that has all detailed specifications from the start. It can be structured using an agile[1] (Chin, 2004) management methodology with core components interacting with each other as well as evolving iteratively. Figure 1 presents the framework that we propose. It is structured as a representative cycle of dataset development. One cycle is composed of five components: the requirements analysis involves the principal stakeholders and consists of defining the needs of the developed dataset; the design involves the domain expert and consists of determining how to structure the dataset and its implementation; the implementation involves data creators (e.g., data/software engineers, labellers) and consists in collecting and transforming the data to be usable; the evaluation involves data scientists and adversarial testers, consists in assessing the quality of the developed dataset concerning its requirements; the distribution and maintenance involve regulation, storage, and network experts and consist of defining the storage and accessibility of the dataset.

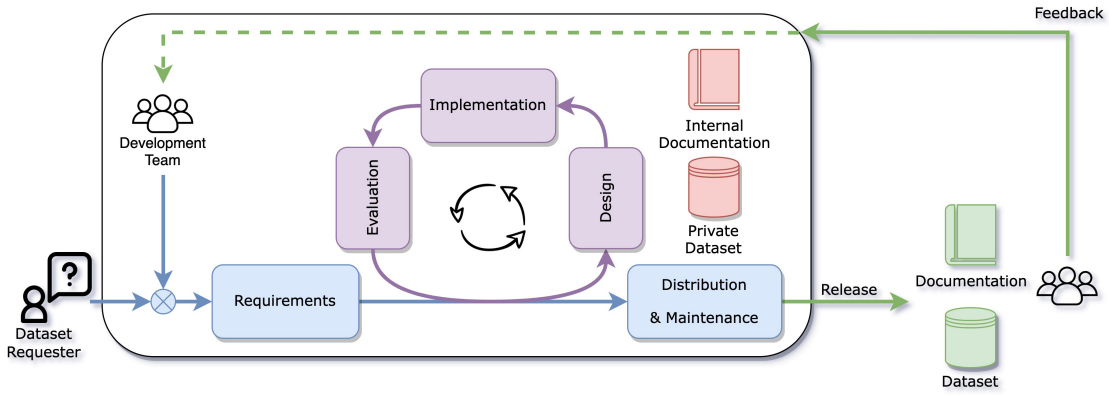


Figure 1: The dataset development cycle.

During this cycle, different aspects need to be documented. The dataset development team should keep track of internal (in the sense that it is meant to be used by the development team) documentation of common assumptions made when developing the dataset (e.g., explaining why the collected data are representative of the target population) as well as the tools or processes designed to acquire the data (e.g., survey, software). In addition, the development team should maintain public documentation explaining the content of the

1. Agile is a common term from software development.

dataset, its purpose, and its technical usage (e.g., software to install, interface to retrieve data samples or metadata of features). This public documentation differs from internal documentation because it could contain fewer details, and sensitive information (such as personal identifiers) should be omitted. Typically, both private (raw) versions of the data and public releases of the data exist.

The framework we present aims to support the development of datasets used in machine learning. While the dataset development cycle consists of many stakeholders and roles, a small group of persons can cover multiple roles in smaller dataset development projects. We substantiated our claims with scientific literature when possible. However, some paragraphs could not be matched with such references but are the results of our personal experience (e.g., data challenges).

2 Documentation

We first review various standards of building documentation (public or private) of the development process and the final produced dataset. Documentation is essential and spans every aspect of the dataset development in order to have better transparency and traceability, which will improve trust and safety. Documentation must be performed with reflexivity, which means its goal is to clarify or uncover obscure discretionary decisions (e.g., power dynamics, differences of perception) in order to understand their effect on the produced dataset (Miceli et al., 2021). In an effort to systematize the documentation of datasets, different initiatives emerged such as new methods (Gebru et al., 2021; Bender and Friedman, 2018; Holland et al., 2018), best practices[2], and formats[3]. The main aspects addressed in dataset documentation include its purpose, composition, collection process, preprocessing, intended usage, distribution, and maintenance (Wilkinson et al., 2016).

Several standards for dataset documentation have been proposed.

Datasheets for datasets: Gebru et al. (2021) proposes specific example problems to document around the core components of the dataset development. It encourages every dataset to be accompanied by documentation of its motivation (requirements), composition (design), collection process (implementation), recommended uses, distribution, and maintenance. The NeurIPS Dataset and Benchmark track adopted these guidelines as well. Similarly, the DC-Check (DC standing for data-centric) framework was released (Seedat et al., 2022) to spawn a broader range of data-centric related tasks on a general range of applications. Bender and Friedman (2018) propose similar guidelines called "data statements", with a focus on natural language processing applications.

FAIR principles: Wilkinson et al. (2016) focus on distribution and maintenance aspects and provide guidelines to improve the findability, accessibility, interoperability, and reuse of datasets. This initiative aims to standardize digital practices around dataset distribution and maintenance, improving the reusability of both software (e.g., data loaders, due to the use of open standards) and datasets.

2. Metadata and data documentation: tinyurl.com/5j5ynu6p
3. AutoML format: tinyurl.com/yc5uk4xh

Dataset nutrition label: Holland et al. (2018) follow the idea of nutrition facts labels to help create a summarized diagnosis of the quality of a dataset. It comprises diverse qualitative and quantitative modules generated through checklists or multiple statistical analyses of the dataset and displayed in a standard fashion.

However, despite the good intentions of standard documentation practices, they often must be refined for the specificities of the target use, and some attributes may not be allowed to be disclosed or stored due to data regulation (e.g., race, belief), which therefore prohibits public verification of some statistical properties.

Use Case: Facial Emotion Recognition

We give an example of potential aspects to be documented when developing a video dataset for facial emotion recognition (Corneanu et al., 2016). In this context, videotapes are recorded from a group of participants.

- *What is the dataset intended to be used for (e.g., for which application and for which population)? –* ***Requirements***
- *How are participants recruited for video recording (e.g., gender, age, hairstyle, use of glasses)? –* ***Design***
- *How is the privacy of recorded participants managed? –* ***Design***
- *For the recording protocol: Are participants following a script during recordings? Are participants stimulated by a particular incentive during recording? –* ***Design***
- *For the annotation protocol: How are the annotations defined (e.g., categorical, discrete, or real values)? How are annotators selected (e.g., gender, age, etc.)? How are data to be annotated defined (e.g., per frame, per video segment, with or without sound)? How many annotators observe each data? –* ***Design***
- *How are acquired data transformed (e.g., calibration)? How are final labels computed? What is the definition of frames and sample rate? –* ***Implementation***
- *For the investigation of social bias one could consider documenting some sensitive information regarding participants and annotators (e.g., gender, age, race) while being cautious to respect data regulation, privacy, and consent. What are the possible biases from the selected populations (participants, annotators)? How is the annotators' agreement evaluated (e.g., metric), and what is its value? –* ***Evaluation***
- *Where are the data hosted? How can the data be accessed? –* ***Distribution*** *&* ***Maintenance***

Finally, we give some practical advantages of good dataset documentation. The dataset development team can benefit from the following aspects: a better management of the dataset development by an improved understanding of why, how, and what is done to produce the dataset; a better traceability of possible bugs or flaws; an improved reusability of developed tools due to clear documentation and principled development; a reduced presence of flaws in the produced dataset; a better dataset quality. On the other hand, the dataset users can benefit from the following aspects: an improved usability of the dataset; an improved understanding and trust of the dataset; a better quality of machine learning models; and a better reporting of possible flaws discovered in the data.

Nevertheless, while documentation helps to improve dataset quality and, therefore, the quality of machine learning models using them, some limitations still exist. Companies often regard some of the information that could or should be documented as confidential, especially if it involves details about the intended product or if some of the processes involved in producing the dataset are a strategic advantage (Miceli et al., 2021). For this reason, we differentiate private (required for audits) and public documentation (required for the user). Documentation is often seen as time-consuming work that is likely to delay the completion of other tasks that are perceived as more important. It is often perceived as an optional, nice-to-have but not must-have component, and therefore, in such cases implemented last (Miceli et al., 2021). Lastly, producing complete but synthetic and clear documentation is challenging. The documentation format may vary for the different stakeholders (engineers, statisticians, business analysts) and create redundancy (pdf document, book, website). We encourage dataset development teams to use tools such as Sphinx[4], ReadTheDocs[5] and Pandoc[6] to automate the build of documentation and navigate between formats.

3 Requirements

The requirements analysis is the first step of the dataset development cycle (see Figure 1), where the dataset requester (representing the main stakeholder requiring the dataset) and the dataset development team (representing who is in charge of producing the dataset) meet to define the requirements. During this phase, the following topics can be addressed:

1. Why is the dataset needed?
    - **Application scenario:** What are the intended purposes and use cases?
    - **Machine learning tasks:** What type of machine learning techniques (e.g., supervised, unsupervised, reinforcement learning) is planned to be used? How will tasks be carved out of data?
    - **Users:** Which group of users do we expect to use the dataset?
2. How is the dataset developed?
    - **Prior work:** Are there already existing datasets filling this need (see Section 4.3 and 5.1.1 about potential sources of already collected data)? Will collecting (more) data solve the problem at hand or help in understanding it better?
    - **Method:** What dataset development protocol is planned to be used?
    - **Ethics:** Are the intended purposes ethical? What are possible fairness and privacy issues, and how will they be evaluated? Is collecting such data considered experimenting on human subjects?
    - **Risks:** What adverse usage could be done from the dataset? Are the risks worth the trouble? How can the risks be reduced?

4. www.sphinx-doc.org
5. readthedocs.org
6. pandoc.org

- **Constraints:** What are the anticipated difficulties limiting the development of the dataset (e.g., recruitment of subjects)? In the case of data involving human subjects or a legal entity, it is crucial to follow regulations from governmental regulations such as GDPR (Voigt and Bussche, 2017). Sometimes, the data needs to be anonymized, cannot be stored anywhere (risk of data leaks), and its legal framework (e.g., application, lifespan) needs to be defined before acquisition.
- **Development team:** Who is composing the dataset development team? Who will lead the effort? How are the roles distributed (design, implementation, evaluation, distribution, and maintenance)?
- **Resources:** How many resources will be required to complete the whole dataset development cycle, including compensation of staff, recruitment of volunteers, payment of annotation services, computational resources, etc.? Also, the environmental impact (CO2 emissions) should be considered.

3. What is the dataset expected to be?

- **Content:** What information does the dataset needs to contain (e.g., features, annotations, metadata)?
- **Baseline:** What baseline modeling methods will be used to evaluate the quality of the dataset? What evaluation measures will be used, including utility, fairness, and privacy?
- **Ownership:** Who will own the rights? Who will be legally responsible?
- **Storage:** Where does the dataset need to be hosted?
- **Distribution:** How is the dataset going to be accessed (e.g., through a web API)?

4. When is the dataset expected to be delivered?

When designing a dataset, one should always keep its envisioned purpose in mind. Also, the interplay between the dataset and the machine learning method is essential. A dataset represents a snapshot of the real world used to train (and possibly evaluate) a learning algorithm on a particular task. It is essential for quality assurance to involve baseline modeling methods and their performance evaluation early in the dataset development process. If the utility, fairness, ethical concerns and privacy of the data cannot be assessed in the context of an actual learning task, the dataset will likely be useless.

Regarding ethical considerations, having a committee that includes diverse members in terms of competence, demographics, and cultural backgrounds is advisable. The committee should include persons competent in the target application area, machine learning/data science, and persons representative of the subject and target population (if human subjects are involved in data collection or are affected by data usage). Additionally, it is advisable to include an ethics expert and a law expert. It is advisable to ensure that at least one member is not affiliated with the organization creating the dataset and that no member has

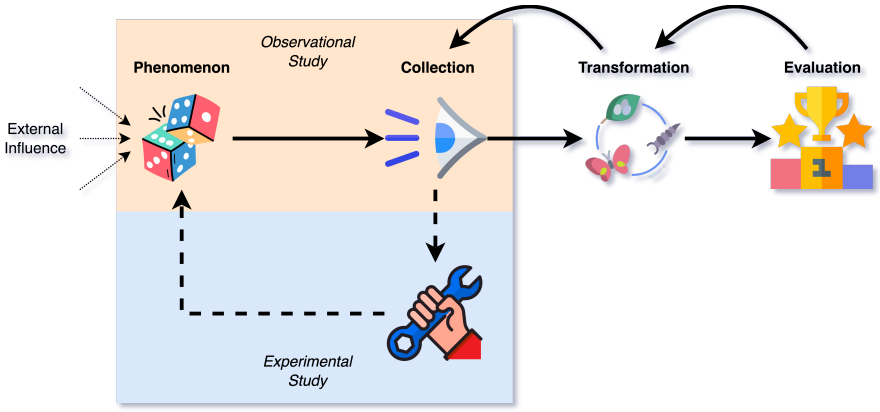


Figure 2: Illustration of observational vs. experimental studies.

a conflict of interest. In the US, such committees are called Institutional Review Board[7] and can be registered officially. Also, some general guidelines[8,9] can be followed.

4 Design

The design phase is about defining more precisely what the dataset should contain and how it will be implemented (i.e., collected and transformed), evaluated, distributed, and maintained. The complexity of implementation design, which comprises data collection and transformation, is directly related to the decision between creating a dataset from scratch or reusing, repurposing, and recycling existing data. For many use cases, processes describing dataset development already exist and are published, such as at the NeurIPS Dataset and Benchmark track[10]. We recommend exploring this literature to look for dataset development-specific methods.

The design of a dataset is often time-intensive while being crucial to innovation. Of course, designing a dataset results directly from the definition of requirements, but it is also possible to update the requirements based on insights from the design step. In fact, any critical aspect missed at the requirements and design stages may compromise the whole dataset development process. Even though these steps should be designed thoroughly, not everything can be foreseen, and flaws will likely be discovered later. For instance, consider the case where data are planned to be collected through an online survey; the development team could forget to ask participants to sign an agreement. Such an agreement can include the rights to process and use the data or to transfer the rights (including copyright). The collected data cannot be used without such an explicit agreement with the participants. Therefore, we encourage the development team to bootstrap the whole development cycle on a small scale to help refine requirements and design. As a result, the dataset development should not be a linear closed loop but an iterative and interactive process.

7. Institutional Review Board: https://tinyurl.com/mvpd292w
8. Ethics Guidelines for Trustworthy AI: https://tinyurl.com/2zc7hfdz
9. Recommendation on the Ethics of Artificial Intelligence: https://tinyurl.com/3ew8xp4e
10. NeurIPS Datasets and Benchmarks track: https://tinyurl.com/37u4cbx3

When developing a dataset, either observational (i.e., the variables of the phenomenon cannot be controlled) or experimental (i.e., the variables of the phenomenon can be controlled) data can be used (Figure 2). Similarly, data can be collected *de novo* (i.e., from scratch) or from existing sources (i.e., reuse, repurpose, and recycle).

4.1 Data Leakage

Before diving into possible designs, we emphasize the critical importance of vigilance against data leakage, which can undermine the integrity of the evaluation process. Unlike typical software bugs, data leakage can irreparably compromise months of model development. Given the myriad ways data leakage can occur, it is prudent to assume that the initial dataset may contain some form of leakage introduced during its development (Figure 1).

Participants in benchmarks and competitions often have diverse motivations beyond scientific advancement, such as monetary incentives and prestige. Some may adopt a "hacker" mindset, prioritizing circumvention over problem-solving, which necessitates the establishment of clear competition guidelines to discourage cheating.

Data leakage generally refers to the inclusion of illegitimate information in the dataset used for model development and selection (Kaufman et al., 2012). Illegitimate information pertains to data that will not be available once the model is deployed in its final environment. This leads to biased model selection and frequently results in overestimating generalization performance. It also affects the ranking of compared models in unexpected ways; for instance, the selected best model may exhibit outstanding performance in the training environment, but it could perform mediocrely in the final environment.

In this chapter, we will not address "adversarial" aspects of data leakage, such as deliberately using the ground truth targets of a test set for training, as it relates more to software security. Simply put, we focus on issues that are difficult to detect, even for someone willing to avoid leakage and who wants to follow the spirit of the task associated with the dataset.

Use Case: Leakage while collecting pictures of animals

Assume a team is organizing a challenge to predict whether an image depicts a cat or a dog. They deliver the data in two folders: one containing all cat images and the other containing all dog images. What types of data leakage should be considered?

Timestamps of the images could leak information if the data were collected on different days. For example, if the person collecting the data spent one week photographing cats and the following week photographing dogs, or if they downloaded all the cat images first, followed by the dog images. It is a good practice to randomize (in a repeatable and deterministic manner) how the files are stored.

If new photographs were taken specifically for the competition, different individuals might have used different cameras preferentially for cats or dogs. Consequently, metadata indicating the camera type (either through actual metadata embedded in the image or a proxy such as resolution and color balancing) could predict the target, thus constituting leakage. Tools like EXIF viewer[11] *can be used to inspect image metadata.*

It is generally advisable to strip files of any associated metadata, especially with images, but more is needed. A camera model can often be inferred from a raw image without using embedded metadata through features such as resolution or the specific way the JPEG is created from the camera sensor data.

11. EXIF Viewer: https://exif.tools/

Therefore, images should be presented in a standardized format that minimizes the possibility of such leakage, or cameras should be randomized to take the pictures.

The situation becomes even more complex in scenarios such as medical imaging competitions, where it may be infeasible to eliminate the effects of varying imaging equipment (e.g., scanners). In these cases, conducting a thorough risk/benefit analysis is crucial to determine whether to include data from different imaging equipment and what metadata to incorporate. In addition, proper model analysis will need to be conducted to ensure that it did not simply learn an artefact of the sensor.

4.2 De Novo Data

This section describes design aspects for *de novo* data. The first setting to consider is which variables of the target phenomenon can be controlled or not and in which quantity they appear (e.g., difference between pixels of images and tabular dataset). We distinguish between an experimental case (where variables can be controlled) and an observational case (where variables can not be controlled). In the experimental case, when only a few variables are available, one can decide to discretize them and explore all possible combinations. When more variables are available, one often resorts to random sampling. In the observational case, even though variables cannot be controlled and impacting factors are often intractable (e.g., describing a patient in healthcare), random sampling is also advocated (random trials) to vary these factors. In this case, it is essential to check that the observed population is randomly sampled and not biased (e.g., selection bias). Other considerations for the dataset design are:

Data quantity: How many samples should the dataset have? An example of character recognition is proposed in (Guyon et al., 1998). However, it may be challenging to have such information beforehand. It is possible to refine the quantity needed by involving the modeling process during development (i.e., baseline in the evaluation step). A default choice is to collect as much data as possible and leave the choice of quantity to the user. An other possibility is to use the Hoeffding bound[12] (Burges, 1998; Bousquet et al., 2003) to estimate the number of samples needed to reach a certain level of confidence in estimating the generalization of a model. Finally, learning curves of machine learning algorithms (Mohr and van Rijn, 2022, 2023) can also be used to extrapolate the quantity of samples needed in order to reach a certain accuracy level.

Data balancing: How to make sure that you have enough samples of each group that should be represented? For instance, depending on the application, one may want to balance groups by gender, age, or educational background. It is advised to pay attention to the possible need to consider cross-sectional groups' gender $\times$ age $\times$ educational background. Unfortunately, the larger the number of grouping factors to be considered, the larger the number of samples to be collected to adhere to a minimal group size.

Data annotations: Are labels required? Do we need more advanced annotations, such as bounding boxes around the subject? If yes, how is the labeling process operated? For

12. Learning Theory, J. Domke: https://tinyurl.com/yc2k99w4

example, this can be achieved through crowd-sourcing, citizen science, or commercial parties. When making use of such services, one should always carefully check the conditions under which these labeling operations are being performed (e.g., are the persons performing the labeling doing this under fair work conditions) and whether this process can be (semi-)automated.

Data representation: How are data represented? Many data structures exist to represent data, it can be tables, images, videos, text files, and graphs. The data can be compressed or not. The data can be accessed incrementally or all at once (we refer the reader to the HDF5 open format and library[13]). Can we provide a data reader? If tabular data are collected, what are the features to collect?

Metadata: What metadata can be collected (e.g., date of recording, operator name, temperature)? Generally, the more metadata, the better. Metadata can help identify bias or spurious correlations (potentially resulting in data leakage). The metadata explains the context of generated data, and therefore it can help determine if the data was well collected in diverse settings (e.g., at different times or temperatures). Also, metadata should not be correlated with the predicted variable. If this is the case, then some spurious correlation can be identified and resolved by modification of the dataset creation process.

4.3 Reusing, Repurposing, and Recycling Data

As discussed before, *de novo* dataset development can be time-intensive. However, many datasets are now publicly accessible (Koch et al., 2021) with a license allowing to use them free of charge (e.g., Creative Common[14]). In addition, search engines (a short list is available at the end of this section) can help find datasets corresponding to specific needs. Therefore, before performing *de novo* data collection, it is essential to investigate such opportunities, which can help save considerable resources. We refer to such methods as data reusing (analogously to reusing a plastic bottle of water by refilling it), data repurposing (analogously to repurposing a plastic bottle of water to collect leaking water from a pipe) and data recycling (analogously to breaking down the plastic bottle of water to make plastic boxes). These concepts are ordered from fewer to more modifications applied to the pre-existing data. However, the boundaries among them are not clear and can overlap.

On the one hand, data reusing is the practice of directly reusing data (i.e., without modification, including its purpose) when the use case is similar, and the required information is already available. Hence, we talk about data reusing when the initial product (i.e., the data) does not need to be transformed before being ingested and the intended purpose remains the same.

In other cases, one may reuse the same data set but repurpose it from being used in research to commercial applications. In this case, many ethical and privacy concerns must be revisited, among other aspects.

When leveraging existing data, it is important to be careful with possible deprecation, bias, and retirement of the source data. A dataset becomes deprecated when it is still

13. https://www.hdfgroup.org/solutions/hdf5
14. https://creativecommons.org

publicly accessible, but for some reason (e.g., social bias), it should not be used in practice. An example of a deprecated dataset is the Boston house prices dataset (Harrison Jr and Rubinfeld, 1978) is kept public for scientific traceability, reproducibility, and social bias study but discouraged from being used otherwise. In this case, the deprecation was due to social bias in the data. Similarly, some datasets can be retired (i.e., removed from public access) such as the Tiny Images dataset[15] (Torralba et al., 2008) due to the presence of derogatory terms as categories and offensive images.

In some cases, light data transformations are required to enable data repurposing. For example, a re-annotation process may be performed if the to-be-predicted target variable changes. In other cases, pre-existing data can be completed by new samples and features. For instance, the "First Impressions V2" (Escalante et al., 2017) dataset is a typical example of repurposing. The original "First Impressions" (Ponce-López et al., 2016) dataset was annotated with Big-Five personality traits with the purpose of analyzing personality from audio-visual data. In version V2, the audio-visual data was kept the same. However, new labels, such as an additional interview variable and transcriptions, were included to enable research on explainable machine learning.

Finally, one can choose to recycle a data set. Data recycling leverages existing material with the possibility of reshaping it entirely. In this case, the dataset's purpose can further differ from its initial purpose. Therefore, it has to be verified that it is allowed by the dataset's license. While data recycling typically requires less effort than *de novo* collection, these additional checks still incur an additional workload not to be underestimated. An example of recycling is the creation of an image dataset with a single face per image per sample given a pre-existing image dataset containing single or multiple people on each image with full or partial bodies (Agustsson et al., 2017).

Examples of Datasets Search Engines and Providers are Dataset Search from Google, Kaggle, OpenML.org (Vanschoren et al., 2014; Bischl et al., 2021), UCI Machine Learning Repository, Hugging Face Datasets, and the NeurIPS: Datasets and Benchmarks track.

5 Implementation

This section focuses on the implementation of the dataset development. We consider this stage to be a set of processes; Figure 3 overviews these. This figure categorizes all processes into two categories: Collection processes (blue) take as input a design (Section 4) and as output a dataset (of an arbitrary size). Transformation processes (yellow) require a dataset as input and will output a transformed version of this dataset. Some processes (green) fall into both categories.

These processes are the functions that will produce a dataset on which machine learning models are trained. Data **collection** (Section 5.1) entails the *gathering*, *acquisition*, *synthesis/generation*, and *annotation* of data. To avoid any confusion with other works that interchangeably use these words, in our chapter we called *collection* the larger class that includes all the others.

Data **transformation** (Section 5.2) includes *cleaning*, *reduction*, *representation*, and *normalization/calibration* of data. The tasks of data *integration/fusion*, and *augmentation*

15. Tiny Images dataset: https://tinyurl.com/2vfa4xve

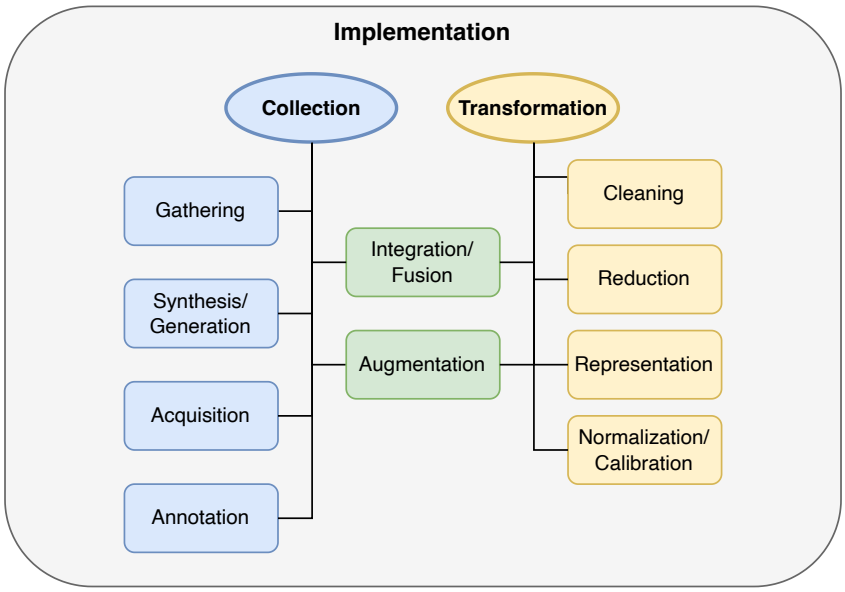


Figure 3: Categorization of sub-tasks included in data collection and transformation. Collection operators (blue) take as input a design and as output a dataset (of an arbitrary size). Transformation operators (yellow) require a dataset as input and will output a transformed version of this dataset. Some operators (green) fall into both categories. A typical data development process combines several operators into a pipeline, always starting with a collection operator.

can be categorized as both collection and transformation operators; therefore, we link them to both in Figure 3.

The implementation phase typically contains various sub-processes defined by the design. An example of such a pipeline of processes is visualized in Figure 4, in which a pipeline combines five sub-processes.

Note that data evaluation can also be considered in the case of a dynamic data collection process, for example, to keep collecting/augmenting data until a pre-defined objective is satisfied. Section 6 details on these evaluation processes. Next, we discuss each part (collection and transformation) in more detail.

5.1 Data Collection

Data collection is the set of processes that can gather, generate, or measure information in order to create a dataset. Therefore, data collection includes data gathering, data synthesis, data acquisition, and data annotation.

Any of these processes can be performed in an observational or experimental setting. In the observational setting, the investigator responsible for data collection does not interfere with the phenomenon. The distribution of collected samples is supposed to reflect the natural distribution of data. For example, a naturalist studying wildlife may set up a camera

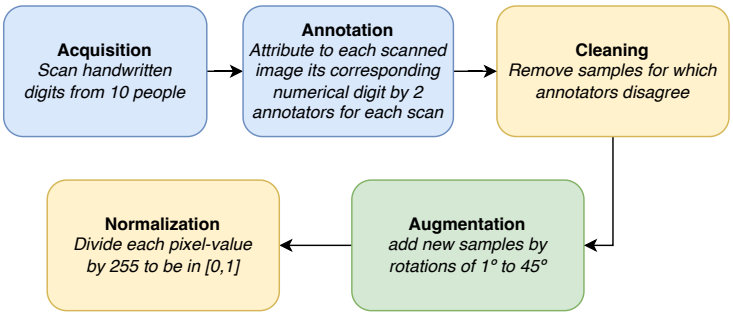


Figure 4: An example flow chart diagram of a data implementation pipeline that creates a dataset for handwritten digits classification.

trap in a forest to take pictures of animals living there. In contrast, in the experimental setting, the investigator may interfere with the phenomenon to achieve desired effects. The distribution of samples collected follows an experimental plan or design. For instance, a pet food company may want to study the influence of certain dog foods on certain dog breeds and conduct a trial, assigning different regimens to dogs of various breeds and then evaluating their energy by videotaping them. There also exists in-between cases in which data are observational, but the investigator samples data and features in an active way (Settles, 2009). For example, a photographer who gathers data by going on a photo safari may use aesthetic criteria to make their shots.

When collecting data, metadata is often available or can be created to describe processed data. It is essential to save as much metadata as possible to perform better data evaluation later.

5.1.1 Gathering

Data gathering is the set of processes by which data are brought together from sources where the data is already stored digitally, and the acquisition (Section 5.1.3) process cannot be influenced (see, e.g., Ullah et al. (2022)). The emergence of the modern web, where one can quickly access a massive amount of public data, has made the cost relatively low. These techniques suffer from high noise (e.g., picture wrongly tagged and therefore wrongly returned by a search engine) as well as ethical and regulatory limitations (e.g., privacy, copyrights, license).

An obvious way of gathering data is by using a search engine (e.g., Google Image, Bing), in which case it is critical to comply with the regulations. Another way is through web scraping, which is the practice of automating the search and downloading of data from the Internet. It provides more granularity to develop the gathering process. Tools such as Scrapy[16] can help to set up such a process. However, it must be performed responsibly, such as by following robot policies[17], which in some cases restrict access to robots.

16. https://scrapy.org
17. How to write and submit a robots.txt file: http://tinyurl.com/a9tvx6n8

Last, the gathering can be performed through crowd-sourcing (Roh et al., 2021; Garcia-Molina et al., 2016) where human workers are generally given micro tasks to gather bits of data that collectively become the generated dataset. More generally, it can be defined as the process of obtaining needed services, ideas, or content by soliciting contributions from a large group of people (Zhang et al., 2016). While crowd-sourcing can be applied both to data gathering and annotation, the methodologies applied differ, and its application to data annotation is detailed in Section 5.1.4. Crowd-sourcing to gather data can be performed implicitly or explicitly (Garcia-Molina et al., 2016). It is implicit when people are unaware of it, such as through website analytics. For example, by watching a movie, an individual provides data about the popularity of this movie. Then, the gathering is explicit when subjects get a request for information. The machine learning community has been using crowd-sourcing as a means to gather massive amounts of data, which can be considered a prominent way to outsource work to people reachable online. One of the most popular platforms for crowd-sourcing in machine learning at the time of writing is Amazon Mechanical Turk[18], where tasks are assigned to remote workers, which are compensated when the task is completed (Roh et al., 2021). Of course, the dataset collector also has the responsibility of checking that the workers who perform this labeling are doing so under ethical work conditions.

5.1.2 Synthesis and Generation

Data synthesis, also known as synthetic data generation, is the process of generating artificial data that mimics real-world observations and can be used to (pre)train machine learning models when actual data is difficult or expensive to get. It has become an attractive field of study because of data-hungry technologies such as deep learning (Nikolenko, 2021). Synthetic data can be generated procedurally (Queiroz et al., 2010; Wood et al., 2021) (i.e., predefined set of rules), through simulations (Dosovitskiy et al., 2017) or using generative models (e.g., generative adversarial networks (Karras et al., 2019)). One of the main motivations behind its use is the reduced cost of creating larger datasets where factors of variability (i.e., parameters of the synthetic generator) can be manipulated on demand. In addition, the generated data can often be directly annotated by leveraging the parameters of the data generator or using post-processing techniques. Other advantages include the possibility of better-controlling privacy (Yale et al., 2020; Kuppa et al., 2021) and fairness (Bhanot et al., 2021, 2022). However, one of the main limitations of synthetic data is that simulated phenomena often need to represent real-world scenarios closely, which is hard to achieve. Another challenge is related to transfer learning (Weiss et al., 2016) and domain adaptation (Ajith and Gopakumar, 2023), since a predictor trained on synthetic data (source domain) must generalize well to real-world data (target domain). Hence, the data used to eventually train the predictor implicitly should have similarities with the target data on which the trained model is deployed. Artificial data (available in large quantities) are frequently used to pre-train the model, which is later fine-tuned to real-world data (available in smaller quantities) before deployment.

The level of abstraction and realism can vary depending on the application domain, context, and needs. For example, in the case of synthetic data generated through simulations

18. https://www.mturk.com

and computer graphics applied to autonomous driving cars (Dosovitskiy et al., 2017), the level of abstraction and realism can be associated with the rendering quality, but also with respect to the behavior of the different simulated agents and phenomena (e.g., interaction between agents, traffic, weather, etc.). Each has a particular impact on the outcomes if the data are used for training a machine learning method. The associated costs and resources (e.g., data experts, designers, software engineers, etc.) required to achieve the desired level of abstraction and the trustworthiness of the generated data are additional barriers that might limit broader usability to train and evaluate machine learning models.

Synthetic data is also reported in the literature to perform data augmentation, combining synthetic and real-world data. However, a recent study on face analysis proposed to train their models with synthetic data only (Wood et al., 2021), opening up new approaches to better address fairness and privacy.

The synthetic data generation process can introduce artifacts, including some that may leak ground truth information. An example of this occurred during the SETI Breakthrough Listen[19] challenge. In this competition, the processing date of the files (the timestamps) leaked information about the ground truth target.

5.1.3 Acquisition

Data acquisition is the process that converts a real-world signal into a digital representation (Emilio, 2013). A basic example is the acquisition of temperature data through a thermometer. Some techniques such as quantization (Gray and Neuhoff, 1998), signal sampling (Higgins, 1996) or real number encoding [20], which have already been studied in depth in the information theory literature for physical signals, are often used for this purpose.

Performing data acquisition usually requires a well-designed experimental protocol (Section 4). In addition, it is always important to record additional metadata. Some basic metadata are the measurement time, location, and model of the device used for acquisition. Metadata can be descriptive, such as an overall description of the dataset and variables (e.g., units of physical quantities, and preferably following international standards). They can also contain copyrights or license terms. In general, metadata is to be understood as data used to describe the dataset to maintain tractability about how the data were acquired. In many cases, they can become the dataset of another dataset, for example, in the case of integration/fusion. In fact, metadata play an important role in identifying bias.

5.1.4 Annotation

Data annotation is the process of mapping existing data samples to other data. It is often performed in the context of supervised learning, which includes two principal variants: classification and regression. In classification tasks, the goal is to train a model that returns one of many possible classes for each sample. In this context, data annotation is referred to as data labeling. In regression tasks, the goal is to train a model that predicts a real number given a sample. Therefore, annotating with continuous variables is much more complex than a set of fixed classes (Roh et al., 2021), which can explain why data annotation research has primarily been focused on data labeling for classification. Other types of supervised tasks

19. SETI Breakthrough Listen Kaggle challenge: https://tinyurl.com/mrv9vtrk
20. https://ieeexplore.ieee.org/document/8766229

exist, such as: (i) describing the title of images (image captioning), (ii) labeling the style of a music record (classification), (iii) predicting the age of an individual (regression), (iv) rating an Amazon product by a score (categorical regression), or (v) drawing a bounding-box on an object in an image (object detection).

Roh et al. (2021) propose the following categories for understanding the data labeling landscape: crowd-based labeling (e.g., via crowd-sourcing or active learning (Tang et al., 2021)), automatically labeling from existing labels (e.g., through semi-supervised learning (van Engelen and Hoos, 2020)), and the use of weak labels (Ratner et al., 2017) (i.e., generating imperfect labels, but in large quantities to compensate for the lower quality).

First, crowd-sourcing techniques (Zhang et al., 2016) are generally focused on running tasks with many workers who are not necessarily experts (either on labeling or on any particular task). Therefore, different solutions have been proposed to collect more accurate and trusted labels (e.g., see (Nowak and Rüger, 2010; Maroto and Ortega, 2018)). In this line, a general procedure for controlling and ensuring the quality of data labeling is to have multiple workers annotate the same sample so that an agreement level can be computed. This way, any bias the workers may have can be identified and mitigated, with the cost of increasing time and resources. However, it does not necessarily include human perception bias, which is much more challenging to identify and mitigate (discussed in Section 6.4). Although various inter-annotator agreement measures exist (Checco et al., 2017) for simple categorical and ordinal labeling tasks, relatively little work has considered more complex labeling tasks, such as structured, multi-object, and free-text annotations (Braylan et al., 2022). Providing adequate instructions and the proper labeling interface is also a critical success factor (Roh et al., 2021).

Then, active learning focuses on iteratively selecting the most informative (according to some pre-defined measure) unlabeled examples for the model to reduce the need for human labor, which can then be outsourced or crowd-sourced (Roh et al., 2021). The workers are expected to be accurate; thus, the key challenge is to choose the proper examples given a limited budget. In addition, semi-supervised learning can complement active learning (Gu et al., 2014; Camargo et al., 2020) by finding the predictions with the highest confidence and adding them to the labeled examples (as pseudo-labels). In contrast, active learning can identify the predictions with the lowest confidence and send them for manual labeling.

Finally, weakly supervised learning (Zhou, 2018; Zhang et al., 2022) can also reduce the amount of human labor to annotate the training samples. This approach is beneficial when there are large amounts of data, and manual labeling becomes infeasible (Roh et al., 2021). In weakly supervised learning, it is possible to automate the labeling process by defining a set of labeling functions (Ratner et al., 2017), hand-crafted rule-based classifiers. An example from the Snorkel tutorial[21] is a labeling function to sort emails as "spam", using simply the presence of "http" in text metadata. A labeling function can leverage metadata or classifiers trained previously on similar tasks. Using multiple labeling functions helps to obtain a label score, which can be interpreted as a label probability, serving as weak supervision.

In conclusion, data annotation can be a time-consuming and expensive process, with many challenges (Rasmussen et al., 2022). However, quantity in some types of machine

21. https://www.snorkel.org/use-cases/01-spam-tutorial

learning, such as deep neural networks, is a key to success. Therefore, much research is trying to alleviate this limitation by learning representations from unlabeled data, which is later discussed in Section 5.2.4.

5.2 Data Transformation

Once the data collection process has provided a set of initial raw data, different transformation techniques can be used to make the data suitable for a particular machine-learning model. This section presents a brief overview and discussion around distinct aspects of data transformation, which includes data integration or fusion, cleaning, reduction, representation, normalization or calibration, and augmentation.

Without considerable care, data leakage can be introduced during the transformation process for a competition. Before discussing the various types of transformations, we outline several ways information can be inadvertently leaked, urging dataset developers (and challenge organizers) to remain vigilant.

The ordering of observations (e.g., how the data is sorted) should not reveal any information about the targets. In our toy example of cats and dogs (mentioned in Section 4.1), it is evident that we would not want the images sorted such that all cat images precede the dog images. However, more subtle ordering issues can also cause leakage. An example occurred in the TalkingData AdTracking Fraud Detection Challenge[22] hosted on Kaggle. The data was sorted in such a way that if multiple events occurred within the same timestamp and any had a positive label, these were sorted below the negative labels. Although these occurrences were rare, they provided a small amount of leakage that some participants were able to exploit.

Another type of order-based leakage can be introduced while processing individual files. When preparing files for the competition, it is sometimes more convenient to process them by label, such as when opening images, removing metadata, and re-saving with a new observation ID (e.g., processing the folder of cats first and then the folder of dogs). However, if files are saved label-wise, participants can use the file timestamps to determine the label. A best practice is to process individual files randomly and reset the timestamps of the processed files after completion. Redundancy provides additional protection in case of a failure in one of the steps.

When creating a competition data processing pipeline, ensure that random sorting is repeatable by explicitly setting random seeds. Avoid using common random seeds (e.g., 0, 1, 123, 42), as these provide opportunities for participants to reverse-engineer the sorting method. Instead, use unique, difficult-to-guess random seeds that will not be reused in the future.

5.2.1 Integration and Fusion

Data integration or data fusion refers to the process of merging data or datasets from various sources into one dataset (Bleiholder and Naumann, 2008). For instance, if data are from the same relational database (e.g., SQL) but from different tables, then the JOIN operation[23] is a way to perform data fusion (which can expand sample and feature dimensions). In

22. TalkingData AdTracking Fraud Detection Challenge: https://tinyurl.com/4c59vjky
23. https://en.wikipedia.org/wiki/Join_(SQL)

this case, there often exists some matching identifier (e.g., an index that the tables can be joined on) that helps perform this task directly. Moreover, it is now possible to collect data from external databases (without matching identifiers), websites, or search engines in order to improve the predictive performance of machine learning models.

The type of algorithm used to perform data integration varies depending on the data type (e.g., tabular, image, sequence). During such operation, some typical problems to resolve are identification of matching entities (e.g., tabular), re-scaling (Section 5.2.3), spatial/temporal alignment or registration (e.g., 3D shapes, images, videos, Section 5.2.5), cleaning (e.g., removal of duplicates, imputation of missing values, Section 5.2.2), calibration and normalization (Section 5.2.5). Machine learning algorithms (Meng et al., 2020) can be used to address these problems and enable performing data integration or fusion. For example, the problem of 3D shape registration is usually resolved through the iterative closest point algorithm (Arun et al., 1987), based on the least-squares method. An advantage of the iterative closest point algorithm is that it iteratively estimates matching points, unlike the Procrustes-based method (Gower and Dijksterhuis, 2004). Another example is the Fuzzy-Join (Wang et al., 2011) literature, which intends to resolve the problem of inexact matching to merge two different datasets of tabular data.

It is important to mention that despite all the efforts, data integration is often far from perfect, and one should keep track of the source of each data as part of metadata to identify possible biases (e.g., the situation where all recorded patients with a given disease are coming from the same hospital).

5.2.2 Cleaning

Data cleaning refers to the process of improving the consistency of data (Ridzuan and Zainon, 2019). Some data records may be corrupted (e.g., download errors), incoherent (e.g., the same entity is represented by different tokens, or the same data has different associated labels), and may include missing data (e.g., partial information about a user). Data cleaning can be applied both on the input data (e.g., images, videos, text, etc.), the metadata, or any annotation (e.g., the target variable), jointly or separately. Processing all simultaneously is necessary to detect sampling bias with respect to individual samples or groups of samples and decide if the chosen data cleaning method is appropriate.

In fact, distinct methods have been proposed in the literature to deal with missing data, such as partial deletion, statistical imputation, interpolation, and Bayesian inference. However, missing data is problematic due to the risk of bias, which depends on the type of missing values, the relative size of the data that are missing, and the way of dealing with these missing values, which can have associated risks (e.g., yield false positives) and benefits (e.g., reduce false negatives) (Seijo-Pardo et al., 2019, 2018). For example, if missing data are missing at random, then they can be imputed (Van Buuren, 2018). However, if they are not missing at random, spectrum bias may be reinforced (e.g., increase bias toward already present patterns) with imputations. Similarly, samples with missing data could be removed (Guyon et al., 1996) but with the risk of introducing an exclusion bias.

Among data cleaning methods, Bayesian inference (Lew et al., 2021) is a family of methods offering good performance and automation capabilities. It can also leverage prior

knowledge from domain expertise (i.e., understanding of the data), which is often the key to success.

5.2.3 Reduction

Data reduction corresponds to processes that reduce the information contained in data. It includes methods to reduce the dimensionality of feature space (often referred to as dimensionality reduction) and the number of samples (often referred to as sub-sampling). In some cases, data reduction can also refer to the re-scaling or re-sampling of signals according to spatial or temporal dimensions. Contrary to data augmentation and feature engineering, the goal of data reduction is to *reduce* spurious information in data, for instance, to accelerate learning or to make the predictor more robust by eliminating redundancy or noise in data. While several machine learning algorithms are naturally designed to separate important patterns from noise in the data, data-centric approaches such as data reduction can further facilitate this.

Concerning features, a canonical way to perform dimensionality reduction is the Principal Components Analysis (Wold et al., 1987) (PCA), which consists of building a subset of features, which are linear combinations of the original features and explains best the variance in data. PCA can be viewed as an ancestor of neural network auto-encoder methods for manifold learning. Indeed, Bourlard and Kamp (1988) have shown that for n inputs, a 2-layer network with $d < n$ hidden units, trained to reproduce its input on its output with mean-square-error, yields a representation projecting the data in the d directions of largest variance. Non-linear auto-encoders and their descendants (such as denoising auto-encoders (Bengio et al., 2013) and variational auto-encoders (Kingma and Welling, 2014)) provide a generalization of PCA. However, when working in a reduced space it can be required to reconstruct data in the original space. For example, in climate applications, the data are often extremely large and it is required to reduce the size of the data representation during learning but the prediction needs to be in the original space (Maulik et al., 2020). This step can be challenging with autoencoders; in many cases, reconstructed data do not satisfy validity constraints (i.e., data are not realistic). Some methods, such as grammar autoencoder (Kusner et al., 2017b) impose additional structure to alleviate this challenge.

Feature selection methods (Guyon and Elisseeff, 2003) are a particular case of dimensionality reduction methods that avoid replacing features with newly constructed features, which can facilitate explainability (i.e., interpretation about how a prediction is constructed from the inputs) in some applications. Like other methods of dimensionality reduction, removing redundancy and noise is the primary goal.

Also, more classical quantization can be performed to compress a signal. Quantization is the process of mapping a variable from an uncountable (e.g., real values) to a countable space. It is the core of lossy-compression algorithms and can be used to reduce the memory size of signals. A typical and fast way to perform quantization is through the k-means algorithm (Pollard, 1982).

Some dimensionality reduction methods are directly applicable to reduce the number of samples, such as PCA or clustering methods (Yen and Lee, 2009). It is also possible to leverage gradient-based methods to detect non-informative samples (Killamsetty et al., 2021). While sub-sampling can be used to balance the proportion of different classes in

a dataset, it is unclear if it is consistently well-performing (García et al., 2012). The Imbalance-Learn Python package (Lemaître et al., 2017) provides a set of algorithms to perform sub-sampling.

The utility of data reduction is particularly important for spatiotemporal data, which quickly grow in size and face computational and memory limitations (Steadman et al., 2021). Although many of the previously introduced areas of data reduction have been extended to this setting, down-sampling of spatial resolution with bi-linear interpolation and strided sub-sampling over the temporal axis are often preferred in practice.

5.2.4 Representation

Data representation refers to a set of techniques that maps data to a numerical representation that is well-suited for the learning method. Basic data types can be real (e.g., height of a person in centimeters), discrete (e.g., number of users on a website), categorical nominal where there is no order on categories (e.g., type of vehicles such as car, scooter, or truck) and categorical ordinal where there is an order on categories without a clearly defined numerical scale (e.g., rating of a restaurant such as 'very bad', 'bad', 'medium', 'good' or 'very good'). In machine learning, data are generally represented as tensors (i.e., a n-dimensional matrix) even in the case of non-regular structures such as graphs (i.e., node-features, edges-features, connectivity which corresponds to $n = 3$). The problem of finding a good representation is key in machine learning (i.e., a representation that makes the model learn and generalize better).

In some cases, one wants to convert images into more high-level features. While techniques like deep neural networks can learn directly from the raw pixel values, there might be reasons to prefer more abstract and semantically richer, higher-level representations. A straightforward way of obtaining those is to use pre-trained models (Weiss et al., 2016) (benefiting from transfer-learning), such as I3D (Carreira and Zisserman, 2017) or R(2+1)D (Tran et al., 2018) for spatio-temporal feature representations. To do this, we can collect the output provided by the penultimate layer of a deep neural network to represent our new features. It is also possible to adapt the representation to our task. For example, we can perform a complete fine-tuning (i.e., all the weights of the neural network continue to be trained on our new task) or a simple linear-probing (i.e., all the weights are frozen we just change the last layer and train it). In particular, for many computer vision tasks, such a strategy provides a good enough initialization and speeds up the training on new datasets while being the most convenient strategy for small vision datasets.

Then, to overcome the difficulties of data annotation by reducing the quantity of annotated data, much research has focused on learning representations from unlabeled data. Learning representations from unlabeled data is now called self-supervised learning (Zhao et al., 2024) but directly corresponds to an extension of works previously classified under unsupervised learning. Self-supervised learning had many successes in natural language processing with methods such as Word2Vec (Mikolov et al., 2013), where a vector representation of words is learned, and arithmetic can be performed on such vectors having some plausible semantic interpretation. For instance, subtracting the vector representing "men" from that representing "king", then adding those of "women", yields a position in vector space close to "queen". The Word2Vec representation is obtained with a neural network, having at

their input a part of the sentence, except for a missing central word, and at the output, the central word to be predicted (this algorithm is referred to as Continuous Bag of Words). Word2Vec has been a leap forward compared to previous bag-of-word representations, only based on frequencies of words in documents, such as TF-IDF (Sammut and Webb, 2010). Many other works followed the steps of Word2Vec and brought new achievements in the area of natural language processing (e.g., Glove (Pennington et al., 2014), fastText (Bojanowski et al., 2017), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLM-R (Conneau et al., 2020)). More recently, large language models have appeared (Brown et al., 2020) as the final realization of this methodology.

In computer vision, self-supervised learning is also applied to extract representations of images, which are later fine-tuned for supervised classification or regression tasks. In fact, using self-supervised learning on medical images has been shown to reduce the learning of spurious correlations (Goel et al., 2021) between input data and annotations, which also results in better performance (about +10% more on accuracy in some cases) (Azizi et al., 2021). The ideas from natural language processing and computer vision are now being generalized to other data representations such as graphs (e.g., Graph2Vec (Narayanan et al., 2017)) and spatiotemporal data.

All primordial methods of self-supervised learning are variants of auto-encoders, which learn a latent representation by first encoding and then decoding. They have recently been renamed "non-contrastive self-supervised learning methods". A recent methodology for non-contrastive self-supervised learning is to use in-painting to learn to predict missing parts (i.e., occlusions) of an input image, where the occluded image is at the input of an auto-encoder and the missing part(s) at the output (Pathak et al., 2016). The limitation of non-contrastive self-supervised learning methods is that the model is not informed about counter examples, i.e., examples which are out-of-distribution or out of the support of the positive examples provided. This limitation motivated the need for contrastive learning, self-supervised learning (Chen et al., 2020), based on pairwise comparisons of similar and dissimilar examples. To that end, a Siamese neural network architecture (Bromley et al., 1993) is used, consisting of two identical networks whose outputs are compared with a contrastive loss function, such that agreement is maximized for similar or compatible inputs (e.g., two images of the same object, but from different views) and minimized in the case of dissimilarity or incompatibility (e.g., inputs represents different objects).

We illustrate the similarities and differences of self-supervised learning methods in Figure 5. Both have in common the mapping of the input data x to a new representation z (in blue). For non-contrastive (orange), often based on reconstruction schemes, the goal is to learn the representation z from x, which helps reconstruct the distorted data z' (e.g., jigsaw puzzle, in-painting). For contrastive (purple), the goal is to learn similar representations for similar entities and different representations for different entities. Similar entities are artificially simulated with data augmentation (e.g., x can be the image of a bird, and x' is the image of the same bird with a rotation), and then a contrastive loss can be used to enforce the contrastive idea.

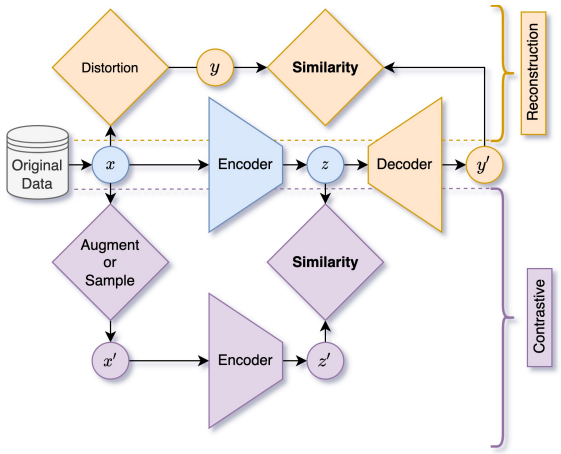


Figure 5: Self-Supervised Learning through Contrastive (purple) or Non-Contrastive (orange) Learning. The input data is x, and the representation learned is z for both.

5.2.5 Normalization and Calibration

Data normalization[24] or data calibration aims to get rid of some systematic bias, which may occur in data collection, due to several uncontrolled factors (e.g., a change in operator, a change in temperature, humidity, luminosity, amount of a certain reagent, etc.). Data calibration should not be confused with model calibration, which focuses on calibrating predictions to improve their probabilistic interpretability.

Linear normalizations or calibration simply amount to shifting and scaling features by an amount determined either by comparing samples to one another (e.g., normalization by dividing by the maximum), by the sample itself (e.g., normalization by diving by the norm of the sample) or by using some reference values (calibration). For example, standard feature normalization consists of removing the mean and dividing by the standard deviation feature-wise. This normalization is commonly performed on input data for neural networks, ensuring all features are unitless and spread over a similar range. In normalization, the quantities used to pre-process the data are directly estimated from these data. Therefore, one needs to be aware that the smaller the dataset is, the higher the variance of these estimates is (i.e., standard error is a function of the $\frac{1}{\sqrt{n}}$ where n is the number of examples), which can create unstable results when the dataset is small. In tabular data, normalizations are often carried out row-wise, column-wise, or both, depending on the nature of the application. Some practical way of performing normalization is through the Scikit-Learn[25] library which provides ready methods more or less sensitive to outliers such as: MinMaxScaler, MaxAbsScaler, StandardScaler, RobustScaler, Normalizer, QuantileTransformer and PowerTransformer.

Calibration can be thought of as a learning problem with a small training set. There are two types of calibrations, i.e., either making use of internal or external calibrants. An

24. not to be confused with "database normalization"
25. Scikit-Learn: https://tinyurl.com/bdfr7z62

internal calibrant is included in every sample. For instance, in chemistry, this would be a compound spiked in known quantity in a sample to adjust the scale of a titrating device; in photography, this would be e.g., landmarks positioned with a given geometry or a known pattern of given shapes and colors, captured together with the scene, serving as a reference to compensate for camera aberrations and adjust the color spectrum of pictures taken. In contrast, an external calibrant is a reference sample (with a well-defined pattern) inserted regularly in between regular samples—for instance, a chessboard image in photography, or the use of a water-solution in a spectrometer.

Internal calibrants are used when measurements constantly change, while external calibrants are suitable for slow drifts in recording conditions. In either case, the calibrants' ground truth (target values) are known. This allows users to train a simple predictive model (often just linear), which maps measured values to target values. The predictor can then interpolate between known values to correct the other measured values in the sample. This type of method is commonly used in computer vision for camera calibration (Zhang, 2000). Calibration is particularly needed for data fusion when samples are obtained from different sources, which relates this problem to data integration introduced in Section 5.2.1. Another example is the calibration of data from multiple sensors, such as in autonomous vehicles (Yeong et al., 2021).

5.2.6 Augmentation

Data augmentation is the process of artificially increasing the size of an already existing dataset (either with respect to samples or with respect to features (Liu et al., 2018)). It potentially enlarges the dataset by orders of magnitude at a reduced cost. It can be rule-based, such as in computer vision, where it is possible to perform changes in resolution, orientation, brightness, execute random crops/shifts, and include noise, among others (Shorten and Khoshgoftaar, 2019). It can also be learning-based, where the underlying distribution of the data is learned via generative models (e.g., generative adversarial networks (Goodfellow et al., 2020), variational auto-encoders (Kingma et al., 2016)), to then artificially sample from it. Although data augmentation can lead to overall improvements in performance (i.e., improved average metrics), it may introduce *selection bias* yielding an increase in performance for some classes or groups at the expense of others. This is related to the fact that designing dataset/task-specific regularizers without introducing selection bias remains an open research question (Balestriero et al., 2022). That is why proper model evaluation and selection needs to be conducted to quantify such effects. Some types of data augmentation are preferably executed during the training to avoid storing the augmented data. Finally, augmentation is often used jointly with other learning methods, especially self-supervised methods such as contrastive learning (Chen et al., 2020) and Siamese networks (Chen and He, 2021) to leverage the knowledge of proximity between original and augmented samples.

6 Evaluation

The goal of dataset evaluation is to assess whether a dataset meets its original dataset requirements, ensuring that it is suitable for training reliable and fair AI models. This includes verifying specified quality and quantity criteria, catching errors in dataset implementation, and identifying flaws in the dataset design. Our evaluation framework is based

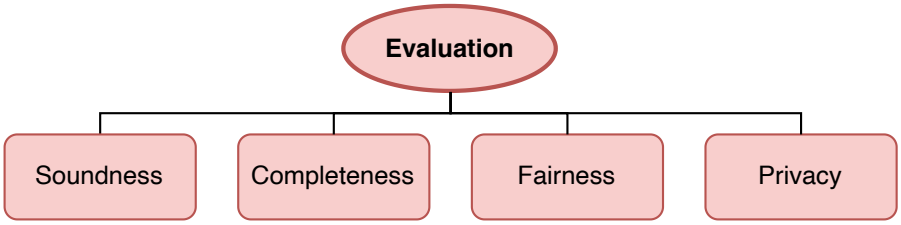


Figure 6: Categorization of sub-processes included in data evaluation.

on four key criteria: soundness, completeness, fairness, and privacy. Each of these criteria benefits from both qualitative (e.g., data visualization) and quantitative assessments (e.g., metrics).

6.1 Preliminaries: Inspection, Visualization, and Baselines

Visualization is key in evaluating a dataset (Figure 7). Data visualization tools should be prepared and used during data collection (Section 5.1) to help identify anomalies early. Common visualization techniques (Figure 7) include:

Heat or cluster maps: Useful for vectorial data to check for anomalous structures.

Pair plots: For datasets with few features, pair plots help visualize class separability. For many features, apply PCA first and use pair plots on the principal components.

Bar plots: Use bar plots to represent the frequency of examples in each class. These can help identify imbalances.

Commonly used library for data visualizations are: Matplotlib[26], Seaborn[27], and Microsoft SandDance[28].

26. https://matplotlib.org/
27. https://seaborn.pydata.org/
28. https://microsoft.github.io/SandDance/

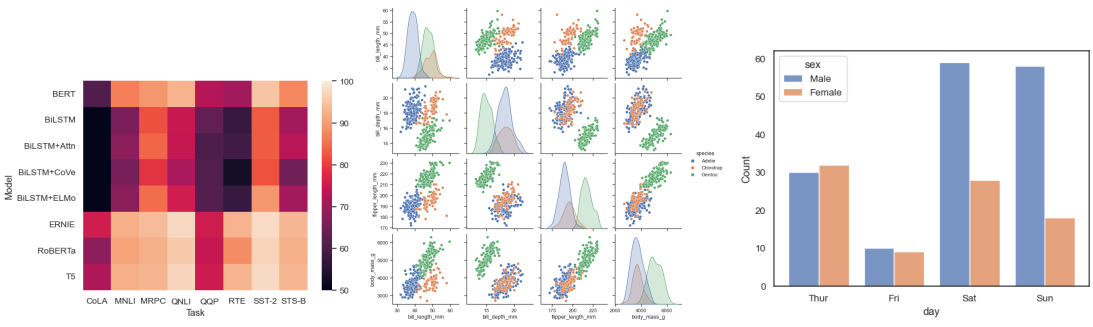


Figure 7: Example visualizations of (left) a heatmap, (middle) a pair-plot, and (right) class frequencies from the Seaborn library.

It is recommended to apply baseline learning methods that tries to solve the task represented by the dataset early in the collection process. Baselines can help assess the difficulty of the dataset, and they can range from very basic to state-of-the-art. Also, the performance gap between the constant predictor (i.e., constant with respect to the input x, e.g., often the mode of target classes in the case of classification and the mean of targets in the case of regression) and state-of-the-art methods can be an indicator of learnable concepts in the data. Then, with trained baselines, other visualization can be done. For example, inspecting the confusion matrices for classification tasks is good practice. This can help determine "unfair" (Section 6.4) predictive performance across classes.

Visualization can be vital in revealing data leakage or bias, for example, by spotting illegitimate information in the input data (e.g., the case in which a red patch is present on all images of cancer cells while a green patch is present on all images of normal cells).

6.2 Soundness

A dataset is considered sound if it correctly results from its premises, which are the requirements, design, and implementation steps, and if these premises are themselves correct. For example, a design choice could be to identify cities by name and location (which would be correct), while another design could be to identify cities solely by name (which would be incorrect because names are not unique identifiers). Therefore, this will include verifying upstream steps and looking for inconsistencies, corruptions, and a good state of collected distributions. Classic unit tests can be implemented to partially check the soundness of the dataset. Next, we discuss some sources of consistency that should receive special attention.

Representation consistency relates to the problem of having a unique representation for the same entity across the dataset. For instance, when collecting articles from the press on the Internet, the same city can be written differently, such as "New York", "N.Y.", and "the Big Apple", which all represent the same city. Similarly, it is essential to check whether physical quantities are all measured using the same unit. In the case of tabular data and storage systems, improving representation consistency is often referred to as data deduplication (Xia et al., 2016).

Labelling consistency refers to self-agreement and inter-agreement among annotators, which is especially important when labeling was performed through crowd-sourcing as mentioned in Section 5.1.4. Suppose various annotators are participating in the annotation process. In that case, it is important to ensure they agree on the exact concept they are labeling and use the same definition and thresholds. Self-agreement is particularly useful in identifying low-quality annotators, while inter-agreement is suitable for estimating the task's difficulty. Krippendorff's Alpha-Reliability (Krippendorff, 2011) is an example of such a metric.

Outlier detection relates to identifying observations that appear inconsistent with other observations of the dataset (Hodge and Austin, 2004). Data visualization is particularly useful for detecting the presence of such samples. Other typical quantitative methods are based on the interquartile range (IQR, represented in box plots) and the Z-score. After identifying outliers, a domain expert can decide whether they result from errors.

Bias detection in the data generally concerns the identification of systematic outcomes of the dataset development process that results in a dataset that is not representative of the true observed phenomenon (Figure 2). For example, a lack of randomization of nuisance factors can result in spurious correlations. Collecting appropriate metadata (i.e., data necessarily not available for training but available for data evaluation) is essential to help detect such bias, including potential nuisance factors (e.g., temperature, humidity, luminosity, recording time, date, collection operator, etc.) and protected groups (e.g., age, gender, ethnicity, etc.) involved in societal bias (fairness, Section 6.4). Subjective bias (Section 6.4) coming from data annotation can also be mitigated by collecting proper metadata to reveal biases with respect to both the annotator and the data being annotated (Jacques Junior et al., 2021; Escalante et al., 2020). A machine learning model may then be trained using variables that are suspected causes of bias, either in isolation or in combinations. Feature selection methods are then applied to determine whether such variables are significantly predictive.

We now present use cases where the dataset was not sound.

Use Case: Spurious Correlation With Video Metadata in the "Looking at People" Challenge

The ChaLearn "Looking at People" Challenge on Self-Reported Personality Recognition (Palmero et al., 2022) adopted a dataset composed of large amounts of data (audio-visual, transcripts, metadata, etc.). However, one competitor team achieved promising results by analyzing the correlation between metadata and the self-assessment personality trait scores (the target variables) while proposing to use a random forest regressor trained solely on metadata features (i.e., age, gender, and number of sessions). That is, the competitor did not utilize the available 60 hours of provided audio-visual data and associated transcripts for training, which data creators believed to be crucial to addressing the problem. In other words, the data creators were not anticipating that a model trained on metadata features could accurately predict someone's personality, indicating that the adopted dataset may include unwanted bias. According to this competitor (and challenge results), a simple random forest regressor based on metadata features only should not be enough to outperform a method based on linguistic, audio-visual, and metadata features like the alternative model they evaluated in such a complex task as personality recognition.

Use Case: Spurious Correlation With File's Metadata in the "Whale" Challenge

A notable example of leakage from file metadata was a competition to detect whether underwater audio contained a Right whale up-call[29]*. A competitor scored an AUROC performance of 0.997 without actually reading the files (let alone making machine learning predictions). This was achieved simply by looking at the size-on-disk of the test files, the timestamp embedded in the audio clip filename, and the chronological order of the clips, which provided enough information to specify which files contained a whale up-call.*

29. Kaggle's Right Whale Challenge: https://tinyurl.com/8d2uw69k

Use Case: Temporal Leakage in the "Predict Future Sales" Challenge

Another common soundness issue is time series leakage, wherein future data provided to competitors inadvertently reveals information. For instance, in the Corporación Favorita Grocery Sales Forecasting Kaggle competition, competitors were given oil prices up to the period of the test set[30]*. This might appear reasonable since the challenge was to predict store sales, not future oil prices. However, since the data originated from Ecuador, an oil-dependent country whose economic health is highly susceptible to oil price fluctuations, this inclusion constituted leakage. In real-world applications, predictive models would not have access to actual future oil prices, leading to inflated performance metrics in the challenge due to this leakage. Ideally, in time series competitions, competitors should only be provided with data up to the prediction horizon. Once their models generate predictions for the next period, they can receive the subsequent increment of data, continuing in this manner. While setting up a competition this way poses practical challenges, organizers might opt to provide the entire test data series upfront. Nevertheless, it is crucial to recognize that this approach introduces a form of leakage.*

6.3 Completeness

The completeness of a dataset is the attribute of a dataset to contain all required features describing sufficiently the problem at hand, as well as to properly select its samples (e.g., the number of i.i.d - independent and identically distributed - samples directly impacts the estimation of the mean estimate - standard error). Therefore, completeness can be evaluated with respect to samples or features but also with respect to metadata or hidden variables, which do not define the problem but have to be adequately sampled (e.g., randomized or factorial design) to avoid bias.

Feature-wise completeness is associated with the notion of causality, which defines that a cause can be necessary or sufficient. If X is a sufficient cause of Y, then the presence of X necessarily implies the subsequent occurrence of Y. However, another cause Z may alternatively cause Y. Thus, the presence of Y does not imply the prior occurrence of X. Then, if X is a necessary cause of Y, the presence of Y necessarily implies the prior occurrence of X. However, the presence of X does not imply that Y will occur. These relations matter to understanding the problem of confounding factors, which relates to a false association between variables such as X causing Y because of a third missing variable Z causing the two (Nunan et al., 2018).

Confounding variables can be divided into omitted variables and correlated noise (i.e., a spurious feature). As an example of an omitted variable: if X is "drinking coffee" and Y is "developing a lung cancer", some data can show that drinking coffee increases the risk factor of developing lung cancer. However, this is happening because a common cause Z, which is "smoking", was not taken into consideration and is associated with both "drinking coffee" and "developing lung cancer". An example of correlated noise is the background (e.g., road, sky, water) with the type of vehicle (e.g., car, plane,

30. Corporación Favorita Grocery Sales Forecasting Kaggle competition: https://tinyurl.com/dttt4e86

boat). A picture of a car will rarely happen with a sky background; therefore, the background (the road) is predictive of the object's class (the car).

It can be tracked by identifying whether some common candidate confounding factors that are spuriously predictive of the target variable (individually or jointly). For sensor data, typical examples are temperature, humidity, luminosity, etc.; for survey data, typical examples are age, gender, ethnicity, etc. Classical feature importance methods can track how significant such associations are (Altmann et al., 2010). Proper metadata collection (also called protected attributes in fairness) can aid in this process (Bellamy et al., 2018). If some of these attributes cannot be recorded directly, then it is advisable to measure them indirectly. For example, suppose one suspects the image background could be a spurious feature. In that case, one may create mini-images containing only the background of the original images and try to predict the target variable from them (excluding the object of interest). If the model now predicts the object of interest, the background was utilized in the predictive model (Tian et al., 2018). If confounding bias is identified, the data collection process must be revised to alleviate it. Missing variables often cause excessive aleatoric uncertainty (or intrinsic randomness), which is the variability of the outcome given input variables because of unknown source factors. This variability needs to be assessed to make sure the prediction is performed within a reasonable confidence interval.

Sample-wise completeness is directly associated with the problem of selection bias, which can be divided into exclusion bias (removal of samples) and spectrum bias (only a subset of the target population was observed). Exclusion bias comes from a choice of the dataset development team. For example, it can result fram data cleaning, which may remove too many samples or may modify the data distribution by removing specific observations (e.g., creating an imbalanced dataset). It can also be a choice of filter in a search engine or a selected window of observation. A well-known bias in an academic dataset is the recruitment of graduate students as subjects. Testing sample-wise completeness depends on the assumption made on the problem. In the i.i.d. samples case, it can be tested through the classic generalization error using learning curves techniques (Mohr and van Rijn, 2022, 2023).

6.4 Fairness

Fairness has recently attracted attention in the machine learning community (Bird et al., 2019; Mehrabi et al., 2021) after several flaws arising from misuse of biased data being reported by the media such as the Guardian[31] or the Washington Post[32]. In the context of decision-making, and according to (Mehrabi et al., 2021), fairness is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics. Therefore, an unfair algorithm is one whose decisions are skewed toward a particular group of people.

Machine learning algorithms can inherently transfer bias from data to model. When black-box models are used (e.g., most deep learning architectures), results are usually diffi-

31. The Guardian, 2022: http://tinyurl.com/yfeeu2sx
32. The Washington Post, 2020: http://tinyurl.com/4kv4nf5v

cult to explain and interpret, making bias mitigation even harder. Hence, a biased dataset used for training or evaluating machine learning methods can negatively impact outcomes. For instance, a face recognition method trained on male faces may not generalize well to female faces. To mitigate such problems, in most countries, the law protects against a discriminatory decision (e.g., about hiring or condemning) based on protected attributes, which include gender, age, and ethnicity, among others (Barocas et al., 2023). One way of ensuring some level of fairness at the data collection level is to ensure that samples are group-balanced, i.e., that there is an approximately equal number of samples in all groups resulting from combinations of protected attributes and labels. Of course, this may be impractical, and it may be more feasible to record the protected attribute to later correct bias at the machine learning level, with in-processing (Wan et al., 2022) (i.e., during learning, which is categorized into explicit and implicit, where the former directly incorporates fairness metrics in training objectives, and the latter focuses on refining latent representation learning) or post-processing techniques (i.e., after learning). However, recording protected variables may raise privacy issues (Section 6.5).

Data annotation may also be a source of social bias. The machine learning and computer vision communities are paying more attention to this problem as it relates to fairness (Bird et al., 2019). Recent work reports different types of subjective biases coming from crowdsourced annotations. Although biases produced by human perception have been widely studied in sociology and psychology (e.g., gender (Oh et al., 2019) or attractiveness (Talamas et al., 2016) bias), little attention has been given to subjective bias analysis (Shen et al., 2019; Quadrianto et al., 2019; Robinson et al., 2020; Yan et al., 2020) beyond the perspective of explainable models (Escalante et al., 2017; Huk Park et al., 2018; Pérez Principi et al., 2019; Escalante et al., 2020). Moreover, as perception depends on the observer, the relationship between annotators and the entity being annotated could explain how some perception biases are produced, which is an almost unexplored area in computer vision and machine learning. However, this would require a dedicated discussion around privacy and ethical issues (Jacques Junior et al., 2021).

Over the past few years, a vast number of scientific events and studies appeared intending to stimulate discussion and advance the state of the art on fairness and bias mitigation methods (e.g., ACM FAccT[33]), explainability and interpretability (e.g., Escalante et al. (2017); Huk Park et al. (2018)). Distinct definitions and metrics for fairness have been proposed and discussed, like "fairness through unawareness", "individual fairness", "demographic parity", "equalized odds", "equality of opportunity" or "counterfactual fairness" (Kusner et al., 2017a; Ashokan and Haas, 2021; Bellamy et al., 2018). Although there is no standard definition of fairness that could be used for all types of problems, researchers must be attentive to possible fairness issues and consult with social scientists and ethics specialists as needed. For instance, it is advisable to have data collection protocols reviewed by an Institutional Review Board (see Section 3) even though this does not guarantee to solve all possible problems.

33. https://facctconference.org

6.5 Privacy

When a dataset contains human-related data, privacy and data protection become mandatory and will impact all aspects of the dataset development. Data protection regulations (e.g., the European General Data Protection Regulation, GDPR (Voigt and Bussche, 2017)) define different levels of protection depending on each type of data, and each level has different requirements for processing and storing. Moreover, the classification of personal data differs in each regulation and can change regularly. Human-related data require collecting an explicit consent form before storing any data, with a clear and understandable description of the collected data, how it will be used, who will have access to it, and how long it will be stored.

Data are considered anonymized if there is no possibility to identify a subject using the provided data. In cases where data curators maintain a correspondence between published data and originally captured data, we cannot consider the data anonymized, as it is possible to recover a subject's identity. In this case, the data is pseudo-anonymized, where without the link between real identity and published data, no one can recover the real identity of a subject in the dataset. Pseudo-anonymized data allow data curators to remove data from an individual if necessary, but they pose a security risk since if this link is compromised, real identities can be recovered. Avoiding data leakage is crucial to maintaining privacy, as it prevents unauthorized access to sensitive information and ensures compliance with legal and ethical standards.

In the context of anonymization and pseudo-anonymization, k-anonymity is an important measure. The measure of k-anonymity implies that given one entry of the datasets (i.e., information of a specific individual), it contains at least $k-1$ identical entries in the dataset corresponding to other individuals. The minimum recommended value for k is three, although larger values ensure better anonymization. For many public datasets, re-identification is easy even when data seems anonymous. For example, Sweeney (2000) demonstrates that 86% of the U.S. population can be uniquely identified with just three "quasi-identifiers": zip code, gender, and date of birth. Although there are easy ways to increase the k-value, e.g., by binning variables, if a dataset contains sensitive data, it is a good idea to apply one of the many existing anonymization algorithms (Casas-Roma et al., 2012). Finally, it is essential to ensure that the anonymization process does not affect the usefulness of the dataset. Carmona et al. (2019) provide an introduction to the subject regarding health data.

Differential privacy (DP) (Dwork, 2008) and the possibility of replacing real-world data with realistic synthetic data providing some privacy guarantees have gained some recent attention. Sablayrolles et al. (2020) propose an apparatus in which an ideal attacker having maximum information evaluates whether such synthetic data are protected against membership inference attacks (i.e., determining whether a sample was or not part of the data used to train the generative model).

7 Distribution and Maintenance

Dataset distribution is about making the developed dataset accessible to others, while maintenance are all tasks and processes required to maintain the dataset accessibility and to perform changes on the dataset or how it is distributed to improve its accessibility or

quality. Those two tasks should often be considered together. Depending on the structure of the dataset (e.g., size, data type, format) those can be simple or complex tasks. Taking into account the requirements of the dataset (Section 3), one may select the maintenance and distribution technologies/strategies that are cost-effective, sustainable, and supportable with available resources (economic and human). Next, we present the aspects we consider the most relevant to be considered:

Ownership and licensing: The distribution of a dataset includes essential legal requirements. The dataset creators have to define the usage and responsibility of distributed data. To this end, the dataset can be distributed under copyrights, specific licensing (Benjamin et al., 2019) (e.g., open-source[34], creative commons[35]) or terms of use (ToU). Licensing not only includes the dataset creators but also the object of shared data, such as information about individuals, whether the information is personal or medical, whether the individual agreed to distribute this data or whether regulation exists for this type of data (e.g., General Data Protection Regulation[36] in Europe, California Privacy Rights Act[37]). Therefore, the maintenance plan may include all the processes required by the data protection regulations, such as the capacity to remove all the data for an individual in case it is required or the partial or total elimination of the dataset and/or original data. The distribution tools may also support such actions.

Hosting platforms: The dataset can be hosted in a Cloud service (e.g., Microsoft Azure, Google Cloud, or Amazon Web Services) and benefit from optimized downloading services if the source (where the dataset is stored) and the target (where the dataset is downloaded) platforms are from the same service (e.g., Google Drive and Google Colab with `gdown`). It is also possible to directly store the data on a web server so that it can be downloaded utilizing the `http` protocol, the File Transfer Protocol (FTP), and the Secure Copy Protocol (SCP). More recent open-science services such as Globus[38] can also be set up to optimize the transfer of data.

Evolution and versioning: Datasets often evolve in time to fix errors or include new data or labels. Also, flaws inside the dataset can be discovered later through active usage (Wang et al., 2022). A good example is linked to the evolution of machine learning research. Recently, privacy preservation and fairness in machine learning have become a priority. Therefore, the ImageNet dataset was updated to anonymize individuals appearing in pictures or filter out problematic samples[39]. Users were informed about these changes through ImageNet's website.

Due to these continuous changes, it is essential to attribute a version or unique identifier to the dataset to differentiate it in case of modification. Open-science (free of

34. https://opensource.org/licenses
35. https://creativecommons.org/
36. https://gdpr.eu/what-is-gdpr
37. California Privacy Rights Act: https://tinyurl.com/mr38ctvu
38. https://www.globus.org/
39. ImageNet: https://tinyurl.com/54ux9bsu

access) websites such as Zenodo[40] and arXiv[41] now provide digital object identifiers (DOI) which can be used for this purpose. Sometimes it is also possible to use data version control[42] if the data is not required to be completely removed (e.g., for scientific reproducibility) so that versions of the data can be tracked. Data version control is essential when having an evolving dataset to keep track of the versions and give the possibility to trace which dataset was used to train a particular model.

Data format: When distributing data, it is often important to compile it under a compressed format so that it can be downloaded faster. Such format can be `.tar`, `.zip`, and `.gz` to cite a few. It is important to inform users about the decompressed size when using a compressed format. Although the format can change due to the evolution of the dataset, it is desirable to maintain backward compatibility as much as possible. This backward compatibility is a general software principle that is also valid for datasets. A significant format change can limit its usability.

Dissemination: Lastly, in the case of a public dataset, it is essential to communicate its existence. Organizing related competitions and events in international conferences can be an excellent opportunity to present the dataset and a first benchmark using it. The NeurIPS conference promoted a specialized track to publish new datasets and benchmarks that can help showcase such datasets. More details are provided by Richard et al. (2024).

8 Conclusion

Datasets are an essential aspect of scientific benchmarks and competitions for machine learning. More importantly, properly designed and evaluated datasets are extremely important for developing trustworthy and robust artificial intelligence systems. In this chapter, we aimed to specify the dataset development process. We have categorized the various steps that should be undertaken in the dataset development cycle, i.e., documentation, requirements, design, implementation, evaluation, and distribution and maintenance. We approached dataset development as an agile process that is often iterative and requires interactions between its sub-processes. However, we acknowledge that every dataset development process can be different, and there is the possibility to emphasize or skip certain parts of this process. For example, in some cases, emphasis is placed on the design phase, whereas in other cases, the maintenance phase is limited (e.g., when there is no ability to improve the dataset after it has been released).

While we have attempted to give a broad overview of the dataset development process, this is by no means exhaustive. When developing a dataset, one must take care to not introduce any bias. Every dataset development process can introduce its distinct type of bias. Furthermore, while this chapter focuses on the dataset development process, many machine learning benchmarks and competitions also include a stage that evaluates the models trained on these data. Typically, this involves splitting the dataset into a train and test set, which, when not appropriately addressed, can induce other types of bias (e.g.,

40. https://zenodo.org/
41. https://arxiv.org/
42. https://dvc.org

information bias and data leakage). While we briefly touched upon the concept of data leakage and how to consider this pro-actively in the dataset development cycle, how to avoid this completely in the concept of competitions is out of the scope of this chapter.

With this chapter, we aimed to harmonize some terminologies from the dataset development process as well as bring together several directions of the literature that we expect to be taken into consideration when developing new datasets.

9 Acknowledgment

This material is based upon work partially supported by the U.S. Department of Energy (DOE), Office of Science, Office of Advanced Scientific Computing Research, under Contract DE-AC02-06CH11357. This material is based upon work partially supported by the ANR Chair of Artificial Intelligence HUMANIA ANR-19-CHIA-0022 and TAILOR EU Horizon 2020 grant 952215. This work has been partially supported by the Spanish projects PID2022-136436NB-I00 and PID2022-138721NB-I00, and by ICREA under the ICREA Academia program.

References

Eirikur Agustsson, Radu Timofte, Sergio Escalera, Xavier Baró, Isabelle Guyon, and Rasmus Rothe. Apparent and real age estimation in still images with deep residual regressors on appa-real database. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 87–94, 2017.

Ashly Ajith and G. Gopakumar. Domain adaptation: A survey. In *Computer Vision and Machine Intelligence - Lecture Notes in Networks and Systems*, pages 591–602. Springer, Singapore, 2023.

André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

K. S. Arun, Thomas S. Huang, and Steven D. Blostein. Least-squares fitting of two 3-D point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 9(5):698–700, 1987.

Ashwathy Ashokan and Christian Haas. Fairness metrics and bias mitigation strategies for rating predictions. *Information Processing and Management*, 58(5):102646, 2021.

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 3478–3488, 2021.

Randall Balestriero, Leon Bottou, and Yann LeCun. The effects of regularization and data augmentation are class dependent. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022*, pages 37878–37891, 2022.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943, 2018.

Emily M. Bender and Batya Friedman. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 12 2018.

Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, pages 899–907, 2013.

Misha Benjamin, Paul Gagnon, Negar Rostamzadeh, Christopher Joseph Pal, Yoshua Bengio, and Alex Shee. Towards standardization of data licenses: The Montreal data license. *CoRR*, abs/1903.12262, 2019.

Karan Bhanot, Miao Qi, John S Erickson, Isabelle Guyon, and Kristin P Bennett. The problem of fairness in synthetic healthcare data. *Entropy*, 23(9):1165, 2021.

Karan Bhanot, Ioana Baldini, Dennis Wei, Jiaming Zeng, and Kristin P. Bennett. Downstream fairness caveats with synthetic healthcare data. *CoRR*, abs/2203.04462, 2022.

Sarah Bird, Krishnaram Kenthapadi, Emre Kiciman, and Margaret Mitchell. Fairness-aware machine learning: Practical challenges and lessons learned. In *International Conference on Web Search and Data Mining*, pages 834–835, 2019.

Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael Gomes Mantovani, Jan N. van Rijn, and Joaquin Vanschoren. OpenML benchmarking suites. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Jens Bleiholder and Felix Naumann. Data fusion. *ACM computing surveys (CSUR)*, 41(1): 1–41, 2008.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4):291–294, 1988.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning, ML Summer Schools 2003*, volume 3176 of *Lecture Notes in Computer Science*, pages 169–207. Springer, 2003.

Alexander Braylan, Omar Alonso, and Matthew Lease. Measuring annotator agreement generally across complex structured, multi-object, and free-text annotation tasks. In *Proceedings of the ACM Web Conference*, pages 1720–1730, 2022.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "Siamese" time delay neural network. In *Advances in Neural Information Processing Systems 6, [7th NIPS Conference]*, pages 737–744, 1993.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, pages 1877–1901, 2020.

Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

Guilherme Camargo, Pedro H. Bugatti, and Priscila T. M. Saito. Active semi-supervised learning for biological data classification. *PLOS ONE*, 15(8):1–20, 2020.

Fer Carmona, Jordi Conesa, and Jordi Casas-Roma. Towards the analysis of how anonymization affects usefulness of health data in the context of machine learning. In *International Symposium on Computer-Based Medical Systems*, pages 604–608, 2019.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

Jordi Casas-Roma, Jordi Herrera-Joancomartí, and Vicenç Torra. Comparing random-based and k-anonymity-based algorithms for graph anonymization. In *Modeling Decisions for Artificial Intelligence*, volume 7647, pages 197–209. Springer, 2012.

Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. Let's agree to disagree: Fixing agreement measures for crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 5(1):11–20, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International conference on machine learning, ICML 2020*, Proceedings of Machine Learning Research (PMLR), pages 1597–1607, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2021.

Gary Chin. *Agile Project Management: How to Succeed in the Face of Changing Project Requirements*. Amacom, 2004. ISBN 9780814427361.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 8440–8451. Association for Computational Linguistics, 2020.

Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F. Cohn, and Sergio Escalera Guerrero. Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(8):1548–1568, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics, 2019.

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.

Cynthia Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, pages 1–19. Springer Berlin Heidelberg, 2008.

Maurizio Di Paolo Emilio. *Data Acquisition Systems: From Fundamentals to Applied Design*. Springer, 2013. ISBN 1461442133.

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 *International Joint Conference on Neural Networks (IJCNN)*, pages 3688–3695, 2017.

Hugo Jair Escalante, Heysem Kaya, Albert Ali Salah, Sergio Escalera, Yağmur Güçlütürk, Umut Güçlü, Xavier Baró, Isabelle Guyon, Julio C. S. Jacques, Meysam Madadi, Stephane Ayache, Evelyne Viegas, Furkan Gurpinar, Achmadnoer Sukma Wicaksana, Cynthia Liem, Marcel A. J. Van Gerven, and Rob Van Lier. Modeling, recognizing, and explaining apparent personality from videos. *IEEE Transactions on Affective Computing*, 13(2):894–911, 2020.

Vicente García, José Salvador Sánchez, and Ramón Alberto Mollineda. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25(1):13–21, 2012.

Hector Garcia-Molina, Manas Joglekar, Adam Marcus, Aditya Parameswaran, and Vasilis Verroios. Challenges in data crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, 28(4):901–911, 2016.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commununications of the ACM*, 64(12):86–92, 2021.

Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. Model patching: Closing the subgroup performance gap with data augmentation. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

John C Gower and Garmt B Dijksterhuis. *Procrustes Problems*. Oxford University Press, 2004. ISBN 9780198510581.

Robert M. Gray and David L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, 1998.

Yingjie Gu, Zhong Jin, and Steve C. Chiu. Combining active learning and semi-supervised learning using local and global consistency. In *Neural Information Processing*, volume 8834 of *Lecture Notes in Computer Science*, pages 215–222. Springer, 2014.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

Isabelle Guyon, Nada Matic, and Vladimir Vapnik. *Discovering informative patterns and data cleaning*, pages 181–203. American Association for Artificial Intelligence, USA, 1996.

Isabelle Guyon, John Makhoul, Richard Schwartz, and Vladimir Vapnik. What size test set gives good error rate estimates? *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 20(01):52–64, 1998.

David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

John Rowland Higgins. *Sampling theory in Fourier and signal analysis: foundations*. Oxford Science Publications, 1996. ISBN 0198596995.

Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126, 2004.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *CoRR*, abs/1805.03677, 2018.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8779–8788, 2018.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 560–575, 2021.

Julio C. S. Jacques Junior, Agata Lapedriza, Cristina Palmero, Xavier Baro, and Sergio Escalera. Person perception biases exposed: Revisiting the first impressions dataset. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 13–21, 2021.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):1–21, 2012.

Krishnateja Killamsetty, S Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, Proceedings of Machine Learning Research (PMLR), pages 5464–5474, 2021.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 4743–4751, 2016.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. Reduced, reused and recycled: The life of a dataset in machine learning research. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*, 2021.

Klaus Krippendorff. Computing krippendorff's alpha-reliability. Technical report, Annenberg School for Computing, 2011.

Aditya Kuppa, Lamine Aouad, and Nhien-An Le-Khac. Towards improving privacy of synthetic datasets. In *Privacy Technologies and Policy*, pages 106–119, Cham, 2021. Springer International Publishing. ISBN 978-3-030-76663-4.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 4066–4076, 2017a.

Matt J. Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research (PMLR)*, pages 1945–1954, 2017b.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

Alexander Lew, Monica Agrawal, David Sontag, and Vikash Mansinghka. Pclean: Bayesian data cleaning at scale with domain-specific probabilistic programming. In *International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research (PMLR)*, pages 1927–1935, 2021.

Bo Liu, Xudong Wang, Mandar Dixit, Roland Kwitt, and Nuno Vasconcelos. Feature space transfer for data augmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9090–9098, 2018.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *CoRR*, abs/1907.11692, 2019.

Javier Maroto and Antonio Ortega. Efficient worker assignment in crowdsourced data labeling using graph signal processing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2271–2275, 2018.

Romit Maulik, Romain Egele, Bethany Lusch, and Prasanna Balaprakash. Recurrent neural network architecture search for geophysical emulation. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '20. IEEE Press, 2020.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021.

Tong Meng, Xuyang Jing, Zheng Yan, and Witold Pedrycz. A survey on machine learning for data fusion. *Information Fusion*, 57:115–129, 2020.

Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. Documenting computer vision datasets: An invitation to reflexive data practices. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 161–172, 2021.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013*, 2013.

Felix Mohr and Jan N. van Rijn. Learning curves for decision making in supervised machine learning - A survey. *CoRR*, abs/2201.12150, 2022.

Felix Mohr and Jan N. van Rijn. Fast and informative model selection using learning curve cross-validation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(8):9669–9680, 2023.

Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. Graph2vec: Learning distributed representations of graphs. *CoRR*, abs/1707.05005, 2017.

Sergey Nikolenko. *Synthetic Data for Deep Learning*. Springer, Cham, 2021. ISBN 978-3-030-75177-7. Part of the Springer Optimization and Its Applications book series (SOIA, volume 174).

Stefanie Nowak and Stefan Rüger. How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, page 557–566, 2010.

David Nunan, Jeffrey Aronson, and Clare Bankhead. Catalogue of bias: attrition bias. *BMJ Evidence-Based Medicine*, 23(1):21–22, 2018.

DongWon Oh, Elinor A. Buck, and Alexander Todorov. Revealing hidden gender biases in competence impressions of faces. *Psychological Science*, 30(1):65–79, 2019.

Cristina Palmero, German Barquero, Julio C. S. Jacques Junior, 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 *Understanding Social Behavior in Dyadic and Small Group Interactions*, volume 173 of *Proceedings of Machine Learning Research (PMLR)*, pages 4–52, 2022.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

Ricardo Darío Pérez Principi, Cristina Palmero, Júlio C. S. Jacques Júnior, and Sergio Escalera. On the effect of observed subject biases in apparent personality analysis from audio-visual signals. *IEEE Transactions on Affective Computing*, pages 1–14, 2019.

David Pollard. Quantization and the method of k-means. *IEEE Transactions on Information theory*, 28(2):199–205, 1982.

Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian A. Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. ChaLearn LAP 2016: First round challenge on first impressions - dataset and results. In *European Conference on Computer Vision Workshop (ECCVW)*, pages 400–418, 2016.

Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8219–8228, 2019.

Rossana Queiroz, Marcelo Cohen, Juliano L. Moreira, Adriana Braun, Julio C. Jacques Junior, and Soraia Raupp Musse. Generating facial ground truth with synthetic faces. In *Conference on Graphics, Patterns and Images*, pages 25–31, 2010.

Christoffer Bøgelund Rasmussen, Kristian Kirk, and Thomas B. Moeslund. The challenge of data annotation in deep learning - a case study on whole plant corn silage. *Sensors*, 22(4), 2022.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, 11(3):269–282, 2017.

Magali Richard, Yuna Blum, Justin Guinney, Gustavo Stolovitzky, and Adrien Pavão. AI competitions and benchmarks, practical issues: Proposals, grant money, sponsors, prizes, dissemination, publicity. *CoRR*, abs/2401.04452, 2024.

Fakhitah Ridzuan and Wan Mohd Nazmee Wan Zainon. A review on data cleansing methods for big data. *Procedia Computer Science*, 161:731–738, 2019.

Joseph P. Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: Too bias, or not too bias? In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1–10, 2020.

Yuji Roh, Geon Heo, and Steven Euijong Whang. A survey on data collection for machine learning: A big data - ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1328–1347, 2021.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Herve Jegou. Radioactive data: tracing through training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research (PMLR)*, pages 8326–8335, 2020.

Claude Sammut and Geoffrey I. Webb, editors. *TF–IDF*, pages 986–987. Springer, Boston, MA, 2010. ISBN 978-0-387-30164-8.

Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. DC-check: A data-centric AI checklist to guide the development of reliable machine learning systems. *CoRR*, abs/2211.05764, 2022.

Borja Seijo-Pardo, Amparo Alonso-Betanzos, Kristin P. Bennett, Verónica Bolón-Canedo, Isabelle Guyon, Julie Josse, and Mehreen Saeed. Analysis of imputation bias for feature selection with missing data. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 655–660, 2018.

Borja Seijo-Pardo, Amparo Alonso-Betanzos, Kristin P. Bennett, Verónica Bolón-Canedo, Julie Josse, Mehreen Saeed, and Isabelle Guyon. Biases in feature selection with missing data. *Neurocomputing*, 342:97–112, 2019.

Burr Settles. Active learning literature survey. *Computer Sciences Technical Report, 1648. University of Wisconsin-Madison, Department of Computer Sciences*, 2009.

Judy Hanwen Shen, Agata Lapedriza, and Rosalind W. Picard. Unintentional affective priming during labeling may bias labels. In *International Conference on Affective Computing and Intelligent Interaction*, pages 587–593, 2019.

Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

Liam Steadman, Nathan Griffiths, Stephen Jarvis, Mark Bell, Shaun Helman, and Caroline Wallbank. Kd-str: A method for spatio-temporal data reduction and modelling. *ACM/IMS Transactions on Data Science*, 2(3), 2021.

Latanya Sweeney. Simple demographics often identify people uniquely. *Carnegie Mellon University, Data Privacy*, 2000.

Sean N. Talamas, Kenneth I. Mavor, and David I. Perrett. Blinded by beauty: Attractiveness bias and accurate perceptions of academic performance. *PLOS ONE*, 11(2):1–18, 2016.

Xiu Tang, Sai Wu, Gang Chen, Ke Chen, and Lidan Shou. Learning to label with active learning and reinforcement learning. In *Database Systems for Advanced Applications*, pages 549–557. Springer International Publishing, 2021. ISBN 978-3-030-73197-7.

Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. Eliminating background-bias for robust person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5794–5803, 2018.

Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(11):1958–1970, 2008.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6459, 2018.

Ihsan Ullah, Dustin Carrión-Ojeda, Sergio Escalera, Isabelle Guyon, Mike Huisman, Felix Mohr, Jan N. van Rijn, Haozhe Sun, Joaquin Vanschoren, and Phan Anh Vu. Meta-album: Multi-domain meta-dataset for few-shot image classification. In *Advances in*

Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, 2022.

Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.

Jesper E. van Engelen and Holger H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

Joaquin Vanschoren, Jan N Van Rijn, Bernd Bischl, and Luis Torgo. OpenML: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

Paul Voigt and Axel von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer, 2017.

Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, 2022.

Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, 130(7): 1790–1810, 2022.

Jiannan Wang, Guoliang Li, and Jianhua Fe. Fast-join: An efficient method for fuzzy token matching based string similarity join. In *IEEE International Conference on Data Engineering*, pages 458–469, 2011.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

Mark D. Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, 2016.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021*, pages 3661–3671. IEEE, 2021.

Wen Xia, Hong Jiang, Dan Feng, Fred Douglis, Philip Shilane, Yu Hua, Min Fu, Yucheng Zhang, and Yukun Zhou. A comprehensive study of the past, present, and future of data deduplication. *Proceedings of the IEEE*, 104(9):1681–1710, 2016.

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P Bennett. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416:244–255, 2020.

Shen Yan, Di Huang, and Mohammad Soleymani. Mitigating biases in multimodal personality assessment. In *International Conference on Multimodal Interaction (ICMI)*, page 361–369, 2020.

Show-Jane Yen and Yue-Shi Lee. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3):5718–5727, 2009.

De Jong Yeong, Gustavo Velasco-Hernandez, John Barry, and Joseph Walsh. Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors*, 21(6):2140, 2021.

Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(9):5866–5885, 2022.

Jing Zhang, Xindong Wu, and Victor S. Sheng. Learning from crowdsourced labeled data: A survey. *Artificial Intelligence Review*, 46(4):543–576, 2016.

Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(11):1330–1334, 2000.

Zehui Zhao, Laith Alzubaidi, Jinglan Zhang, Ye Duan, and Yuantong Gu. A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. *Expert Systems with Applications*, 242:122807, 2024.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

How to judge a competition: Fairly judging a competition or assessing benchmark results

Adrien Pavão ADRIEN.PAVAO@GMAIL.COM
LISN, CNRS
Université Paris-Saclay
France

Reviewed on OpenReview: *No*

Abstract

This chapter addresses how to minimize randomness in competition or benchmark judging. We discuss scoring metrics, size of test data, error bars, splitting into phases, and score aggregation methods. Our approach blends theoretical insights with practical guidelines, aiming to provide a clear framework for effective decision making and reduced uncertainty.

Keywords: evaluation, metric, ranking, error bars, staged competitions

1 Introduction

Machine learning competitions, in many ways, resemble sport events. From the perspective of the organizers, much like in sports, there's a pursuit to rank participants fairly based on their skillset and adherence to a specific set of rules. One of the primary objectives in these scientific competitions is not just to crown a winner, but to address a particular problem or answer a scientific question. The emphasis lies in judging the participants justly and, even more crucially, in assessing the merit of their proposed solutions—the models—to make noteworthy advancements in problem-solving. It's vital to design tasks and evaluate contributions in a manner that fosters competitions with substantive outcomes rather than ones that can be exploited, intentionally or otherwise. A well-structured scientific competition can serve various purposes, such as stimulating research in a field or promoting a specific research line.

Beyond the goal of generating significant, reproducible, and universal results, there's a legal aspect to consider. In many jurisdictions, gambling is stringently regulated. As competition organizers, it's essential to distinguish these contests from games of chance. Often, the competition rules explicitly state: "this is a skill-based contest in which chance plays no role." However, this isn't always entirely accurate. Indeed, the cash prize might be awarded to a winner whose score is not significantly distinct from other competitors, as illustrated by the Wheat Detection Challenge (David et al., 2020), later in this chapter.

This chapter explores key aspects of organizing such competitions, including the selection of the scoring metric (Section 2), determining the statistical significance of results (Section 3), and addressing the challenge of score aggregation across various criteria (Section 4). For clarity, the structure of the chapter is illustrated by the Figure 1.

Figure 1: Structure of the chapter 4.

2 What metric for what purpose?

The evaluation metric is, obviously, the core of a challenge: it produces the value that everybody is trying to optimize. This is why the choice of the metric is one of the most important part of the definition of a problem. We have observed that the final ranking of methods is usually very sensitive to the choice of metric (Caruana and Niculescu-Mizil, 2004). This is why this choice needs to be made carefully: an unsuitable metric results in unsuitable solutions. Let's take some examples of such errors of design.

So, how to choose a metric that fits the problem? The variety of metrics that have been proposed in the literature is so large that is hard to find one's way. Thus, we present here the most used metrics in several area of machine learning and explain their main qualities and shortcomings. General view and survey of evaluation methods are provided by survey (Raschka, 2018; Hernández-Orallo et al., 2012).

In the following, we distinguish between performance metrics (e.g. accuracy), ethical and societal impact metrics (e.g. measure of fairness), resources consumption metrics (e.g. time consumption) and evaluator-centric metrics (e.g. human evaluation). Note that many classification, regression and clustering metrics are implemented[1] in the famous machine learning Python package Scikit-Learn (Pedregosa et al., 2011).

2.1 Performance metrics

In this section, we describe metrics commonly used as primary objective in classical machine learning problems: classification, regression, reinforcement learning and unsupervised learning.

Classification

A prediction task is called a *classification problem* when the possible outcomes to predict are grouped in different classes (Grandini et al., 2020). The simplest setting involves only two classes (binary classification, reviewed by Berrar (2019); Canbek et al. (2017)); classification tasks involving more than two classes are called *multi-class classification*. For instance, the classical problem of handwritten digits recognition (LeCun and Cortes, 2005) is a multi-class classification. In *multi-label classification*, each data point can be classified into several classes at the same time. The goal is to use available data called X to obtain the best prediction $\hat{Y}$ of the outcome variable Y. In multi-class classification, Y and $\hat{Y}$ can be seen as two discrete random variables that assume values in $\{1, \ldots, K\}$ where each number represents a different class, and K is the number of distinct classes.

When classification algorithms output the probability that a sample from X belongs to a given class; a classification rule is then employed to assign a single class. In binary classification, a threshold can be used to decide the predicted class. In the multi-class case, there are various possibilities, the most employed technique being selecting the highest probability value, commonly computed using the softmax function (Grandini et al., 2020).

To give a general overview of the resemblance between classification metrics, we conducted an experiment to empirically evaluate the correlation between common classification scoring metrics: we computed rankings of the final models from AutoDL Challenge, independently on all 66 datasets formatted for this competitions, and compared these rankings using Euclidean distance. The results are presented graphically in Figure 2, scaled in a two-dimensional plot. *Jaccard score* and *F1-score*

1. `https://scikit-learn.org/stable/modules/model_evaluation.html`

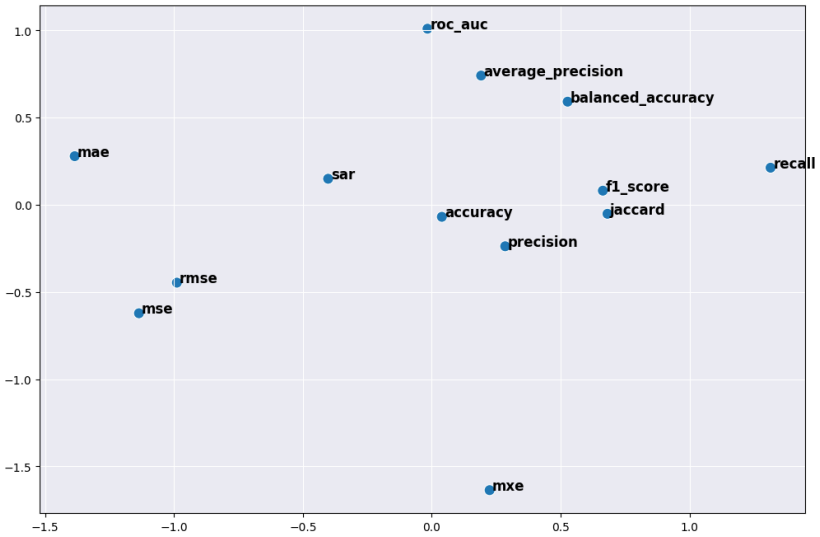


Figure 2: Multidimensional scale (MDS) plot illustrating the degree of correlation between scoring metrics in a 2D space. The metrics are compared by computing Euclidean distance between the rankings they produced. This experiment was performed on the classification tasks and models from AutoDL Challenge (Liu et al., 2021).

are confirmed to be very close[2]. The *SAR* metric (Caruana and Niculescu-Mizil, 2004), an average between *accuracy*, *AUC* and *root mean squared error* (using $1 - RMSE$), is well centralized, as it was designed to be. Interestingly, *balanced accuracy* seems to be centralized between *accuracy* and *AUC*. Loss functions, such as *mean absolute error* (MAE) and *mean squared error* (MSE), are clearly distinct from classification metrics.

Selecting the right scoring metric for a classification task is important as it directly impacts the evaluation and interpretation of the model's performance. The main points we identified as relevant to guide the selection process are the type of problem, the class balance and the real world objectives.

Problem type: Consider the type of problem. Most classification metrics are defined as binary classification metrics; however, they can be used to score multi-class problems, by breaking the problem down into multiple binary problems. The problems can be broken down using either *One-vs-One* (OVO) or *One-vs-Rest* (OVR) approaches. In the OVO approach, the pairwise score of all pairs of classes is computed. In the OVR approach, the scores for each class is computed separately, treating each class as the positive class and all other classes as the negative class. In both cases, the problem is broken down into a series of binary problems, and the final score is obtained by averaging all the scores, either using a simple average, or a weighted average. The OVO approach is computationally expensive, as the number of scores to compute is $\frac{K(K-1)}{2}$ with K being the number of different classes. For this reason, the OVR approach is mostly used in practice, needing only K computations, one for each class. We consider only the OVR approach for the rest of this section.

2. https://stats.stackexchange.com/questions/511395/are-jaccard-score-and-f1-score-monotonically-512378#512378

Beyond its computational benefits, the OVR approach is favored due to its simplicity, making it more interpretable for non-experts. It scales linearly as the number of classes grows, contrasting with the exponential complexity of the OVO method. Furthermore, its prominence in modern machine learning tools, given the optimized implementations in popular frameworks, emphasizes its relevance for real-world applications.

The simplest idea when it comes to score classifiers is to use *rates of success*. *Precision*, *recall* and *accuracy* are metrics that simply count the successes and failures of the classifier. Intuitively, *accuracy* is the likelihood that a randomly chosen sample will be correctly classified by the model. The fundamental component of this metric is the individual units in the dataset, each of which holds equal weight and contributes equally to the score. However, when considering classes instead of individuals, some classes may have a high number of units while others have only a few. In such cases, the larger classes will carry more weight compared to the smaller ones. When the dataset is imbalanced, meaning that most units belong to one particular class, *accuracy* may overlook significant classification errors for classes with fewer units as they are less significant compared to the larger classes.

Class imbalance: Consider the distribution of classes in the dataset. If there is a class imbalance, using accuracy as a metric might not provide an accurate picture of the model's performance. In such cases, metrics like *balanced accuracy* or *area under ROC curve* (AUC) are more appropriate. *Balanced accuracy* addresses this issue by giving each class equal impact on the score. This is simply done using a weighted average of each class accuracy, weighted by the proportion of the class in the test set. Despite having fewer units, smaller classes may have a disproportionately larger influence on the formula. When the dataset is relatively balanced, meaning the class sizes are similar, *accuracy* and *balanced accuracy* tend to produce similar results. The main distinction between the two metrics becomes apparent when the dataset exhibits an imbalanced distribution of classes.

Real world objective: Consider the real world impact of the problem and the cost of false positive and false negatives. For instance, in medical diagnosis, a false positive (e.g. a healthy person is diagnosed with a disease) can lead to unnecessary medical procedures and treatments, causing harm to the patient. On the other hand, a false negative (e.g. a sick person is not diagnosed) can result in a delay in treatment and a potentially fatal outcome. In this case, *recall* (the ability of the model to identify all positive cases) is a relevant metric. Another example: in the finance sector, false positives can lead to substantial revenue loss from incorrect trades or transactions, while false negatives might result in missed opportunities. In such contexts, *precision* (the model's capability to correctly identify positive instances) and *F1 score* (a harmonized metric between precision and recall) could be considered better choices for assessing the model's efficiency.

Regression

The prediction task is called a regression problem when the outcome is a continuous numeric value, in contrast to a classification problem where the variable to predict falls into discrete categories. For instance, predicting temperatures or road traffic from support variables are regression tasks. A review of metrics for regression is given by (Botchkarev, 2018). Another survey concerns evaluation methods for time series forecasting (Cerqueira et al., 2020).

The main points we identified as relevant to guide the selection process are the type of problem, the scale of the target variable and the real world objectives.

Problem type: Consider the type of problem you are trying to solve. For example, for time-series forecasting problems, metrics like *mean absolute error* (MAE), *mean squared error* (MSE), and *root mean squared error* (RMSE) are relevant. For problems involving predicting counts, metrics like *mean absolute percentage error* (MAPE) and *symmetric mean absolute percentage error* (SMAPE) might be more appropriate.

Scale of the target variable: Consider the scale of the target variable. If the target variable is large, the absolute difference between the actual and predicted values will also be large, making *MAE*, *MSE* and *RMSE* less interpretable. In such cases *R-squared* (coefficient of determination) might be more appropriate metrics. *R-squared* has no units, can be compared among different tasks, and has an intuitive interpretation.

Real world objective: Once again, consider the real world problems. For instance, if the problem involves predicting stock prices, the magnitude of the error is more important than the direction of the error. In such cases, *RMSE* or *MSE* might be appropriate metrics. Another example: in the field of climate modeling, the right choice of scoring metric vary depending on the problem. If the goal is to predict global temperatures, metrics like *mean absolute error* (MAE) could be used to evaluate the performance of the model. However, in predicting regional precipitation patterns, metrics like *R-squared* or *explained variance score* might be used to evaluate the ability of the model to capture the complex spatial patterns of precipitation.

Reinforcement learning

Reinforcement learning (RL) consists, for an autonomous agent (e.g. robot), in learning what actions to take, based on experiences, in order to optimize a quantitative reward over time. The agent is immersed in an environment and makes its decisions based on its current state. In return, the environment provides the agent with a reward, which can be either positive or negative. The agent seeks, through iterated experiments, an optimal strategy or *policy*, which is a function that associates the current state with the action to be performed, to maximize the sum of rewards over time. This setting makes the scoring procedure particularly problem-specific and prone to design flaws.

In reinforcement learning, overfitting occurs when the agent becomes too specialized to the conditions of the training environment and is unable to generalize to unseen situations. This is a common problem in RL because the training and the evaluation processes are usually conducted on the same environment, rather than in two separate environments. This creates a situation where the agent can memorize the optimal actions in the training environment without truly understanding the underlying dynamics. As a result, the agent's performance may appear to be much better than it actually is, leading to a biased evaluation. This can be a serious issue in RL, as it undermines the validity of the evaluation process and may result in the selection of sub-optimal algorithms or policies. To mitigate this problem, it is important to separate the training and testing environments and use different metrics and simulators for evaluation. This helps to unbias the evaluation process and to ensure that the results accurately reflect the true performance of the agent.

Some research studying the evaluation of RL algorithms show that behavioral metrics play a crucial role in determining the quality of a state representation, and in learning an optimal representation (Jordan et al., 2020; Lan et al., 2021). Existing methods, such as *approximate abstractions* and *equivalence relations*, aiming at reducing the size of the state or action space by aggregating similar states, are not effective for continuous-state reinforcement learning problems, due to their inability to maintain the continuity of common RL functions and their tendency to generate overly

detailed representations that lack generalization. A behavioral metric in reinforcement learning is a measure of an agent's performance in an environment, based on its actions and observed rewards. It can be used to evaluate and compare different reinforcement learning algorithms or policies. Examples include *average reward per episode*, *success rate*, and *convergence speed*.

According to Henderson et al. (2018), multiple trials with different random seeds are necessary to compare performance, due to high variance. To ensure reproducibility, it is crucial to report all hyperparameters, implementation details, experimental setup, and evaluation methods for both baseline comparisons and novel work.

Unsupervised learning

Numerous tasks in machine learning lack a definitive ground truth for evaluating solutions. Termed as unsupervised learning, this category includes a diverse range of tasks, from clustering and dimensionality reduction to data modeling, generation, and feature extraction. It also covers areas with a more subjective nature, like automatic music composition, identifying molecules that bind to COVID-19, and text summarization.

It is particularly challenging to design competitions for such unsupervised learning tasks, due to the following reasons: the lack of ground truth, the diversity of solutions, and the subjectivity in evaluation. Indeed, unlike supervised learning, unsupervised learning often lacks clear ground truth, making it difficult to evaluate the results objectively. Also, unsupervised learning models can often produce diverse solutions that are equally valid, making it challenging to select a single best solution. Finally, the evaluation of unsupervised learning solutions can be subjective as it depends on the understanding and interpretation of the evaluators.

In the case of distribution modelling and clustering, some clear performance metrics can be identified. Prior research investigate the evaluation metrics for unsupervised learning (Palacio-Niño and Berzal, 2019) and clustering (Ben-Hur et al., 2002; von Luxburg et al., 2012). Typically, when the goal in unsupervised learning is to **learn the underlying data distribution**, various loss functions can be employed depending on the specific type of model:

- *Maximum Likelihood Estimation* (MLE): In probabilistic models, the goal is often to maximize the likelihood of the observed data.
- *Kullback-Leibler (KL) Divergence* (Kullback and Leibler, 1951): Measures the difference between two probability distributions. It is often used with Variational Autoencoders (VAEs) (Kingma and Welling, 2019) where one tries to minimize the divergence between the learned distribution and the true data distribution.
- *Reconstruction Loss*: In models like autoencoders (Liou et al., 2014), the objective is to reconstruct the input data from a compressed representation. The loss measures the difference between the original input and its reconstruction.
- *Wasserstein Distance* (Earth Mover's Distance) (Villani, 2009): Used in Wasserstein Generative Adversarial Networks (WGANs) (Arjovsky et al., 2017) to measure the distance between the generated distribution and the true data distribution.

To assess the performance of **clustering** methods, when no ground truth is available, the following metrics and techniques can be used:

- *Inertia* (Sum of Squared Errors): Inertia measures the sum of squared distances between each data point and its closest cluster center. Lower inertia values indicate tighter clusters and better performance.
- *Silhouette Score* (Rousseeuw, 1987): This metric computes the average silhouette value for all data points, measuring how similar a data point is to its own cluster compared to other clusters. A silhouette score close to 1 indicates a well-partitioned dataset, whereas a score close to -1 indicates poor clustering.
- *Davies-Bouldin Index* (Davies and Bouldin, 1979): This index measures the ratio of within-cluster scatter to between-cluster separation. Lower values of Davies-Bouldin Index indicate better clustering performance.
- *Calinski-Harabasz Index* (Caliński and Harabasz, 1974): This index evaluates clustering by comparing the ratio of between-cluster dispersion to within-cluster dispersion. Higher values of the Calinski-Harabasz Index suggest better clustering performance.

Although these metrics are valuable for tasks like distribution learning and clustering, they fall short in assessing the performance of machine learning models in areas like text summarization or artistic creation. In these scenarios, it is challenging to score the success. Interesting and effective methods include human evaluation, employing other machine learning models as evaluators, and using interactive adversarial frameworks.

Indeed, **human evaluation** can play a crucial role in assessing the performance of unsupervised learning algorithms when traditional quantitative metrics may not fully capture the desired outcomes, or when ground truth labels are not available. Human evaluation can provide valuable qualitative insights and help ensure that the resulting patterns or structures discovered by the algorithms align with human intuition and domain knowledge. To perform human evaluation effectively, it is essential to establish clear guidelines for the evaluators, provide proper training, and, when possible, recruit multiple evaluators to increase the reliability of the assessment. Human evaluation can be time-consuming and resource-intensive, so it is often used in combination with quantitative metrics to balance the efficiency and quality of evaluation. This approach is further discussed in Section 2.4.

Another interesting technique, that may become more popular in the future, is to use a **model used as a metric**. For instance, a classifier trained on discriminating between two distributions (e.g. fake and real) can be used to evaluate the performance of generative models. This can be compared to the functioning of Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), where the output of a binary classifier is used to guide the learning of a generative model. When using another machine learning model as a metric for unsupervised learning, it is important to remember that the evaluation depends on the performance of the supervised model and the quality of the labeled data. Therefore, the results should be interpreted with caution, as the evaluation might be biased or limited by the chosen supervised model or the available data. This approach is discussed in Section 2.4.

Finally, adversarial challenges, where the solutions proposed by the participants are then used as input data in the next phase, is potentially a good way of organizing challenges on unsupervised learning tasks. Adversarial challenges design are explored in details in Chapter 12.

More generally, to overcome these challenges and organize unsupervised learning competitions and benchmarks, it is essential to articulate the problem and describe the data comprehensively.

Evaluation methods must be defined considering the missing ground truth. Promoting collaboration among participants can diversify the solutions. Involving domain experts in the evaluation can enhance the result objectivity. Keeping participants informed about competition progress and being open to adjustments can optimize the process. Lastly, providing evaluations using diverse metrics can help participants gauge the pros and cons of their methods. By following these steps, competitions can be designed to effectively evaluate the solutions to an unsupervised learning task while overcoming the challenges of lack of ground truth, diversity of solutions, and subjectivity in evaluation.

2.2 Ethical and societal impact metrics

Traditional evaluation metrics like *accuracy*, *precision*, and *recall* have long held the spotlight, but there is so much more to consider when measuring a model's real-world impact. In this section, we dive into the lesser-known, unconventional metrics that are reshaping the way we assess machine learning models. From fairness and privacy to interpretability and calibration, these innovative evaluation techniques change how we think about model performance and pave the way for a more responsible, holistic approach to machine learning. Indeed, the performance metrics we have reviewed in this chapter so far reflect only one aspect of the performance of the models: their predictive abilities. Yet, in many applications, our concerns extend to other aspects, like the trustworthiness of the algorithms. This is particularly true in sensitive applications, where an algorithmic decision could mean life or death.

FAIRNESS

Even if the mathematical definition of machine learning models does not necessarily contains unfair or biased elements, trained models can be unfair, depending on the quality of their input data or their training procedure. A model trained on biased data may not only lead to unfair and inaccurate predictions, but also significantly disadvantage certain subgroups, and lead to unfairness. In other words, the notion of fairness of models describes the fact that models can behave differently on some subgroups of the data. The issue is especially significant when it pertains to demographic groups, typically defined by factors such as gender, age, ethnicity, or religious beliefs. As machine learning is increasingly applied in society, this problem is getting more attention and research, and is subject to debate (Benz et al., 2020; Vasileva, 2020; Chouldechova and Roth, 2018; Chen et al., 2018; Boratto et al., 2021). Some interesting ways to quantify fairness include:

Demographic Parity: This measure checks if the positive classification rate is equal across different demographic groups. The formula is as follows:

$$DemographicParity : P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$$

where A is a protected attribute (such as race or gender), Y is the target variable (such as approval or denial) and where $\hat{Y}$ is the predicted value of Y. Demographic parity is a condition to be achieved: the predictions should be statistically independent of the protected attributes.

Statistical Parity Difference: Related to the demographic parity, it measures the difference between positive classification rate across different demographic groups. The formula is:

$$StatisticalParityDifference = P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)$$

Disparate impact: It calculates the ratio of the positive classification rate for a protected group to the positive classification rate for another group. It is similar to the *statistical parity difference*, but it is a ratio instead of a difference:

$$DisparateImpact = \frac{P(\hat{Y} = 1|A = 0)}{P(\hat{Y} = 1|A = 1)}$$

A value of 1 indicates that the positive classification rate is the same for both groups, suggesting fairness. A value greater than 1 indicates a higher positive classification rate for the group with $A = 0$, while a value less than 1 suggests a higher positive classification rate for the group with $A = 1$. However, it is important to note that disparate impact is a limited measure of fairness and should not be used on its own. There may be cases where a higher positive classification rate for one group is justifiable, for example if the group is underrepresented in the training data. Additionally, disparate impact does not consider other factors such as false positive and false negative rates, which may provide a more comprehensive view of fairness.

Equal Opportunity: This metric checks if the true positive rate is equal across different demographic groups. The formula is:

$$EqualOpportunity : P(\hat{Y} = 1|Y = 1,\ A = 0) = P(\hat{Y} = 1|Y = 1,\ A = 1)$$

As for *demographic parity*, it is a condition to be achieved.

Equal Opportunity Difference: This metric measures if the true positive rate is equal across different demographic groups. The formula is:

$$EqualOpportunityDifference = P(\hat{Y} = 1|Y = 1,\ A = 0) - P(\hat{Y} = 1|Y = 1,\ A = 1)$$

The same idea can be applied to false positive rates.

These are just a few of the metrics that can be used to quantify fairness in classification tasks. It is important to note that fairness is a complex issue, and these metrics should not be used in isolation. Instead, they should be considered in the context of the specific problem and the desired outcome.

CALIBRATION

Classifiers usually return probabilities to indicate the confidence levels across different classes. However, the question arises whether these confidence levels accurately reflect the classifier's actual performance. As defined by Naeini et al. (2015); Guo et al. (2017), the notion of miscalibration represents the difference in expectation between the confidence level (or probability) returned by the algorithm, and the actual performance obtained. In other words, calibration measurement answers the following question: is the confidence of the algorithm about its own predictions correct? Promoting well calibrated models is important in potentially dangerous decision making problems, such as disease detection or mushroom identification. The importance of calibration measurement lies in the fact that it is essential to have a clear understanding of the confidence level that the algorithm has in its own predictions. A well-calibrated algorithm will produce confidence levels that accurately reflect the likelihood of a prediction being correct. In contrast, a miscalibrated algorithm

will either over or under estimate its confidence in its predictions, leading to incorrect or unreliable outcomes. In applications where the consequences of incorrect decisions can be severe, it is of utmost importance to have a well-calibrated algorithm. Misclassification of a disease can lead to incorrect medical treatment and harm to the patient. Similarly, misidentification of a mushroom can result in serious health consequences. In these scenarios, well-calibrated models can help ensure that the decisions are made based on reliable predictions.

The calibration can be estimated using the **Expected Calibration Error** (ECE): this score measures the difference between the average predicted probability and the accuracy (i.e., the proportion of positive samples) in bins of predicted probability. The formula for the ECE is given by:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc_m - conf_m|$$

where M is the number of bins, B_m is the set of samples in the m^{th} bin, n is the total number of samples in the test data, acc_m is the accuracy of the m^{th} bin, and $conf_m$ is the average predicted probability in the m-th bin.

When computing the calibration, we derive the performance prediction directly from the model's output. Another interesting possibility is to ask the participants to provide an estimation of the generalization score of their method. Indeed, we can make a connection between the calibration and the prediction of generalization error, more commonly estimated by a separated method. The Performance Prediction Challenge (Guyon et al., 2006) focused on this problem.

Interpretability and explainability

Given the complexities of machine learning models, the assessment of their interpretability and explainability emerges as a considerable challenge. While both concepts are crucial for ensuring trust and understanding in model predictions, especially in critical applications, measuring them accurately is difficult, especially when models vary widely in their structures and underlying mechanisms. Interpretability and explainability are related but distinct concepts in machine learning.

Interpretability refers to the degree to which a human can understand the cause of a model's predictions. It refers to the ability to understand the internal workings of the model and how it arrived at its decisions.

Explainability refers to the ability to provide a human-understandable explanation of the model's decision making process. It is concerned with the presentation of the reasons behind the predictions to humans in a understandable form, e.g. through feature importance.

In summary, interpretability focuses on the transparency of the model itself, while explainability focuses on the communication of the model's behavior to a human audience. A wide survey on interpretability is proposed by Carvalho et al. (2019). They stressed out how interpretability is greatly valuable in one hand, but hard to define in the other hand. Another way to explain algorithms, automatically, is the sensitivity analysis (Iooss et al., 2022). Sensitivity analysis is a technique used to determine how changes in input variables of a model or system affect the output or outcomes of interest.

Past competitions have been exploring the development and evaluation of explainable models, such as the *Job Candidate Screening Challenge* (Escalante et al., 2017, 2018). It is a challenge of first impressions and apparent personality analysis, on audio-visual data. The candidate models

have to predict apparent traits of people[3] (e.g. friendly or reserved, imaginative or practical) from short videos, with a focus on the explanatory power of techniques: solutions have to "explain" why a given decision was made. To this end, participants had to provide a textual description that explains the decision (i.e. the prediction) made. Optionally, participants could also submit a visual description to enrich and improve clarity and explainability. Performance was evaluated in terms of the creativity of participants and the explanatory effectiveness of the descriptions. For this evaluation, a set of experts in the fields of psychological behavior analysis, recruitment, machine learning and computer vision was invited. We can note that, this way, the explainability component of the challenge requires qualitative evaluations and hence human effort.

It is worth noting that some models are interpretable by nature, such as logistic regression or decision tree. Some researchers make the point that there is a trade-off between interpretability and models' performance, especially for complex tasks, that seem to be requiring blackbox models – huge deep learning neural networks. However, Rudin (2019) argues that this "accuracy-interpretability trade-off" is an unfounded myth.

Privacy

Privacy must typically be measured when the candidate algorithms are generative models, modelling a distribution of potentially confidential data. The goal in such a case is to use the generative models in order to create artificial data that reassemble sufficiently the real data to be used in actual applications, but not too much for private information to be leaked. A metric to estimate this trade-off is the **adversarial accuracy**, that we introduced in Yale et al. (2019). Here is its definition:

$$AA_{TS} = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left(d_{TS}(i) > d_{TT}(i)\right) + \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left(d_{ST}(i) > d_{SS}(i)\right)\right)$$

where the indicator function $\mathbf{1}$ takes the value 1 if its argument is true and 0 otherwise, T and S are true and synthetic data respectively. d is an arbitrary chosen distance function, such as the Euclidean distance. $d_{TS}(i)$ represents the distance between the i^{th} point of T and its closest neighbor from S. $d_{TT}(i)$ is the distance between this i^{th} point and its closest neighbor in T. Subsequently, $d_{ST}(i)$ and $d_{SS}(i)$ compare the i^{th} point of S to its closest neighbors in T and in S.

It is basically the accuracy of a 1-nearest-neighbor classifier, but the ideal score is not 1 (perfect classification accuracy) but 0.5. Indeed, a perfect score means that each generated data point has its closest neighbor in the real data, which means that the two distributions are overly similar. A score of 0 would mean that the two distributions are too different, thus the practical utility is low. Hence, a 0.5 score, where the closest neighbor of each data point can either be fake or real with the same probability, is what guarantees a good privacy. These principles are illustrated with a toy example in Figure 3. Alaa et al. (2022) proposes a similar approach.

One limitation of this method is that a proper measure of distance is needed. This is also a strength because it means that the method is general and can be applied in different fields, by selecting an adequate distance measure.

In the study of privacy, *differential privacy* and *membership inference attacks* are core concepts.

Differential privacy provides a robust framework to ensure that a trained model does not get substantially influenced by the inclusion or exclusion of a single data sample from the dataset. It employs a parameter ε to quantify the privacy, with smaller ε values meaning more privacy

3. The data was labeled by around 2500 annotators.

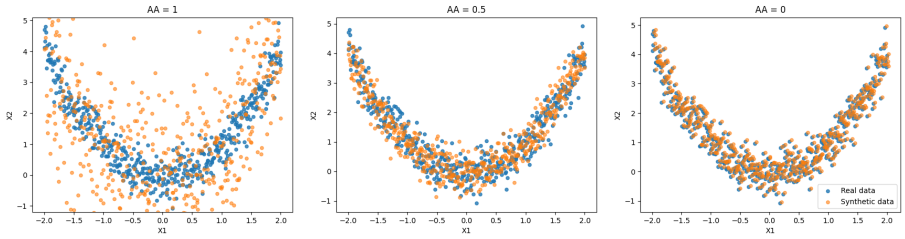


Figure 3: Adversarial Accuracy (AA) is the performance of an nearest neighbor classifier that distinguishes between real data vs synthetic data. The ideal value is $AA = 0.5$ (represented by the plot in the middle). This is a bi-variate ($X1$ and $X2$) illustrative example.

guarantees. It relates to the metric we proposed, which estimates the proximity between generated data and training data. While our approach evaluate the privacy preservation at the scale of the dataset, differential privacy focuses on individual-level privacy.

Inference attacks represent methods by which attackers deduce sensitive information using the models' output (or predictions). The adversarial accuracy metric intends, in some way, to measure the sensitivity of the model to these inference attacks. If generated data were nearly identical to training data, adversaries might be able to get sensitive information. By ensuring a degree of distance between original and generated data, the resistance against such attacks is improved.

Additionally, privacy concerns are not limited to synthetic data or generative models. Predictors too can be the target of privacy attacks, as highlighted by Pedersen et al. (2022).

2.3 Resources consumption metrics

Resource consumption metrics quantify the energy, time, and data utilized by models, thereby providing insights into their efficiency and sustainability.

Time and memory consumption

It is useful to include the consumption of time, memory and energy of ML models in their evaluation and comparison. There are two main approaches to take these into account: **limit the resources** and **track the use of resources**. Both may imply in practice the use of code submission, as opposed to results submission. In code submission competitions, the participants submit their models which then get trained and tested on the servers, while in results submission competitions, the participants work locally and upload their predictions to the platform. Code submission is therefore advised if limiting or tracking the resource usage is part of the competition design.

The **training and inference time**, the **size of the model**, the **memory used** during the process or even the **energy consumption** are variables that can be limited by design or measured and shown on the leaderboard. Obviously, using the same hardware and evaluation conditions for all participants is needed in order to have a fair evaluation. The number of lines of code, or the number of characters, can also be used as an indicator of the **simplicity and practicability** of the solution. However, obviously, this indicator can be easily tricked by calling external packages and may need

a manual review. The simplest models that solve the task are preferable, for being less harmful for the environment, less costly, deployable in weaker devices and easier to interpret.

A model that can produce the same results in less time is more desirable, as it reduces the computational resources required and can lead to cost savings. This is especially important in light of the current ecological crisis, as reducing energy consumption in computing can have a significant impact on reducing the carbon footprint of technology. Additionally, models that are faster to train and make predictions are more scalable and can be deployed in real-time applications, further enhancing their utility. Thus, optimizing time consumption is a key factor in the development of efficient and environmentally sustainable machine learning models.

ANYTIME LEARNING

Anytime learning refers to a learning paradigm in which a machine learning model or algorithm incrementally improves its performance as it receives more data or training time. The key aspect of anytime learning is that the model can produce meaningful results at any point during the learning process, with its performance generally improving as it acquires more data or spends more time on training. Anytime learning algorithms are particularly valuable in situations where resources, such as time or computation power, are limited, or when it is essential to provide real-time or near-real-time insights. These algorithms can be employed in various machine learning settings, including classification, regression, and reinforcement learning tasks.

To evaluate the score in this framework, one can compute the Area under the Learning Curve (ALC):

$$ALC = \int_{t_0}^{t_f} s(t)dt$$

where $s(t)$ is the performance score (obtained form a metric depending on the task) at timestamp t. t_0 and t_f refers to the first and last timestamps, and should be fixed to allow a fair evaluation and comparison of the scores.

The time can be linear, or transformed at any scale:

$$ALC = \int_{t_0}^{t_f} s(t)d\tilde{t}(t)$$

where $\tilde{t}$ is the transform function. For instance, a logarithmic scale transform can be used, in order to give more importance to the first steps of training:

$$\tilde{t}(t) = \frac{\log(1+t)}{\log(1+t_f)}$$

The metric is depicted in Figure 4. Although both *Model 1* and *Model 2* converge to the same score, *Model 1* is favored in an anytime learning context due to its rapid performance improvement across epochs. This advantage is evident from the larger area under its learning curve.

Any-time learning can be linked to multi-fidelity. Multi-fidelity methods in machine learning refer to strategies that use various "fidelities" or qualities of data or models to speed up the learning process. For instance, one might use a simpler, faster-to-evaluate but less accurate model (low fidelity) to guide the learning process in the early stages, and a more complex, accurate but computationally intensive model (high fidelity) later on. The idea is to use less expensive resources to get

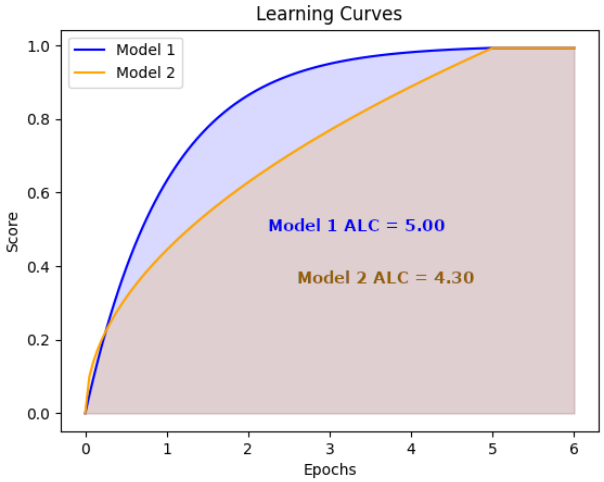


Figure 4: Example of learning curves for two models. While both *Model 1* and *Model 2* converge to the same score, *Model 1* boasts a larger area under the learning curve (ALC). Thus, in an anytime learning context, *Model 1* is the preferred choice.

an initial idea, and then refine it with more accurate but costlier methods. One could combine anytime learning and multi-fidelity methods to create a learning process that is both time-efficient and increasingly accurate. For example, in the early stages of an anytime learning algorithm, one might employ low-fidelity models or data to quickly get a "good enough" model. As time allows, the algorithm could then switch to using high-fidelity models or data to refine its understanding further. This way, one gets the best of both worlds: a quickly usable model, while aiming for high accuracy given more time. This combination could be particularly useful in scenarios where computational resources or time are constrained but where the model quality also needs to be maximized, such as real-time analytics, robotics, or complex simulations.

Data consumption

In machine learning, data is a key resource, and examining its consumption is relevant. The data input of models can be divided into two dimensions: the number of samples, and the number of features.

Number of training samples.

The amount of training data required is an interesting metric to consider when comparing different models. As the saying goes, "data is the new oil", but not every situation allows for the luxury of vast datasets. In many real-world applications, gathering sufficient labeled data can be time-consuming, expensive, or even impossible. Tracking and limiting data consumption is, therefore, an essential aspect of model evaluation.

Monitoring data consumption can help identify algorithms that perform well with limited data, making them more suitable for scenarios with small datasets. On the other hand, constraining the quantity of available training data can encourage the development of models that are more efficient

in learning. This is typically called few-shots learning, where meta-learning techniques, such as the k-shot n-way approach, are used. In this method, models are trained to quickly adapt to new tasks using only a limited number of examples k from each class n. This k-shot n-way design was used in MetaDL competition (El Baz et al., 2021). By intentionally limiting data consumption, meta-learning promotes the development of models capable of generalizing better from smaller datasets, ultimately enhancing their utility and adaptability in diverse situations.

Feature selection.

Feature selection is the process of selecting a subset of relevant features, or variables, for use in model construction. Even if recent trends in deep learning tend to use the raw data without further preparation, feature selection is an essential aspect of machine learning that aims at selecting the most informative features for a given model. Proper feature selection can lead to simpler, more interpretable, and faster-performing models that may also have improved generalization.

Two primary methods exist to assess feature selection: **evaluation metrics** and **intrinsic metrics**. Evaluation metrics, also called the *wrapper approach*, assess a model's performance after feature selection, based on the assumption that effective model performances signify well-chosen features. The specific metrics used vary depending on the nature of the problem, such as classification or regression, as elaborated previously in this chapter. On the other hand, intrinsic metrics, also called the *filter approach*, evaluate the inherent quality or relevance of features without necessarily training a model. They act as heuristics of the fundamental information contained within each feature. Intrinsic metrics evaluate the significance of individual features without necessitating a full model. These include the *correlation coefficient*, which gauges the linear relationship between a feature and the target; *mutual information*, indicating the relevance of a feature based on how much it informs about another variable; *feature importance* from tree-based models such as decision trees and random forests; and *variance threshold*, where low-variance features, assumed less informative, might be discarded. While Kohavi and John (1997) advocates for the wrapper approach (evaluating trained algorithms), Tsamardinos and Aliferis (2003) argue that neither approach is inherently better, and that both the learner and the evaluation metric should be considered. Hybrid approaches, such as maximizing a performance score while minimizing the number of features, are efficient in balancing between model accuracy and model simplicity.

In the NIPS Feature Selection Challenge (Guyon et al., 2004; Guyon et al.), participants were ranked on the test set results using a score combining *Balanced Error Rate* (BER), the fraction of features selected F_{feat}, and the fraction of probes found in the feature set selected F_{probe}. The aim of feature selection is often to reduce the feature set's size without a significant loss in predictive performance. Hence, a lower F_{feat} could be seen as favorable if the model's performance remains strong. "Probes" in the context of this challenge refer to "dummy" or "non-informative" features that were intentionally added to the dataset. These features don't have any relation or correlation with the target variable and are essentially noise. Thus, a good feature selection algorithm should ideally avoid selecting these probes, minimizing F_{probes}.

Briefly: they used the McNemar test (McNemar, 1947) to determine whether classifier A is better than classifier B according to the BER with 5% risk yielding to a score of 1 (better), 0 (don't know) or -1 (worse). Ties (zero score) were broken with F_{feat} (if the relative difference in F_{feat} was larger than 5%). Remaining ties were broken with F_{probe}. The overall score for each of the five datasets was the sum of the pairwise comparison scores.

These methods of feature selection, from intrinsic metrics to challenge-specific criteria such as those in the NIPS Feature Selection Challenge, play a role in optimizing machine learning models simplicity, performance and interpretability, and reducing their data consumption.

2.4 Interactive and evaluator-centric metrics

Evaluator-centric metrics represent a paradigm in machine learning evaluation where the benchmarking process actively involves specific evaluators, be they humans or models. Within this category, there is a distinction: human-centric approaches primarily leverage human judgment and perspectives, while model-centric approaches utilize predefined algorithms or models for evaluation.

HUMAN-CENTRIC APPROACHES

In addition to the quantitative evaluation metrics discussed previously, it is essential to consider more "human" evaluation techniques when assessing machine learning models. These approaches place emphasis on qualitative aspects and subjective interpretation, bringing a human touch to the evaluation process. For instance, in the case of text-to-image algorithms, manual inspection of generated images can help determine whether the outcomes are visually appealing, coherent, and contextually relevant. More generally, models that produce art, such as automatic music generators, benefit from manual evaluation. AI art competitions typically involve human evaluation through voting, such as the *Deep Art* competition of NeurIPS 2017[4]. IEEE CEC also hosts regular art competitions [5]. Similarly, large language models can be subjected to psychological or behavioral tests, where human evaluators rate the model's responses based on factors such as coherence, empathy, and ethical considerations. Such human-centric evaluation methods can reveal insights that purely numerical metrics might overlook, providing a more nuanced understanding of a model's strengths and weaknesses. It is important to distinguish between human evaluation and the comparison to human performance. While both are human-centric approaches, the latter specifically uses human abilities as a baseline for performance comparisons. A clear example of this approach is how the Generative Pre-trained Transformers (GPT) (OpenAI, 2023), the famous large language models, was tested using psychology tests (Uludag and Tong, 2023; Li et al., 2022), high-school tests (de Winter, 2023) and mathematics tests (Frieder et al., 2023). By integrating these human-oriented techniques into our evaluation toolbox, we can ensure that our machine learning models are not only effective in solving problems but also resonate with the multifaceted nature of human experiences.

MODEL-CENTRIC APPROACHES

A model can be used as a metric to evaluate the performance of other models, offering a dynamic and specialized approach to scoring complex tasks. This approach can be called model-centric, as opposed to human-centric approaches discussed in the previous section.

Typically, discriminative models can be used to assess the performance of generative models. Examples of this were given with the use of k-nearest neighbor adversarial accuracy to compute privacy, and the use of a classifier to score an image generation task. More generally, in this case, the

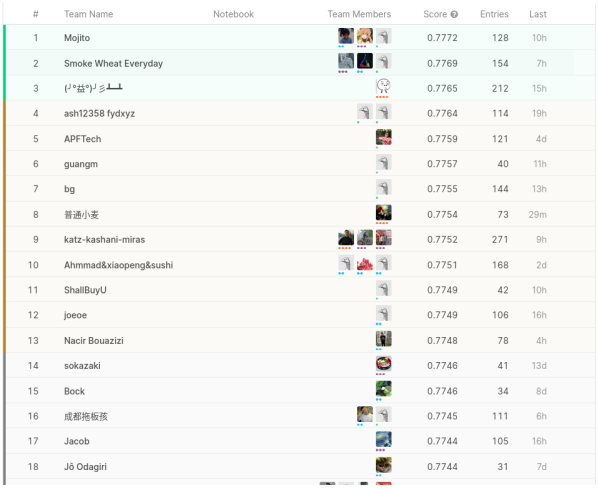
4. https://nips.cc/Conferences/2017
5. https://sites.google.com/view/ieeecec2021ecmac/

discriminant is trained to distinguish between real data from the target distribution and artificially generated data, thereby judging how "realistic" the generated data appears to be. This approach is similar to the learning framework of generative adversarial networks (Goodfellow et al., 2014). However, as underscored by the metric of adversarial accuracy for privacy, a generative model that completely deceives the discriminative model – in the sense that the discriminant gets an extremely low score – is an indication of privacy leakage of the training data. In other words, if the generative model performs exceptionally well within this adversarial framework, it raises concerns about its ability to generate general and original data. To avoid this, an "originality" or "privacy" metric should be invoked to measure the similarity, as mentioned in the Section 2.2. One distinct advantage of employing a discriminative model for performance measurement is its ability to output numerical values. Consequently, this form of performance assessment can easily be incorporated into a ranking system or any quantitative evaluation framework. An illustration of this protocol can be found in the Dog Image Generation challenge (Kan et al., 2019).

Another, more qualitative, way of measuring performance using models emerges through the use of language models. Large Language Models (LLMs) can serve as evaluators on various Natural Language Processing (NLP) tasks. For instance, in text summarization, an LLM can be employed to measure the semantic coherence and relevance of generated summaries by comparing them with the original text. The LLM could produce a likelihood score or even generate textual critiques to indicate how well the summary captures the essence of the source material. When it comes to explainability, an LLM can analyze the output explanations of complex models to assess their clarity and coherence. Naturally, as with any model-based metric, the initial prerequisite is to have confidence in the reliability of the evaluating model. More broadly, LLMs can act as judges for smaller models in a variety of NLP tasks. They can evaluate the quality of machine-generated translations, assess the sentiment consistency in chatbot dialogues, or even measure the relevance of answers generated by a question-answering system. These ideas were explored by the innovative competition *Auto-Survey Challenge 2023*[6] (Khuong and Rachmat, 2023). In this challenge, the participants propose AI agents capable of composing scientific survey papers and reviewing them. Such AI agents thus operate either as authors or reviewers. API calls to chatGPT were used to output the scores of *conclusion* (how well the conclusion highlights the main findings in the text) and *contribution* (relevance of the paper). To bridge between qualitative and quantitative outputs, the organizers asked the LLM to provide a number in Likert scale (Likert, 1932), for better differentiation between good and bad results (for instance, *1 - Strongly Disagree*, *2 - Disagree*, *3 - Neutral*, *4 - Agree*, *5 - Strongly Agree*). They also made clear and complete prompts, detailing how the characteristics should be evaluated.

Using a model as a metric offers many advantages and may even be essential for certain applications, but it also comes with its notable drawbacks and challenges. One of the immediate concerns is the additional layer of complexity and computational cost involved in using one model to evaluate another, which becomes particularly problematic when computational resources or time are limited. This issue is closely followed by questions regarding the reliability and consistency of the metric model itself. If the model employed as a metric has inherent weaknesses or biases, these could be transferred to the evaluation of the models being evaluated. The method's potential unreliability, complexity, and sensitivity to data and hyperparameters can result in difficult interpretations and risk misleading evaluations, especially in critical fields like healthcare and legal decision-making.

6. `https://www.codabench.org/competitions/1145/`



#	Team Name	Notebook	Team Members	Score	Entries	Last
1	Mojito			0.7772	128	10h
2	Smoke Wheat Everyday			0.7769	154	7h
3	(ﾉ°益°)ﾉ彡┻━┻			0.7765	212	15h
4	ash12358 fydxyz			0.7764	114	19h
5	APFTech			0.7759	121	4d
6	guangm			0.7757	40	11h
7	bg			0.7755	144	13h
8	普通小麦			0.7754	73	29m
9	katz-kashani-miras			0.7752	271	9h
10	Ahmmad&xiaopeng&sushi			0.7751	168	2d
11	ShallBuyU			0.7749	42	10h
12	joeoe			0.7749	106	16h
13	Nacir Bouazizi			0.7748	78	4h
14	sokazaki			0.7746	41	13d
15	Bock			0.7746	34	8d
16	成都拖板孩			0.7745	111	6h
17	Jacob			0.7744	105	16h
18	Jō Odagiri			0.7744	31	7d

Figure 5: The (very tight) leaderboard from the Global Wheat Detection challenge (David et al., 2020) on Kaggle (Goldbloom and Hamner, 2010). Even if the scores are close, only the top-3 candidates share the $15,000 cash prize.

This vulnerability introduces the risk of "circular reasoning", particularly when the metric model is trained on similar data or shares architectural components with the model being evaluated, potentially leading to overly optimistic results. Typically, the organizers of the *Auto-Survey Challenge 2023* reported that chatGPT was usually overly optimistic and tended to grade its own work better than actual human work.

Despite these various challenges, using models as metrics can provide nuanced and context-specific insights that are hard to capture with traditional evaluation methods. This approach should be applied with caution and rigorous methodology to ensure the most reliable and informative results.

3 How to make a statistically significant evaluation

We stressed out that the selection of appropriate evaluation metrics is critical. Equally critical is ensuring that the evaluation is statistically significant. For a robust evaluation, a sufficiently large test set is essential, along with the computation of score error bars. The figure 5 shows an example of a tight leaderboard from a past competition. In this particular competition, the third place candidate, qualified for a prize, only has a 0.0001 difference in score with the fourth place candidate. This thin margin between the third and fourth place highlights potential concerns regarding the evaluation methodology. This section presents methods for computing error bars, and addresses questions such as the ideal size of the test set. The goal is to provide methods to minimize the influence of randomness in competitions.

3.1 Error bars

Error bars are the representation of the uncertainty of a measurement, allowing to distinguish score estimations between candidate models. There are three common types of error bars: *standard deviation* (STD), *standard error of the mean* (SEM) and *confidence interval* (CI) (Krzywinski and Altman, 2013).

The **standard deviation** consists in the average distance between each sample and the mean:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

The use of $n-1$ in the denominator when computing the sample standard deviation is a result of what's called Bessel's correction (Bishop, 2006; Murphy, 2012). The main reason for using $n-1$ instead of n is to provide an unbiased estimate of the population variance and standard deviation when computed from a sample.

The **standard error of the mean**, if the n observations are statistically independents, is the standard deviation divided by the square root of the sample size:

$$SEM = \frac{\sigma}{\sqrt{n}}$$

The **confidence interval** is calculated by using the standard deviation to create a range of values likely – to a given probability (commonly 95%) – to contain the true population mean. This technique requires the computation of approximations in practice, and is not commonly used to analyze competitions and benchmarks in machine learning.

Given this context, what are the sources of variability in our area of interest? Specifically, in supervised learning, performance estimation often involves comparing model predictions with a test set's ground truth. Here, **the variability comes from the model in one hand, and the data in the other hand**.

The model can be stochastic during three different processes: the initialization, the learning process and the prediction process. Note that each of these processes is not necessarily stochastic in nature, and many models are completely deterministic. Even models that involve random initialization, such as neural networks, can be made deterministic by fixing a random seed[7]. This raises a question: should we impose to participants to fix a seed in order to reduce the variability of their methods? The main benefit of fixing the random seeds being improving the reproducibility of the methods. In general, having high variation of the results due to initialization can be a sign of low generalization capabilities. However, fixing the seed is controversial as it overlooks variability factors. In previous contests, we executed participants' code multiple times, choosing their poorest performance to motivate variance reduction. Although averaging multiple runs decreases variance, organizers shouldn't do this as it may favor high-variance methods; participants should ensure their methods have low variance.

The data is inherently stochastic, originating from real-world observations that can be considered as samples from unknown distributions. This randomness is compounded by variations in labeling quality, splitting into training and test sets, and other factors. One effective ways to mitigate these sources of variance is through high quantities of data. While it is often stated that the

7. A random seed is a number used to initialize a pseudo-random number generator, which is then used to generates the weights. A fixed random seed means that the number generator will always returns the same values.

quality of a machine learning model is largely determined by the quantity of available data, it is at least safe to say that abundant data enhances the reliability of model evaluations, as will be explored in subsequent sections.

The authors of Bouthillier et al. (2021) shows that most evaluations in deep learning focus on the impact of random weight initialization, which is only a small source of variance, comparable to residual fluctuations from hyperparameter optimization. However, this variance is much lower compared to the variance caused by splitting the data into training and test sets.

Statistical hypothesis testing

Statistical hypothesis tests are used to decide whether the data at hand sufficiently support a particular hypothesis. In our area of interest, hypotheses often involve comparisons, such as whether algorithm A outperforms algorithm B or if the performance of various algorithms aligns with that of the baseline method. We mostly make multiple comparisons of multiple algorithms, multiple comparisons between two algorithms, or comparison between algorithms to a control (the baseline). The tests used for different scenarios are detailed in Japkowicz and Shah (2011). Even if the classical null hypothesis statistical tests (NHSTs) are widely used, recent research advocate for the use of Bayesian analysis instead (Benavoli et al., 2017). The authors present the Bayesian correlated t-test, the Bayesian signed rank test and a Bayesian hierarchical model that can be used for comparing the performance of classifiers, arguing that it solves the drawbacks of the frequentist tests. One of the drawbacks underlined is the fact that NHST computes the probability of getting the observed (or a larger) difference between classifiers if the null hypothesis of equivalence was true, which is not the probability of one classifier being more accurate than another, given the observed empirical results. Another common problem is that the common usage of NHST relies on the wrong assumptions that the p-value is a reasonable proxy for the probability of the null hypothesis (Demăr, 2008). Other areas of science are also moving from NHSTs to Bayseian approaches, as evidenced by the journal *Basic and Applied Social Psychology*, which in 2015 banned the use of NHSTs and related statistical procedures (Trafimow and Marks, 2015).

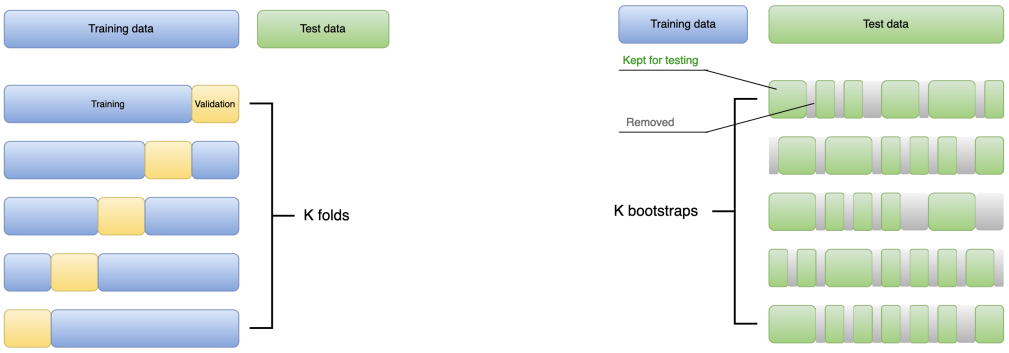
In practice, in competitions, the statistical testing boils down to ranking participants and declaring ties. In the following, we examine the use of bootstrap and cross-validation as estimators of the variance in models performance.

Bootstrap and cross-validation

In the field of machine learning, *cross-validation* and *bootstrapping* are widely used ways of computing the bias and variance of models performances. These two methods are inherently different, as cross-validation involves re-training the model from scratch several times, while bootstrapping only uses re-shuffling of the test samples and predictions, making it quicker to compute. Validation methods help to prevent overfitting, a common issue in machine learning where a model performs well on the training data but poorly on new unseen data.

The **K-fold cross-validation** (CV) (Hastie et al., 2009) involves dividing the original data into several parts (folds), where one part is used for testing and the rest for training. This process is repeated multiple times, each time with a different part used for testing, and the performance metrics are averaged across all iterations to get a final evaluation of the model.

Bootstrapping (Efron and Tibshirani, 1993) involves generating multiple subsets of the original data set, using sampling with replacement, each having the same size as the original set. The



(a) K-fold cross-validation (b) Bootstrapping

Figure 6: Schema of K-fold cross-validation (left) and bootstrapping (right). The cross-validation implies training on subsets of data with size determined by K. Bootstrapping involves sampling with replacement the test data, thus implying duplicated and missing samples in each evaluation. There is no limit to the number of re-sampling (bootstraps) that can be performed.

algorithm is then evaluated on each subset (known as a "bootstrap sample"). The performance metrics are averaged across all bootstrap samples to get a more robust evaluation of the algorithm.

K-fold CV and bootstrapping are illustrated and compared in the Figure 6. In both cases, the variance can be computed on the set of scores obtained. Note that, the bootstraps and the folds not being independents, we can't compute the standard error (dividing by the square root of the number of scores), as mentioned in Section 3.1.

The question of determining the best methods for estimating a model's generalization error received substantial attention in the literature, as evidenced by numerous studies (Nadeau and Bengio, 2003; Bengio and Grandvalet, 2004b; Markatou et al., 2005; Kohavi, 1995a; Zhang and Yang, 2015a; Dietterich, 1998; Tsamardinos et al., 2018; Esbensen and Geladi, 2010; Molinaro et al., 2005; Langford, 2005b,a; Forman and Scholz, 2010). This problem is deep and the suitability of an estimator appears to depend both on the evaluated models and the data they are evaluated on. Some empirical evaluations of generalization error estimators have been conducted (Kohavi, 1995b; Zhang and Yang, 2015b), and advise for a 10-fold CV. Top participants of the Performance Prediction Challenge (Guyon et al., 2006) used various cross-validation techniques to minimize average guess errors. The top performer employed virtual leave-one-out (VLOO) cross-validation for kernel classifiers, optimizing loss function through intensive cross-validation and using fresh data splits for re-estimation. While many preferred standard 10-fold cross-validation for hyperparameter tuning, others experimented with methods like bagging with bootstrap re-sampling or using challenge validation sets for predictions. Bengio and Grandvalet (2004a) demonstrate that, when dealing with simple cases, neglecting the dependencies between test errors can result in a bias that is roughly equal to the variance. These experiments highlight that one must exercise caution when interpreting the significance of differences in cross-validation scores.

Bootstrapping is more practical due to its computational efficiency and can be applied directly to results, eliminating the need to access the underlying algorithms (e.g. when evaluating results

submissions). Therefore, bootstrapping is highly valuable in the context of competitions and benchmarks, as it enables variance calculations without the need for computationally intensive operation.

WITHIN GROUP AND BETWEEN GROUP VARIANCE

It is common to have multiple levels of granularity when computing scores and error bars. The highest granularity level is the level of data points, or samples. Samples can be grouped in lower granularity levels, such as tasks or datasets. Indeed, each task has a unique test set, leading to a distribution of scores for each algorithm on each task. This can also be defined as *within group variance*, the variance between the samples of a dataset, and *between group variance*, the variance between the tasks.

In this situation, the variance can be studied by invoking the law of total variance. The law of total variance (Weiss, 2005), also known as Eve's law, decomposes the variance of a random variable into two parts: the expected value of the variances conditioned on another random variable, and the variance of the expected values conditioned on that same random variable. Formally, let X and Y be two random variables. The law of total variance states:

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}(\mathbb{E}[X|Y])$$

It naturally resonates with the concept of multi-granularity variance in the context of machine learning and algorithm evaluation. The overall variance in the performance of the algorithms (without considering tasks) is represented by $\mathrm{Var}(X)$. The expected variance within each task (given the task) is similar to $\mathbb{E}[\mathrm{Var}(X|Y)]$, where Y denotes the specific task. The variance in the average performance of the algorithms across tasks is represented by $\mathrm{Var}(\mathbb{E}[X|Y])$. It dissects the total variance of algorithm performance into parts: one due to inherent variability within each task, and another due to the variability in the algorithm's relative performance across different tasks. This view allows researchers and practitioners to understand the robustness of an algorithm across tasks and the variability of performance within specific tasks.

To dive more in-depth into this multi-level granularity scenario, we conducted experiments on the models' predictions from the AutoML (Guyon et al., 2019) and AutoDL (Liu et al., 2021) Challenges benchmarks. For each model, we estimated the within group and between group variance and compared the results. Our empirical experiments suggest that the highest granularity level exhibits lower variance of scores. Figure 7 shows the standard deviation of the ranks obtained by each participant of the AutoML and AutoDL challenges. In these challenges, the candidates are evaluated on a set of datasets, with a test set for each dataset. We can therefore compute the standard deviation of the ranks of each candidate, by varying the samples (high granularity) or the datasets (low granularity). This computation was performed using bootstraps, and highlights the difference in deviation depending on the granularity.

3.2 Size of test set

A crucial consideration is determining the test set size that provides a reliable error rate estimation. This should be the number one worry of the organizer: having enough test data to enable a robust judgement of the candidates. In the case of classification, Guyon et al. (1998) suggest as a rule of thumb to use $n = \frac{100}{p}$, where n is the test set size and p is the error rate of the best classifier, as estimated, for instance, by the human error rate. Hence, the better your classifier, the bigger your test set needs to be in order for you to compute precisely the error rates. In the case of imbalanced

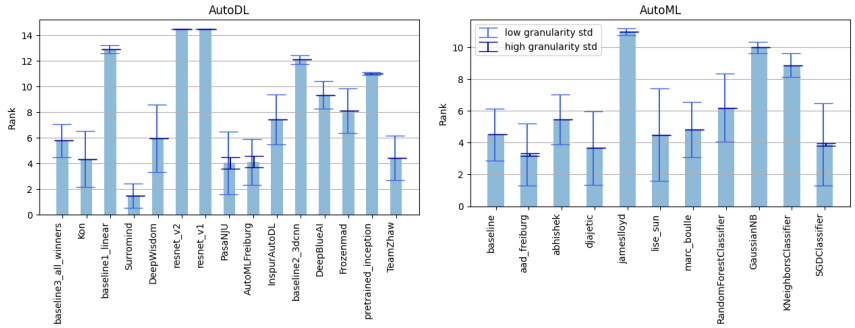


Figure 7: Average and standard deviation of ranks of AutoDL (left) and AutoML (right) candidates. The deviation is computed on bootstraps of data samples (high granularity) and bootstraps of datasets (low granularity).

classes, one can base this analysis on the size of the smallest group, or even regroup the least represented classes. More generally, outside of classification, the absolute precision on the scores or means can be used to separate the participants.

Recently, Guyon introduced a refined formula, which holds for all additive losses. The formula gives the sample size n required to get a given precision ν and confidence k:

$$n = \frac{\sigma^2}{\mu^2} \times k^2 \times 10^{2\nu}$$

Where μ represents the mean error of the model evaluated, and σ represents the standard deviation of the error rates. Interestingly, to increase the precision ν by one decimal, 100 times more test examples are needed. k, μ and σ are squared, also indicating that the number of samples needed grows quickly under the influence of the error rate, the variance and the targeted confidence.

We highlight the importance of the number of test samples in an experiment conducted on tasks from the AutoDL Challenge (Liu et al., 2021). Only the datasets having less than 50,000 test samples were kept, for improved readability of the results. The experiment, done independently on each dataset, consist in increasing gradually the number of test samples used to compute the metrics and rank the candidate models. The size of the test set is therefore increased between 1 and m, m being the total number of test samples of the dataset. For each intermediate value i, we sample with replacement (bootstrap) i samples from the test set, and compute a ranking of the algorithms according to the scores obtained on this test set of i samples. We perform $t = 5000$ trials of this procedure, resulting in t different rankings, on which we compute the ranking stability using the Kendall W concordance measure. The number of candidates n is fixed in this experiment. n should not have an impact on the value of the stability itself, but only the variance of this value.

The results, given in Figure 8, indicate that, unsurprisingly, the stability increase with the number of test samples. A value of stability of 1 means that the ranking of all methods does not change when bootstrapping the test samples. In this experiment, most datasets converge into a stability near 1 when the number of test samples reaches 10^4. This indicates that the proposed models are well separated by the tasks. In the presence of ties, the stability converges to a value below 1. Under 1,000 samples, the rankings are unstable, meaning that there is an insufficient number of samples

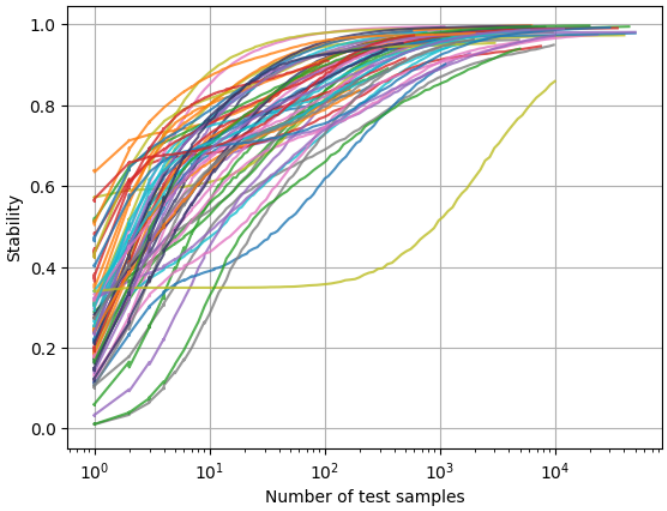


Figure 8: Evolution of the ranking stability depending on the number of test samples used to score the candidates. The stability is the Kendall W measure computed on repeated trials. Each line is an independent dataset, and colors are displayed for readability.

to significantly rank the candidates by performance in the case of this benchmark. For benchmarks where the models are numerous and harder to separate, more than 10^4 may be required in order to obtain a significant final ranking.

3.3 Avoid overfitting: staged evaluation

PUBLIC AND PRIVATE LEADERBOARDS

In order to do a reliable evaluation, you must divide the data in **at least** three sets: train, validation and test. Note that validation and test sets are also commonly named "development and final phases"[8] or "public and private leaderboard". In some competitions, these sets contain distinct tasks or datasets; the validity of the argument still holds in these cases.

The **training set** is fundamental for model building. It includes both the features and the labels (i.e., the ground truth), which enable the model to learn the underlying patterns in the data. The **validation set** serves as an immediate check for how well the model is generalizing to unseen data. Although the ground truth is hidden, participants can get feedback on their performance. This enables them to tweak their models for improvement. It acts as a 'sandbox' for understanding how the model performs on data it hasn't seen before but could potentially overfit to if used improperly. The **test set** is the ultimate arbiter of a model's generalization capability. No feedback on performance is provided, preventing any last-minute tweaking that could artificially inflate the model's evaluation metrics. This is summarized in table 1. Ideally, these sets should share a similar data distribution, unless concept drift or shift is an inherent aspect of the problem being addressed.

8. Other synonyms of development phase include feed-back phase and practice phase.

The test set plays a critical role in this ecosystem by serving as a "firewall" against overfitting, since participants don't get feedback on their test set performance until the competition concludes. Indeed, receiving a repeated feedback from the leaderboard after each submission can lead participants to overfit their models to the validation data. The purpose of the test set is to evaluate performance on entirely new, unseen data, thereby ensuring that the winning solution is general rather than only excelling on the validation data. On a positive note, this empirical study (Roelofs et al., 2019), conducted on 120 Kaggle competitions, suggests that the overfitting between development and final phases (public and private leaderboards) is not common. This could either mean that most participants are adhering to best practices or that the dataset sizes and complexities are sufficient to mitigate the risks of overfitting.

The importance of splitting data into at least three sets — train, validation, and test — cannot be overstated for ensuring both the reliability and generalizability of machine learning models. This is even more crucial in competitive settings, such as machine learning competitions, where the temptation to fine-tune models based on leaderboard performance can potentially lead to overfitting. Some past competitions allowed participants to fine-tune their models during the final phase, which is generally not a good practice. It blurs the lines between validation and testing, compromising the integrity of the evaluation process. A well-structured competition should aim to measure a model's ability to generalize to new, unseen data, and letting participants fine-tune their models based on test set performance undermines this objective.

	Train	**Validation**	**Test**
Can participants access ground truth?	YES	NO	NO
Can participants obtain a score on it?	YES	YES	NO
Can organizers obtain a score on it?	YES	YES	YES

Table 1: Train, validation and test sets. Here, the **validation set** refers to testing data hidden from participants; not to be confused with the validation procedure they can perform on the **train set**. The **test set** is for the final evaluation, avoiding leaderboard overfitting.

Filtering participants to final phase

While we do not declare the winner based on the validation set results in order to avoid "**participants overfitting**", this does not prevent another type of overfitting: "**organizer overfitting**".

The ambition of competitions is generally to recommend algorithms that could perform well on new tasks resembling that of the competition. Thus competitions are a problem in which the *organizers* perform a learning task: from the task(s) of the challenge, they select an algorithm that should perform well on new future tasks. Organizer overfitting occurs when the number of participants is large and rankings are noisy, increasing the chance of poorly selecting a winner. Competition organizers face a sad paradox: *the larger the number of participants, the more "successful" their competition*, but also *the greater the risk to overfit the particular competition setting*.

A heuristic often employed in sports, chess, and other types of competition is to use eliminatory trial runs to filter participants for the final competition phase. It has been highlighted that generalization can be improved by using the first phase of competition as a filter in machine learning competitions (Pavao et al., 2022b). The method simply consists in keeping only the *top-k* participants of the development phase into the final phase. However, determining the optimal k, the

number of top participants that we allow to access the final evaluation to optimize generalization, is hard to determine in practice.

A conservative choice of k that is preferable in practice, is to **eliminate participants who do not outperform the baseline methods provided by the organizers** with the "starting kit" (which may include well performing methods from previous challenges). This should be more acceptable to the participants than setting a hard threshold on the number of entrants of the *final phase*, and will at least eliminate the least serious participants who just submit the "starting kit".

THE DANGERS OF SPLITTING

We have seen that splitting the data or tasks for training, validation and testing is necessary in order to evaluate participants fairly and avoid overfitting. It is common practice to completely shuffle randomly all data samples before splitting. This strategy assumes that data samples are independently and identically distributed across sets, avoiding to bias the evaluation towards one set or another. However, this approach can often yield misleading results, particularly in specialized domains where data naturally clusters into groups.

Consider a dataset composed of n microscopic images of cells collected from m different patients. These images aim to train a model for automated diagnosis, designed to generalize to new, previously unseen patients. If data splitting is executed at the image level, rather than the patient level, there's a high risk of overestimating model performance. In such cases, the model's apparent success in validation may not translate into effective generalization. This misleading effect is sometimes referred to as "voodoo machine learning" (Saeb et al., 2016). To better evaluate model performance in such scenarios, it's crucial to perform data splitting at the patient level, as illustrated in Figure 9, ensuring that all images from the same patient are grouped together in one of the training, validation, or testing sets.

The First Impressions Dataset (Escalante et al., 2018), utilized in various challenges, presents a similar data-splitting dilemma. The dataset consists in different videos of the same individuals, captured at different time intervals. to group all videos of the same individual into a single set—be it training, validation, or testing—to get a more accurate measure of the model's ability to generalize to new, unseen individuals. This approach reduces the risk of overfitting and better prepares the model for real-world applications.

More generally, all applications where the data are stratified in two or more levels must be split in a manner that respects these hierarchical structures to ensure accurate evaluation and robust generalization.

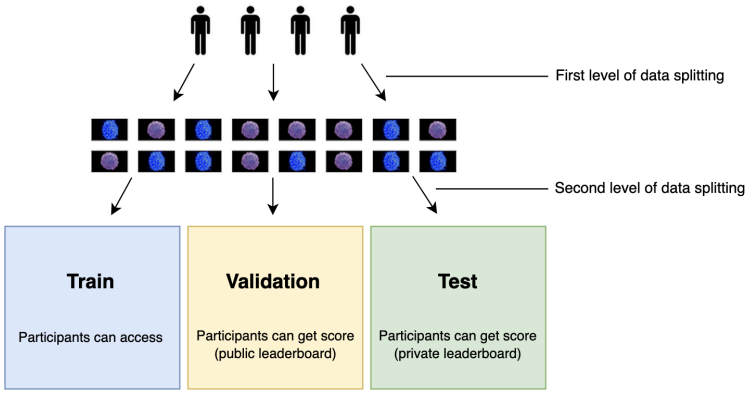


Figure 9: Illustration of several possible levels of data splitting. Here, the samples coming from the same source (the same person or element of the first level of splitting) should be kept in the same subsets of data to avoid overfitting.

4 How to fuse multiple scores

When judging and ranking the participants of a contest, we often need to combine the results from multiple criteria. This multi-score setting can emerge from different scenarios in machine learning competitions: when the models are tested on a set of tasks or datasets, when there are several evaluation trials (e.g. using cross-validation), or when multiple metrics or measures are used to rank the models. The case of having multiple heterogeneous metrics can typically arise when combining primary objective and secondary objective metrics when ranking candidate models. This problem, in a more general form, can be referred as the *problem of ranking* (Kendall and Smith, 1939): the goal is to rank a set of "candidates" $\mathscr{C}$, using the scores attributed to each of them by a set of "judges" $\mathscr{J}$. A judge here is simply any scoring procedure, hence producing a list containing one score for each candidate.

As each judges attributes a score to each candidates, the data of the problem can be represented by a score matrix M, as shown in Table 2. The problem then consists in using a ranking function $f : \mathbb{R}^{n \times m} \to \mathbb{R}^n$ to obtain a single ranking of candidates $\mathbf{r} = \texttt{rank}(f(M))$, with the function $\texttt{rank} : \mathbb{R}^n \to \mathbb{R}_+^n$ is defined as follows: $\forall i \in \{1, ..., n\}$, $\texttt{rank}(\mathbf{v})_i = 1 + \sum_{j \neq i} \mathbb{1}_{\mathbf{v}_j > \mathbf{v}_i} + \frac{1}{2} \sum_{j \neq i} \mathbb{1}_{\mathbf{v}_j = \mathbf{v}_i}$.

	Judge 1	...	Judge m
Candidate 1	$score_{11}$	...	...
...	...	$score_{ij}$	...
Candidate n	...	...	$score_{nm}$

Table 2: Score matrix. The judges can be of various nature (tasks, metrics, etc.). The output scores can also be of various types (real numbers, integers or ranks).

Table 3 shows an example of such problem. Depending on the ranking function f chosen, the final ranking $\mathbf{r}$ will vary. Therefore, what is the good method to use? Intuitively, we want the final

ranking to represent as best as possible the opinions of all judges and to be congruent in this sense. However, this is a ill-defined objective. We can only propose methods that aim at answering this problem, and try to understand the underlying properties of the proposed methods.

	Judge 1	Judge 2	Judge 3	Judge 4
Candidate 1	0.8	0.5	0.7	0.5
Candidate 2	0.6	0.9	0.4	0.5
Candidate 3	0.4	0.7	0.8	0.5

Table 3: An example of a score matrix. A ranking function f takes a score matrix as input and returns a **final ranking r** of the candidates.

We present the most common ranking functions in Section 4.1, their properties in Section 4.2 and give guidelines in Section 4.3.

4.1 Ranking functions

RANDOM DICTATOR

A straightforward approach to deriving a ranking from a matrix of scores involves uniformly selecting a judge at random and then adopting their judgment as the definitive ranking. This method is referred to as the *random ballot* or the *random dictator*. While it may initially seem counter-intuitive or even absurdly incorrect, it's astonishing how prevalent this method is in reality. In essence, the **random dictator is omnipresent**. Whenever we don't tackle a ranking problem head-on, but rather rely on a singular score to rank objects, we effectively permit the outcome to be governed by chance. This isolated score is typically drawn from a "mother distribution" and is consequently chosen at random. Examples of such scenarios include instances without re-runs, those lacking bootstrap resampling, or cases focusing on just one task. In such situations, the influence of the random dictator becomes evident.

MEAN AND MEDIAN

Mean and *Median* are average judges, obtained by either taking the mean (average value) or the median (middle value) over all judges, for each candidate. These two approach are fairly simple and very common in practice, especially the *mean*.

A potential issue with the *mean* is its sensitivity to extreme values, meaning that all judges don't have an equal impact on the outcome, especially if the scores are not normalized on the same scale, or are of different nature. The median leverage a bit this bias. On the other hand, if the scores are of similar nature, i.e. independently sampled from the same distribution, for examples several re-runs of the same experiment, then the *mean* naturally computes and converges to the expected value as the number of judges m increases, as stated by the central limit theorem (Anderson, 2010) and the law of large numbers (Evans and J.S.Rosenthal, 2004).

AVERAGE RANK

Average rank, or *Borda count*, is defined as follows:

$$f(M) = \frac{1}{m} \sum_{\mathbf{j} \in \mathscr{J}} \texttt{rank}(\mathbf{j})$$

It has the interesting property of computing a ranking which minimizes the sum of the Spearman distance with all the input judges, as shown by Kendall and Gibbons (1990).

Pairwise comparisons

Pairwise comparisons methods give scores based on comparisons of all pairs of candidates:

$$f(M) = \left(\frac{1}{(n-1)} \sum_{j \neq i} w(\mathbf{c}_i, \mathbf{c}_j) \right)_{1 \leq i \leq n}$$

where $w(\mathbf{c}_i, \mathbf{c}_j)$ represents the performance of $\mathbf{c}_i$ against $\mathbf{c}_j$. We can define different pairwise methods by designing different w functions:

- *Copeland's method*: $w(\mathbf{u}, \mathbf{v}) = 1$ if the candidate $\mathbf{u}$ is more frequently better than the candidate $\mathbf{v}$ across all judges, 0.5 in case of a tie, and 0 otherwise.
- *Relative Difference*: $w(\mathbf{u}, \mathbf{v}) = \frac{1}{m} \sum_{k=1}^{m} \frac{u_k - v_k}{u_k + v_k}$.

In pairwise comparison methods, when a candidate beats all other candidates, it is a clear winner to be ranked first. If a candidate beats all other according to *Copeland's method*, it is said to be the *Condorcet winner*. However, there is no always a candidate that outplay all its opponents. This is because the majority preferences can be cyclic, thus exhibiting what is called a Condorcet paradox (Gehrlein, 1997). This property can be illustrated using Condorcet graphs, a graphical representation of pairwise comparisons between candidates. An arrow is drawn from one candidate to another when it performs better. Examples of such graphs, with a clear Condorcet winner, and exhibiting a cycle, are given in Figure 10.

Optimal rank aggregation

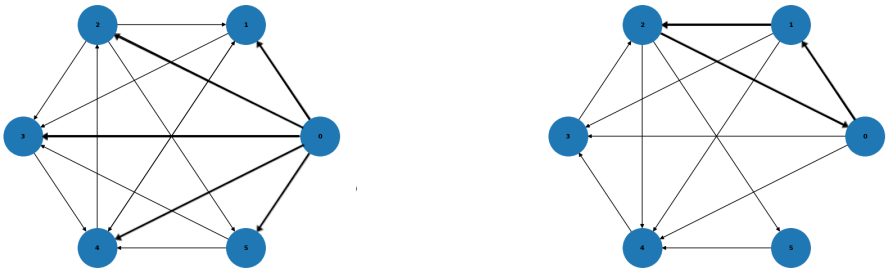
Optimal rank aggregation (ORA) methods are a family of ranking methods that consist in proposing a distance function $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ and finding a ranking r which minimizes the following objective function:

$$l(\mathbf{r}) = \sum_{\mathbf{j} \in \mathscr{J}} d(\mathbf{r}, \mathbf{j})$$

Some well-known distance functions that can be used are Kendall's τ distance, Spearman's distance or the Euclidean distance.

The ORA using Kendall's τ as a distance function is known as the *Kemeny-Young* method. It has interesting properties such as being a Condorcet method and satisfying Local IIA (defined below); however, its computation is NP-Hard. The high complexity of the *Kemeny-Young* method prevented us from including it in the experiments.

The ORA using the Spearman distance also has interesting properties and is computationally linear as it produces the same ranking as the *average rank* method (Kendall and Gibbons, 1990), as mentioned earlier.



(a) *Candidate 0* is the Condorcet winner. (b) There is no Condorcet winner.

Figure 10: Condorcet graphs where vertices represents candidates, and where there is an arrow between candidates **u** and **v** if **u** is more frequently better than **v** according to the judges' preferences. In the left example, *Candidate 0* is a clear Condorcet winner, beating all other candidates. The right example exhibit a Condorcet paradox, as *Candidate 0* beats *Candidate 1*, *Candidate 1* beats *Candidate 2* and *Candidate 2* beats *Candidate 0*, resulting in a cycle. Bold arrows are highlighted for clarity.

In practice, the optimization can be performed using differential evolution (Storn and Price, 1997). A good overview of ORA and rank distance functions is given in Heiser and D'Ambrosio (2013).

4.2 The problem of ranking is not trivial

No ranking function perfectly captures the judges preferences when there are more than two candidates. This idea is well highlighted by an important result from social choice theory: Gibbard's theorem (Allan, 1973), a generalization of Arrow's theorem (J., 1950).

Theorem 1. ***Gibbard's theorem.*** *Any deterministic ranking method holds at least one of the following three (unwanted) properties:*

1. *The process is dictatorial*[9]*,*
2. *The ranking is limited to only two candidates,*
3. *The process is open to "tactical voting": the preferences of a judge may not best defend their interest.*

In practice, this imply incompatibilities between several desired properties of ranking methods. Some of the theoretical properties satisfied or not by the methods defined here are summarized in Table 4. These properties are defined below.

9. In a dictatorial process, a single judge can fully dictate the outcome.

	Winner		Judge perturbation		Candidate perturbation		
	Majority	Condorcet	Consistency	Participation	IIA	LIIA	Clone-proof
Random			✓	✓	✓	✓	✓
Mean			✓	✓	✓	✓	✓
Median					✓	✓	✓
Average rank			✓	✓			
Copeland	✓	✓					
Kemeny-Young	✓	✓				✓	

Table 4: Main properties satisfied or not by the ranking functions.

Majority criterion (Rothe, 2015): If one candidate is ranked first by a majority (more than 50%) of judges, then that candidate must win.

Condorcet criterion: The Condorcet winner is always ranked first if one exists. The Condorcet winner is the candidate that would obtain majority against each of the others when every pair of candidates is compared. The Condorcet criterion is stronger than the Majority criterion.

Consistency: Whenever the set of judges is divided (arbitrarily) into several parts and rankings in those parts garner the same result, then a ranking on the entire judge set also garners that result.

Participation criterion: The removal of a judge from an existing score matrix, where candidate u is strictly preferred to candidate v, should only improve the final position of v relatively to u.

Independence of irrelevant alternatives (IIA): The final ranking between candidates u and v depends only on the individual preferences between u and v (as opposed to depending on the preferences of other candidates as well).

Local IIA (LIIA) (weaker): If candidates in a subset are in consecutive positions in the final ranking, then their relative order must not change if all other candidates get removed.

Independence of clones (clone-proof): Removing or adding clones of candidates must not change the final ranking between all other candidates.

4.3 Guidelines

We have presented different ranking functions, and learned that none of them satisfies all the desired theoretical properties mentioned above. If some of these theoretical properties are absolutely required in your benchmark or competition, then they can dictate the function to chose. However, in most cases, this theoretical analysis shade some light but does not settle the problem once for all. To give insights in practical and machine learning related scenarios, (Brazdil and Soares, 2000) and later (Pavao et al., 2021b) have conducted empirical studies, comparing ranking functions using empirical criteria and running experiments on machine learning benchmarks. The results points that the *average rank* method fares well, in terms of meta-generalization and stability.

Weighted average rank, in which we attribute a weight to each judge before computing the average, can also be a good choice when there is a clear difference in the importance of each judge in the evaluation. The *weighted average rank* function can be expressed as following, with $\mathbf{w}$ the list of weights:

$$f(M) = \frac{1}{m} \sum_{\mathbf{j} \in \mathscr{J}} w_j \times \texttt{rank}(\mathbf{j})$$

5 Conclusion

This chapter offers a categorization of evaluation metrics, covering performance, ethics and societal impact, resource utilization, and evaluator viewpoints (human or model based evaluation). It is clear that real-world constraints must be a part of any evaluation, as they ensure that the results are practical and applicable. To ensure statistically reliable evaluations, it is recommended to have a consequent test set and to use bootstrap methods to assess error bars, given their computational efficiency and accuracy. The required number of test set samples grows quadratically with the mean error, the standard deviation of error rates and the confidence, while it grows exponentially with the targeted precision level.

We also stressed out the importance of having a distinct final phase prevents potential overfitting. Filtering participants accessing the final phase improve the generalization of the winner selection (Pavao et al., 2022a). To enhance the relevance of results, as a rule of thumb, only participants that exceed a predefined baseline should be allowed into this phase.

Finally, we studied the methods for aggregating results from multiple scores. The average rank method showed its efficiency (Pavao et al., 2021a), while remaining simple to compute and to interpret.

References

Ahmed M. Alaa, Boris van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 290–306. PMLR, 2022. URL https://proceedings.mlr.press/v162/alaa22a.html.

Gibbard Allan. Manipulation of voting schemes: A general result. *Econometrica*, 1973.

Carolyn J. Anderson. *Central Limit Theorem*, pages 1–2. John Wiley & Sons, Ltd, 2010. ISBN 9780470479216. doi: https://doi.org/10.1002/9780470479216.corpsy0160. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470479216.corpsy0160.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

Asa Ben-Hur, André Elisseeff, and Isabelle Guyon. A stability based method for discovering structure in clustered data. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, and Teri E. Klein, editors, *Proceedings of the 7th Pacific Symposium on Biocomputing, PSB 2002, Lihue, Hawaii, USA, January 3-7, 2002*, pages 6–17, 2002. URL http://psb.stanford.edu/psb-online/proceedings/psb02/benhur.pdf.

Alessio Benavoli, Giorgio Corani, Janez Demsar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *J. Mach. Learn. Res.*, 18:77:1–77:36, 2017. URL http://jmlr.org/papers/v18/16-305.html.

Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *J. Mach. Learn. Res.*, 5:1089–1105, 2004a. URL http://jmlr.org/papers/volume5/grandvalet04a/grandvalet04a.pdf.

Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *J. Mach. Learn. Res.*, 5:1089–1105, 2004b. URL http://www.jmlr.org/papers/volume5/grandvalet04a/grandvalet04a.pdf.

Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Robustness may be at odds with fairness: An empirical study on class-wise accuracy. *CoRR*, abs/2010.13365, 2020. URL https://arxiv.org/abs/2010.13365.

Daniel Berrar. Performance measures for binary classification. In Shoba Ranganathan, Michael Gribskov, Kenta Nakai, and Christian Schönbach, editors, *Encyclopedia of Bioinformatics and Computational Biology - Volume 1*, pages 546–560. Elsevier, 2019. doi: 10.1016/b978-0-12-809633-8.20351-8. URL https://doi.org/10.1016/b978-0-12-809633-8.20351-8.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, 2006. Ch. 1.2.4 The Gaussian distribution.

Ludovico Boratto, Gianni Fenu, and Mirko Marras. Interplay between upsampling and regularization for provider fairness in recommender systems. *User Model. User Adapt. Interact.*, 31(3): 421–455, 2021. doi: 10.1007/s11257-021-09294-8. URL https://doi.org/10.1007/s11257-021-09294-8.

Alexei Botchkarev. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *CoRR*, abs/1809.03006, 2018. URL http://arxiv.org/abs/1809.03006.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In Alex Smola, Alex Dimakis, and Ion Stoica, editors, *Proceedings of Machine Learning and Systems 2021, MLSys 2021, virtual, April 5-9, 2021*. mlsys.org, 2021. URL https://proceedings.mlsys.org/paper/2021/hash/cfecdb276f634854f3ef915e2e980c31-Abstract.html.

Pavel Brazdil and Carlos Soares. A comparison of ranking methods for classification algorithm selection. In *ECML 2000*, Lecture Notes in Computer Science, pages 63–74, 2000. doi: 10.1007/3-540-45164-1_8. URL https://doi.org/10.1007/3-540-45164-1_8.

T. Caliński and J Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. doi: 10.1080/03610927408827101. URL https://www.tandfonline.com/doi/abs/10.1080/03610927408827101.

Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, and Nazife Baykal. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In

2017 International Conference on Computer Science and Engineering (UBMK), pages 821–826, 2017. doi: 10.1109/UBMK.2017.8093539.

Rich Caruana and Alexandru Niculescu-Mizil. Data mining in metric space: an empirical analysis of supervised learning performance criteria. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 69–78. ACM, 2004. doi: 10.1145/1014052.1014063. URL https://doi.org/10.1145/1014052.1014063.

Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *MDPI Electronics*, 2019. URL https://www.mdpi.com/2079-9292/8/8/832/pdf.

Vítor Cerqueira, Luís Torgo, and Igor Mozetic. Evaluating time series forecasting models: an empirical study on performance estimation methods. *Mach. Learn.*, 109(11):1997–2028, 2020. doi: 10.1007/s10994-020-05910-7. URL https://doi.org/10.1007/s10994-020-05910-7.

Irene Y. Chen, Fredrik D. Johansson, and David A. Sontag. Why is my classifier discriminatory? In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 3543–3554, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/1f1baa5b8edac74eb4eaa329f14a0361-Abstract.html.

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *CoRR*, abs/1810.08810, 2018. URL http://arxiv.org/abs/1810.08810.

Etienne David, Simon Madec, Pouria Sadeghi-Tehran, Helge Aasen, B. Zheng, Shouyang Liu, Norbert Kirchgeßner, Goro Ishikawa, Koichi Nagasawa, Minhajul A. Badhon, Curtis Pozniak, Benoit de Solan, Andreas Hund, Scott C. Chapman, Frédéric Baret, Ian Stavness, and Wei Guo. Global wheat head detection (GWHD) dataset: a large and diverse dataset of high resolution RGB labelled images to develop and benchmark wheat head detection methods. *CoRR*, abs/2005.02162, 2020. URL https://arxiv.org/abs/2005.02162.

D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979. ISSN 0162-8828. doi: 10.1109/TPAMI.1979.4766909.

Joost de Winter. Can chatgpt pass high school exams on english language comprehension? *Preprint*, 2023.

Janez Demăr. On the Appropriateness of Statistical Tests in Machine Learning . 2008. URL http://www.site.uottawa.ca/ICML08WS/papers/J_Demsar.pdf.

Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. 1998. URL https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1005.1391&rep=rep1&type=pdf.

Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. Springer, 1993. ISBN 978-1-4899-4541-9. doi: 10.1007/978-1-4899-4541-9. URL https://doi.org/10.1007/978-1-4899-4541-9.

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 2021. URL https://hal.science/hal-03688638.

Kim H. Esbensen and Paul Geladi. Principles of proper validation: use and abuse of re-sampling for validation. *Journal of CHEMOMETRICS*, 2010. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.1310.

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, 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. Explaining first impressions: Modeling, recognizing, and explaining apparent personality from videos. *CoRR*, abs/1802.00745, 2018. URL http://arxiv.org/abs/1802.00745.

M.J. Evans and J.S.Rosenthal. *Probability and Statistics - The Science of Uncertainty*. W.H.Freeman and Company, New York, 2004.

George Forman and Martin Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *SIGKDD Explor.*, 12(1):49–57, 2010. doi: 10.1145/1882471.1882479. URL https://www.kdd.org/exploration_files/v12-1-p49-forman-sigkdd.pdf.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. Mathematical capabilities of chatgpt. *CoRR*, abs/2301.13867, 2023. doi: 10.48550/arXiv.2301.13867. URL https://doi.org/10.48550/arXiv.2301.13867.

William V. Gehrlein. Condorcet's paradox and the condorcet efficiency of voting rules. 1997.

Anthony Goldbloom and Ben Hamner. Kaggle. 2010. URL https://www.kaggle.com/docs/competitions.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *CoRR*, abs/2008.05756, 2020. URL https://arxiv.org/abs/2008.05756.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017. URL http://proceedings.mlr.press/v70/guo17a.html.

I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh. Feature extraction, foundations and applications.

Isabelle Guyon, John Makhoul, Richard M. Schwartz, and Vladimir Vapnik. What size test set gives good error rate estimates? *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(1):52–64, 1998. doi: 10.1109/34.655649. URL https://doi.org/10.1109/34.655649.

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. Performance prediction challenge. 2006.

Isabelle Guyon, Lisheng Sun-Hosoya, Marc Boullé, Hugo Jair Escalante, Sergio Escalera, Zhengying Liu, Damir Jajetic, Bisakha Ray, Mehreen Saeed, Michèle 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.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. Springer, 2009. ISBN 9780387848570. doi: 10.1007/978-0-387-84858-7. URL https://doi.org/10.1007/978-0-387-84858-7.

Willem J. Heiser and Antonio D'Ambrosio. Clustering and prediction of rankings within a kemeny distance framework. In Berthold Lausen, Dirk Van den Poel, and Alfred Ultsch, editors, *Algorithms from and for Nature and Life - Classification and Data Analysis*, pages 19–31. Springer, 2013. doi: 10.1007/978-3-319-00035-0_2. URL https://doi.org/10.1007/978-3-319-00035-0_2.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In Sheila A. McIlraith and Kilian Q. Weinberger,

editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3207–3214. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16669.

José Hernández-Orallo, Peter A. Flach, and César Ferri. A unified view of performance metrics: translating threshold choice into expected classification loss. *J. Mach. Learn. Res.*, 13: 2813–2869, 2012. doi: 10.5555/2503308.2503332. URL https://dl.acm.org/doi/10.5555/2503308.2503332.

Bertrand Iooss, Vincent Chabridon, and Vincent Thouvenot. Variance-based importance measures for machine learning model interpretability. In *Actes du Congrès*, Oct 2022. URL https://hal.archives-ouvertes.fr/hal-03741384.

Arrow Kenneth J. A difficulty in the concept of social welfare. *Journal of Political Economy*, 1950.

Nathalie Japkowicz and Mohak Shah, editors. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011. ISBN 9780521196000.

Scott M. Jordan, Yash Chandak, Daniel Cohen, Mengxue Zhang, and Philip S. Thomas. Evaluating the performance of reinforcement learning algorithms. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4962–4973. PMLR, 2020. URL http://proceedings.mlr.press/v119/jordan20a.html.

Wendy Kan, Phil Culliton, and Douglas Sterling. Dog image generation competition on kaggle. *Competitions in Machine Learning (CiML) workshop at Neural Information Processing Systems (NIPS) 2019*, 2019.

M. Kendall and J.D. Gibbons. *Rank Correlation Methods.* 5th Edition, Edward Arnold, London., 1990.

M. G. Kendall and B. Babington Smith. The problem of m rankings. *The Annals of Mathematical Statistics*, 10(3):275–287, 1939. ISSN 00034851. URL http://www.jstor.org/stable/2235668.

Thanh Gia Hieu Khuong and Benedictus Kent Rachmat. Auto-survey challenge: Advancing the frontiers of automated literature review. In *Junior Conference on Data Science and Engineering 2023*, Orsay, France, Sep 2023. URL https://inria.hal.science/hal-04206578. ⟨hal-04206578⟩.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *CoRR*, abs/1906.02691, 2019. URL http://arxiv.org/abs/1906.02691.

Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 Volumes*, pages 1137–1145. Morgan Kaufmann, 1995a. URL http://web.cs.iastate.edu/~jtian/cs573/Papers/Kohavi-IJCAI-95.pdf.

Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 Volumes*, pages 1137–1145. Morgan Kaufmann, 1995b. URL http://ijcai.org/Proceedings/95-2/Papers/016.pdf.

Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artif. Intell.*, 97(1-2):273–324, 1997. doi: 10.1016/S0004-3702(97)00043-X. URL https://doi.org/10.1016/S0004-3702(97)00043-X.

Martin Krzywinski and Naomi Altman. Error bars. *Nature Methods*, 10(10):921–922, Oct 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2659. URL https://doi.org/10.1038/nmeth.2659.

S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22: 79–86, 1951.

Charline Le Lan, Marc G. Bellemare, and Pablo Samuel Castro. Metrics and continuity in reinforcement learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 8261–8269. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17005.

John Langford. The cross validation problem. In Peter Auer and Ron Meir, editors, *Learning Theory, 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005, Proceedings*, volume 3559 of *Lecture Notes in Computer Science*, pages 687–688. Springer, 2005a. doi: 10.1007/11503415_47. URL https://doi.org/10.1007/11503415_47.

John Langford. Tutorial on practical prediction theory for classification. *J. Mach. Learn. Res.*, 6: 273–306, 2005b. URL https://www.jmlr.org/papers/volume6/langford05a/langford05a.pdf.

Yann LeCun and Corinna Cortes. The mnist database of handwritten digits. 2005.

Xingxuan Li, Yutong Li, Linlin Liu, Lidong Bing, and Shafiq R. Joty. Is GPT-3 a psychopath? evaluating large language models from a psychological perspective. *CoRR*, abs/2212.10529, 2022. doi: 10.48550/arXiv.2212.10529. URL https://doi.org/10.48550/arXiv.2212.10529.

Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55, 1932.

Cheng-Yuan Liou, Wei-Chen Cheng, Jiun-Wei Liou, and Daw-Ran Liou. Autoencoder for words. *Neurocomputing*, 139:84–96, 2014.

Zhengying Liu, Adrien Pavao, Zhen Xu, Sergio Escalera, Fabio Ferreira, Isabelle Guyon, Sirui Hong, Frank Hutter, Rongrong Ji, Júlio C. S. Jacques Júnior, Ge Li, Marius Lindauer, Zhipeng Luo, Meysam Madadi, Thomas Nierhoff, Kangning Niu, Chunguang Pan, Danny Stoll, Sébastien Treguer, Jin Wang, Peng Wang, Chenglin Wu, Youcheng Xiong, Arber Zela, and Yang Zhang.

Winning solutions and post-challenge analyses of the chalearn autodl challenge 2019. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(9):3108–3125, 2021. doi: 10.1109/TPAMI.2021.3075372. URL https://doi.org/10.1109/TPAMI.2021.3075372.

Marianthi Markatou, Hong Tian, Shameek Biswas, and George Hripcsak. Analysis of variance of cross-validation estimators of the generalization error. *J. Mach. Learn. Res.*, 6: 1127–1168, 2005. URL http://www.jmlr.org/papers/volume6/markatou05a/markatou05a.pdf.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, June 1947. doi: 10.1007/bf02295996. URL https://doi.org/10.1007/bf02295996.

Annette M. Molinaro, Richard Simon, and Ruth M. Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinform.*, 21(15):3301–3307, 2005. doi: 10.1093/bioinformatics/bti499. URL https://pubmed.ncbi.nlm.nih.gov/15905277/.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA, 2012. Ch. 6.4.2 Unbiased estimators.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Mach. Learn.*, 52(3):239–281, 2003. doi: 10.1023/A:1024068626366. URL https://papers.nips.cc/paper/1661-inference-for-the-generalization-error.pdf.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2901–2907. AAAI Press, 2015. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9667.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/arXiv.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

Julio-Omar Palacio-Niño and Fernando Berzal. Evaluation metrics for unsupervised learning algorithms. *CoRR*, abs/1905.05667, 2019. URL http://arxiv.org/abs/1905.05667.

Adrien Pavao, Michael Vaccaro, and Isabelle Guyon. Judging competitions and benchmarks: a candidate election approach. *European Symposium on Artificial Neural Networks (ESANN) Proceedings*, 2021a.

Adrien Pavao, Michael Vaccaro, and Isabelle Guyon. Judging competitions and benchmarks: a candidate election approach. In *29th European Symposium on Artificial Neural Networks*, 2021b.

Adrien Pavao, Isabelle Guyon, and Zhengying Liu. Filtering participants improves generalization in competitions and benchmarks. In *(ESANN) 2022 - European Symposium on Artificial Neural Networks*, Bruges, Belgium, 2022a. URL https://www.esann.org/sites/default/files/proceedings/2022/ES2022-72.pdf.

Adrien Pavao, Zhengying Liu, and Isabelle Guyon. Filtering participants improves generalization in competitions and benchmarks. In *30th European Symposium on Artificial Neural Networks*, 2022b.

Joseph Pedersen, Rafael Muñoz-Gómez, Jiangnan Huang, Haozhe Sun, Wei-Wei Tu, and Isabelle Guyon. Ltu attacker for membership inference, 2022.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *CoRR*, abs/1811.12808, 2018. URL http://arxiv.org/abs/1811.12808.

Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9175–9185, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/ee39e503b6bedf0c98c388b7e8589aca-Abstract.html.

Jörg Rothe. *Economics and Computation: An Introduction to Algorithmic Game Theory, Computational Social Choice, and Fair Division.* Springer, 2015. ISBN 9783662479049.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: 10.1016/0377-0427(87)90125-7.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x. URL https://doi.org/10.1038/s42256-019-0048-x.

S Saeb, L Lonini, A Jayaraman, DC Mohr, and KP Kording. Voodoo machine learning for clinical predictions. biorxiv, 059774, 2016.

Rainer Storn and Kenneth V. Price. Differential evolution - A simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.*, 11(4):341–359, 1997. doi: 10.1023/A:1008202821328. URL https://doi.org/10.1023/A:1008202821328.

David Trafimow and Michael Marks. Editorial. basic and applied social psychology. 2015.

Ioannis Tsamardinos and Constantin F. Aliferis. Towards principled feature selection: Relevancy, filters and wrappers. In Christopher M. Bishop and Brendan J. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, AISTATS 2003, Key West, Florida, USA, January 3-6, 2003*. Society for Artificial Intelligence and Statistics, 2003. URL http://research.microsoft.com/en-us/um/cambridge/events/aistats2003/proceedings/133.pdf.

Ioannis Tsamardinos, Elissavet Greasidou, and Giorgos Borboudakis. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Mach. Learn.*, 107(12):1895–1922, 2018. doi: 10.1007/s10994-018-5714-4. URL https://arxiv.org/pdf/1708.07180.pdf.

Kadir Uludag and Jiao Tong. Testing creativity of chatgpt in psychology: interview with chatgpt. *Preprint*, 2023.

Mariya I. Vasileva. The dark side of machine learning algorithms: How and why they can leverage bias, and what can be done to pursue algorithmic fairness. In Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash, editors, *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3586–3587. ACM, 2020. doi: 10.1145/3394486.3411068. URL https://doi.org/10.1145/3394486.3411068.

Cédric Villani. *The Wasserstein distances*, pages 93–111. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-540-71050-9. doi: 10.1007/978-3-540-71050-9_6. URL https://doi.org/10.1007/978-3-540-71050-9_6.

Ulrike von Luxburg, Robert C. Williamson, and Isabelle Guyon. Clustering: Science or art? In Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham W. Taylor, and Daniel L. Silver, editors, *Unsupervised and Transfer Learning - Workshop held at ICML 2011, Bellevue, Washington, USA, July 2, 2011*, volume 27 of *JMLR Proceedings*, pages 65–80. JMLR.org, 2012. URL http://proceedings.mlr.press/v27/luxburg12a.html.

Neil A. Weiss. *A Course in Probability*. Addison–Wesley, 2005.

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. Privacy preserving synthetic health data. In *27th European Symposium on Artificial Neural Networks, ESANN 2019, Bruges, Belgium, April 24-26, 2019*, 2019. URL http://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2019-29.pdf.

Yongli Zhang and Yuhong Yang. Cross-validation for selecting a model selection procedure. 2015a. URL http://users.stat.umn.edu/~yangx374/papers/ACV_v30.pdf.

Yongli Zhang and Yuhong Yang. Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, 187(1):95–112, 2015b. URL https://EconPapers.repec.org/RePEc:eee:econom:v:187:y:2015:i:1:p:95-112.

Towards impactful challenges: post-challenge paper, benchmarks and other dissemination actions

Antoine Marot ANTOINE.MAROT@RTE-FRANCE.COM
RTE AI Lab, Paris, France

David Rousseau ROUSSEAU@IJCLAB.IN2P3.FR
Université Paris-Saclay, CNRS/IN2P3, IJCLab, 91405 Orsay, France

Zhen (Zach) Xu ZACH.XU@UCHICAGO.EDU
University of Chicago, USA

Reviewed on OpenReview: *https://openreview.net/forum?id=AAqyNe12di*

Abstract

The conclusion of an AI challenge is not the end of its lifecycle; ensuring a long-lasting impact requires meticulous post-challenge activities. The long-lasting impact also needs to be organised. This chapter covers the various activities after the challenge is formally finished. This work identifies target audiences for post-challenge initiatives and outlines methods for collecting and organizing challenge outputs. The multiple outputs of the challenge are listed, along with the means to collect them. The central part of the chapter is a template for a typical post-challenge paper, including possible graphs and advice on how to turn the challenge into a long-lasting benchmark.

Keywords: post-challenge, analysis, paper, dissemination

1 Introduction

If winners are announced at the end of a competition, and everyone simply returns to their usual activity, even the best-conceived competition would have been of limited use. Indeed, many AI/ML challenges shine briefly and are then forgotten, leading to missed opportunities for long-term impact. This chapter discusses the various post-challenge activities necessary to ensure lasting effects from such events, some of which should be prepared even before the challenge starts.

AI/ML challenges have become increasingly popular over the last decade, engaging not only researchers but also enthusiasts and industry practitioners. According to the public data from platforms such as MLContests,[1] which gathers information of challenges across platforms, thousands of competitions solving real-world or academia problems are hosted each year. Conferences like NeurIPS, and KDD have also introduced official competition tracks in recent years[2][3], aiming to foster collaboration between experts and provide tangible benchmarks for new methodologies. Blog posts from conference organizers have explained their rationale for hosting more challenges, typically highlighting their value in addressing open problems, engaging communities, and producing state-of-the-art solutions[4].

1. https://mlcontests.com/
2. https://neurips.cc/Conferences/2024/CompetitionTrack
3. https://kdd.org/kdd-cup
4. https://blog.neurips.cc/2024/06/04/neurips-2024-competitions-announced/

Four types of stakeholders typically play specific roles in a challenge: the *organisers*, the *application domain experts*, the *AI/Machine Learning experts* on techniques relevant to the problem posed, and the *participants*. Participants themselves come from diverse backgrounds—ranging from domain-specific experts and AI specialists to students and experienced data scientists—each contributing unique perspectives that enrich the challenge outcomes. Input from all these people should be gathered and made available for posterity so that the community can **build upon results and lessons**, **identify remaining gaps/research directions**, and **access new materials** (benchmarks, codes, tutorials) to further the state of the art.

1.1 Why Post-Challenge Activities Matter

Ensuring long-term community engagement requires sharing outcomes and resources with both participants and the wider research community:

- Results and lessons on which they can build upon,
- Remaining gaps and research directions that can push the boundaries of current knowledge,
- Materials (benchmarks, solution codes, visualizations, tutorials) that facilitate continued and reproducible research.

Post-challenge activity is therefore necessary, especially as the raw and meta outputs of challenges can be numerous and complex. Assigning sufficient resources for structuring and evaluating these outputs helps extract meaningful analysis, discussion, and conclusions.

A **post-challenge paper** represents one cornerstone of such activities. It conveys results, lessons, gaps, and research directions in a concise, intelligible form. Such a paper often keeps the community engaged by disseminating challenge outcomes, contributing new insights, and incentivizing future work to push the state of the art. Ideally, this post-challenge paper is complemented by:

- A white paper for a broader audience, explaining the challenge background,
- A challenge design paper describing the modelling and problem implementation,
- Code or dataset documentation for reproducibility and further experimentation.

1.2 Types of Post-Challenge Papers

In practice, many different types of post-challenge papers are possible, each providing distinctive benefits:

- A short analysis paper by organisers only (based on a fact sheet) comparing the best solutions' performance and reflecting on challenge design.
- A federated paper that includes top participants, with a stronger focus on best solution descriptions.
- An introduction to a book or special issue that collates multiple participant papers.
- An introduction to the proceedings of a workshop.

- A journal paper, possibly following several iterations of a competition series, providing a more in-depth analysis of how the challenge altered the scientific landscape and emphasizing the applicable current state of the art.

The exact nature of a post-challenge paper depends on the competition's objectives: addressing a fundamental AI/ML question, or tackling an applied problem, or introducing a novel formulation of an existing issue, or focusing on feasibility, benchmarking, and pushing the current state of the art.

When planning such a paper, organizers should consider the diversity of possible readers and carefully position the content. Different audiences may include:

- AI/ML researchers looking to apply their approaches to various problems (federated paper or introduction to workshop proceedings),
- AI/ML researchers wanting to learn about state-of-the-art methods (federated paper, introduction to special issue/book, or journal paper),
- Domain expert researchers interested in outcomes relevant to their field and comparisons with non-AI methods (short or journal paper),
- Domain expert scientists seeking models and problem framings that best attract the AI community (short paper),
- Challenge organizers searching for best practices and innovations in setting up and running competitions (short or journal paper),
- Scientists investigating evaluation frameworks (short or journal paper),
- Research managers looking for fruitful collaboration with skilled teams (federated paper),
- Vendors/investors seeking promising approaches in next-generation applications (short or journal paper),
- Science popularizers/reporters (short or journal paper).

Other additional aspects might also be of interest, such as considerations of ethics, diversity, AI for good, AI democratisation, framework development, and resource requirements.

The chapter guides readers through critical aspects of post-challenge planning: organizing raw outputs (Sec. 2), facilitating workshops (Sec. 3), drafting post-challenge papers (Sec 4), and establishing enduring benchmarks (Sec 5), concluding with actionable recommendations in (Sec. 6).

2 Challenge raw output

Having the above mentioned high-level considerations in mind and before describing typical post-challenge papers more concretely, we will now review the different kinds of challenge outputs that could be made available and of good use. The raw output of the challenge is all the material produced during the challenge, which must be analysed. The various types of outputs are listed in this section.

2.1 Competition platform output

A typical challenge platform can provide to the organisers, for each participant (or team) and all their submissions, many pieces of information: the time of submission, public and private scores and other similar quantities and most likely the detailed content of the submission, including the actual code in case of code submissions. The final (or selected) submissions are the most interesting, but intermediate ones can also inform the improvement process. Also, the time evolution of the public and private leaderboards is part of the challenge narrative.

2.2 Participants fact sheet

At the end of the competition, it is good practice for organisers to submit a form, the "fact sheet", to the participants to (i) have details on their best submission, which is not automatically provided by the platform, and (ii) know more about the participant themselves, who they are and how they managed their participation. It is best to advertise the form already in the last few days of the competition before participants move on to other activities and to insist that all inputs are worthwhile, even from non-top performers. The form should have a good balance between closed form Multiple Choice Questions, more straightforward to analyse, and free fields to gather specific feedback. The form should be anonymous by default, but it should also include the possibility of indicating the platform user name or leaving an email to continue the dialogue. The goal is to provide an overview of all participants (at least the ones who answer), while the bulk of the post-challenge activity will (and rightly so) focus on the best or most original contributions.

One might want to know more about:

- techniques and tools used
- resource used, in particular, training resources
- estimate time spent on the competition
- participant's background and initial knowledge about the competition. How diverse were the participants? From which age range? From which regions of the world? From academia or industry? Is there initial expertise on the problem or the domain? Which family of methods do they come from? Were they here to win, learn or find a dataset to use?
- How did the participants initially learn about the competition? This helps understand to whom the competition was eventually best addressed and disseminated and if that matches what was expected.
- Which available material and resources (papers, documentation, tutorials, baselines, tools, compute credits) did participants know of, and how useful and easy to use were they? This could help understand if the main competition features were understood or if any were missed, if that helps participants stay active and engaged, and if everything was eventually there to help them learn and participate. This would help organisers improve those resources for future benchmark or competition iterations.
- How interesting and challenging did participants find the competition and its format, with any pros and cons? What does the entry cost and learning curve look like for most participants? How resource-intensive was the competition to reach competitive performance? How difficult

the competition eventually was? These can explain participant activity during competition. This should bring some lessons on the competition formulation and calibration. You can eventually ask them how satisfied they were with the competition, if they'd want to stay engaged and how.

- Other questions can consider organisational aspects during the competition, such as ease of competition platform and submission protocol, competition length and phases, communication clarity, forum activity, prize distribution and incentives for participant investment, and opportunities for collaboration among participants.

This form and questions will help better assess which aspects the competition was most successful at and could be improved later.

2.3 Code

In a code-less platform, participants train their algorithm on a training dataset, apply it (the inference part) to a test dataset and submit the results as a data file. In this case, asking participants to submit their software, including training and inference, is customary on a platform like GitHub. This, however, can only be made mandatory for participants who can claim a price, which should have been specified in the competition rules; the others would often not bother. Also, if the inference code can be tested relatively quickly, it is much less the case for the training code, as the final model is often the result of many iterations. The code should be runnable, well documented, and accompanied by a short document describing its functionality.

For a platform accepting code submissions, the code (guaranteed to be the one producing the ranked results) is already available. There still needs to be a submission of commented human-readable code accompanied by a short document. It is the same if the code submission is only the inference part; the training part must be submitted separately.

There is a difference between publicising the code and releasing it to the organisers. The former is best for dissemination, but people in some communities might prefer to avoid it. It is also important to recommend that the code has an open-source license so that others using it don't risk copyright infringement.

2.4 Competition log

When the challenge is running, a competition logbook should be updated with the main events so that the narrative of the challenge can be told afterwards. Possible salient events: significant changes in the leaderboard, popular posts in the forum, for example, advertising a technique that many participants adopt, updates in challenge documentation, possible challenge reset after a flaw was uncovered and fixed, reports or the discovery of cheating, social media visibility, media coverage (and impact on participation), etc.

2.5 Unorganized raw output

A flow of information from the competition not formatted by the platform or online forms needs to be analysed. They can also represent a measure of how active the competition is. The competition forum is the primary source of such information: data exploration posts and notebooks, insights, code sharing, and documentation gleaned on the web on the topic. One can, for example, see

an idea appear in a forum post, followed by a code implementation (not necessarily by the same person), followed by a general increase in the scores on the leaderboard. However, relevant technical discussions may happen outside the competition forum, such as on social networks, blog posts, or arXiv papers. Also, competitions are often used as practical projects for (e.g., data science) courses. Compared to typical participants, students are usually compelled to write a hopefully clear document about their techniques, which might be public. Such spontaneous output should be harvested regularly, which is much easier if the competition has a unique acronym to be googled and for which alerts can be set up.

3 Post-challenge workshops and discussion

One or more post-challenge workshops are often organised as part of a conference or in a dedicated venue[5][6][7]. Participating in it is part of the incentive for participants. It allows participants to switch from competition to cooperation. They can discuss their techniques between them and with the application or Machine Learning experts. The connections created during such workshops encourage participants to remain engaged with the competition topic instead of just moving on to a different one. The discussions during the workshops and continued online are crucial to deriving the scientific lessons from the workshops and the avenues for further study and possible future challenges.

4 Post challenge paper template

This section details a complete template of a typical post-challenge paper. It should only be considered as guidelines. The emphasis on the different sections would vary widely depending on the type of challenge. Also, a decision to be made early on is how to include the authors of the most interesting contributions: should they write a sub-section about their methods (as done in this template) and sign the paper, or should they write their own paper citing the post-challenge paper? Examples of graphs relevant to a post-challenge paper illustrate the template.

4.1 Introduction

As for every paper, a proper introduction should first recall the context of the challenge and highlight the problem at hand. It should further explain why it is essential to solve it, how impactful it can be, and what the bottlenecks to addressing it have been so far. Eventually, objectives and expectations for such a challenge could be shared.

4.2 Challenge Description

This section should ideally come as a reminder and synthesis of a prior paper on the challenge design. One should first review the problem of the challenge to give a good enough understanding to the targeted audience for the remainder of the paper, particularly highlighting its main dimensions of complexity and variations. One should further present the specific task formulated for the challenge with some description of the underlying datasets, framework (such as a chatbot or reinforcement learning framework if relevant) and problem modelisation (possibly relying on some simulator).

5. https://autodl.chalearn.org/neurips2019
6. https://llm-efficiency-challenge.github.io/schedule
7. https://fair-universe.lbl.gov/

One can remind if the challenge aims to deliver a new problem or formulation to the Machine Learning community or advance and benchmark state-of-the-art. A section on related works and similar challenges can be written to best position this challenge in the scientific landscape and possibly build upon previous work.

The scoring metric and protocol should be discussed, and some considerations should be shared as to why they were chosen a priori to evaluate any advances towards problem-solving best. A brief description of the challenge platform , as well as possible specific choices (like resource allocation) or developments that can be justified there.

Finally, one can describe the challenge organisation and materials available to the participants, highlighting any innovations to increase participant engagement during the competition.

4.3 Challenge Narrative

Once the competition is over, it is interesting to understand retrospectively what happened during the competition, leading to the final leaderboard and results. Was the competition tight or not? How long did it take participants to reach a good enough performance? How many teams showed sustained activity, and how many eventually performed well? Did any innovation in solutions occur within the competition, breaking some performance ceiling? Or were most of the solutions derived from an adopted model published in the baselines or by a participant? These can be extracted from raw competition output, highlighting the competition dynamics, attractivity and difficulty.

For example, Fig. 1 shows the progress of participants in one competition as a function of time; one can see the bulk of participants progressing as a swarm, following community understanding of the problems, while a few isolated outliers obtained the best scores.

Fig. 2 shows participants' progress in a different competition in the (accuracy, speed) plan. Various strategies are evident: some tried to optimise both simultaneously, why some others (the best), `fastrack` and `gorbuno` have first reached the best accuracy, then improved their speed without accuracy loss.

A graph like Fig. 3 can help summarise the main events and competition activity.

4.4 Post challenge checks

In most cases, the performance of submissions can be evaluated against a held-out (or private) dataset, providing the final ranking. Generally, this is automatically done on the challenge platform, but it may require some manual actions, mainly when doing it automatically requires too many resources. The stability of the ranking can then be evaluated as in Fig. 4. The graph shows that the rankings among teams are stable, except for `4th Rek`, which performs significantly worse in the second phase.

A more detailed analysis is also possible, as in Fig. 5. The comparison of the private curves shows that a numerical analysis confirms that `Gabor` (the winner) is clearly above the others. The comparison of private and public curves of different participants shows a clear overfitting case for `Lubozs`, who was first on the final public leaderboard but slipped to seventh position in the private leaderboard. It turned out that he had indicated in his blog that he had set up an automatic cron job, which was automatically re-submitting new submissions (five times per day, which was the maximum allowed by the platform) with slightly altered parameters to maximise his public score. This engineering feat allowed him to select a lucky spike and grab the top of the public leaderboard, but this did not fool the private leaderboard measurement. It should be noted that this

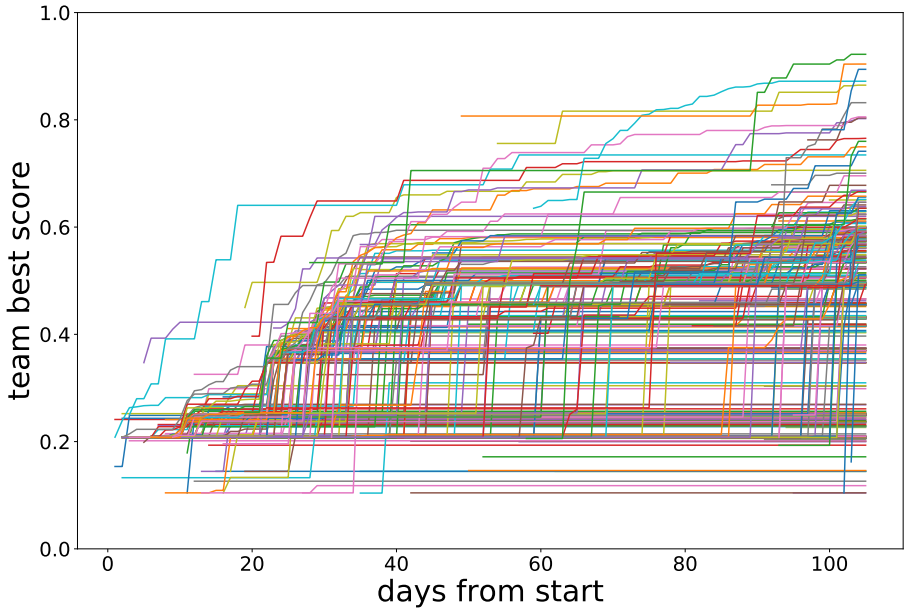


Figure 1: Participants' best score evolution as a function of days in the competition (Amrouche et al., 2019).

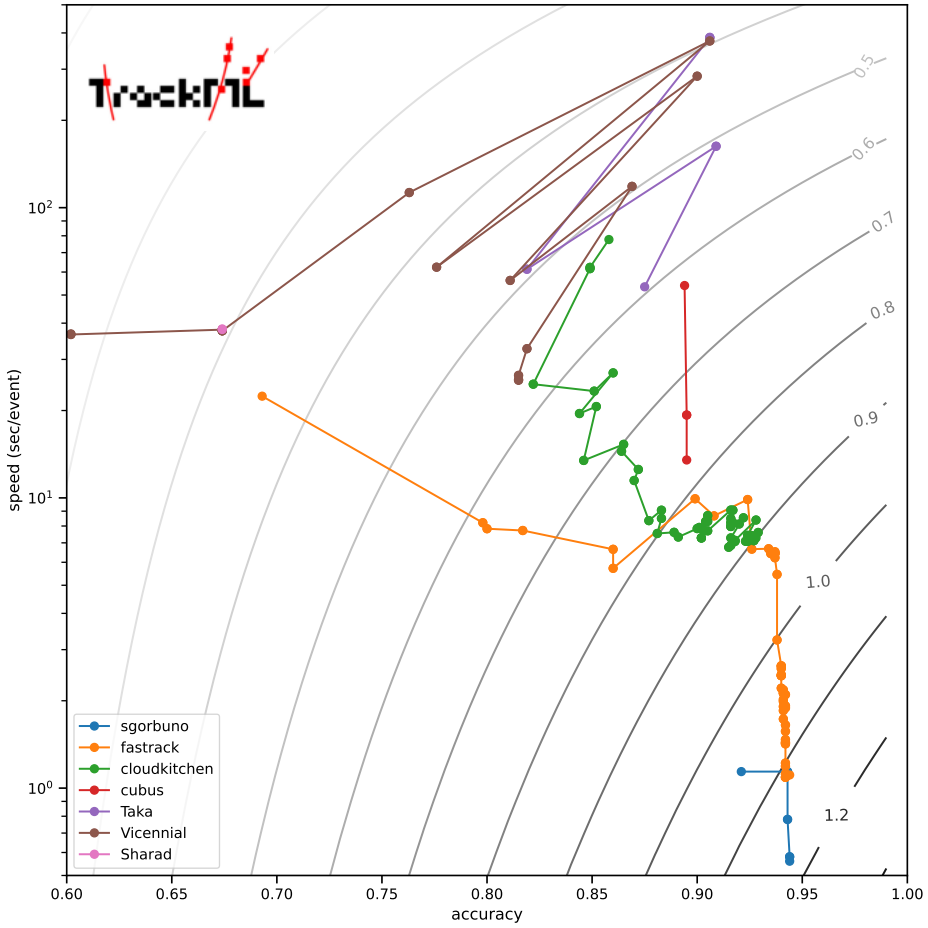


Figure 2: Participants score evolution: the horizontal axis is the accuracy, and the vertical axis is the inference speed. The total score, a function of both variables, is displayed in grey contours. Each colour/marker type corresponds to a contributor; the lines help to follow the score evolution(Amrouche et al., 2021).

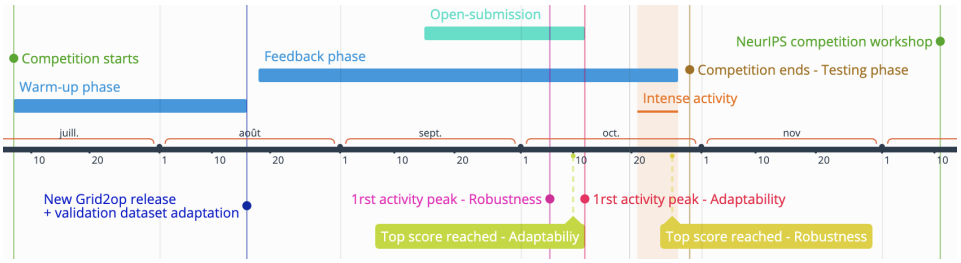


Figure 3: A high-level retrospective competition period timeline can help understand the competition organisation through phases and noteworthy events such as competition adjustments, peak of activities, performance outbreaks, and collaborative periods. This can support the competition narrative. (Marot et al., 2021)

piece of information was not reported through the standard channels or forum; it was discovered by googling, illustrating the point of analysing unorganised information as discussed in Sec. 2.5.

Additional graphs can be produced from by-products of solution submission gathered by the platform. For example, Fig. 6 compares the performance on different datasets, giving for each the intrinsic and modelling difficulties. Ideally, organisers should choose datasets of low intrinsic difficulty and high modelling difficulty, which would be in this example `Isaac2`, `Caucase` or `freddy`.

One can also analyse the solutions to evaluate their originality. For example, Fig 7 shows the dendrogram of a clustering competition. The dendrogram shows how similar or not the clusters found are, which correlates well with what is known on the participant's method or background: the set of six participants at the top of the diagram (#12 to #9) have highly optimised the starting kit using generic clustering algorithms; #3 and #4 are domain experts with the same background, who have found similar clusters with optimised domain methods; #1 is a CS student who has spent a lot of time studying the domain literature, has seen similar clusters as #3 and #4; on the other hand, the diagram sets apart #2 who is a CS expert who has developed a very original approach (which turned out to be impractical because of the significant resource it requires).

4.5 Deeper analysis of the submissions

After a challenge finishes, we often need a systematic and deep analysis of the winning solutions. The analysis could be very case-specific depending on the challenge task and application. As a consequence, we only mention a few general analyses here.

Reproducibility As an essential aspect of machine learning, reproducibility often should be considered in challenge organisation. Indeed, in post-challenge analysis, we should first reproduce winning methods to have a sanity check when the code was not submitted to the platform. The training should be reproduced even if it can be in practice challenging.

Metrics. In a challenge, organisers often use a single metric to evaluate people's submissions. The choice of the metric might come from organisers' interests, but we often need additional metrics to fully evaluate, understand and conclude with winning solutions. For example, in a classification task, we may need accuracy and balanced accuracy to pay attention to the skewness of dataset classes; in a regression task, we may need to mean squared error (MSE) and mean absolute error

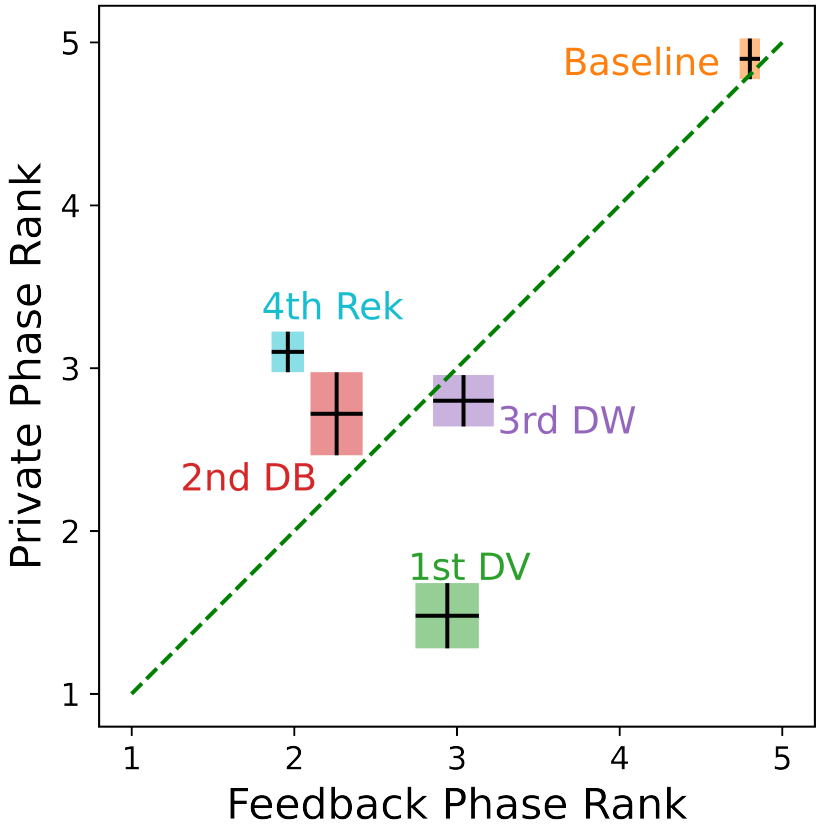


Figure 4: Overfitting/Ranking stability plot from (Xu et al., 2021). The comparison of the x/y axes shows overfitting. In a challenge of two phases, the feedback (or public) phase is the first phase, and the private phase is the second. By showing how submissions perform in the consequent two phases, we demonstrate the overfitting of algorithms. The rectangle shows ranking stability around each team. This rectangle is calculated by the average ranks of multiple reproductions of submissions' performances.

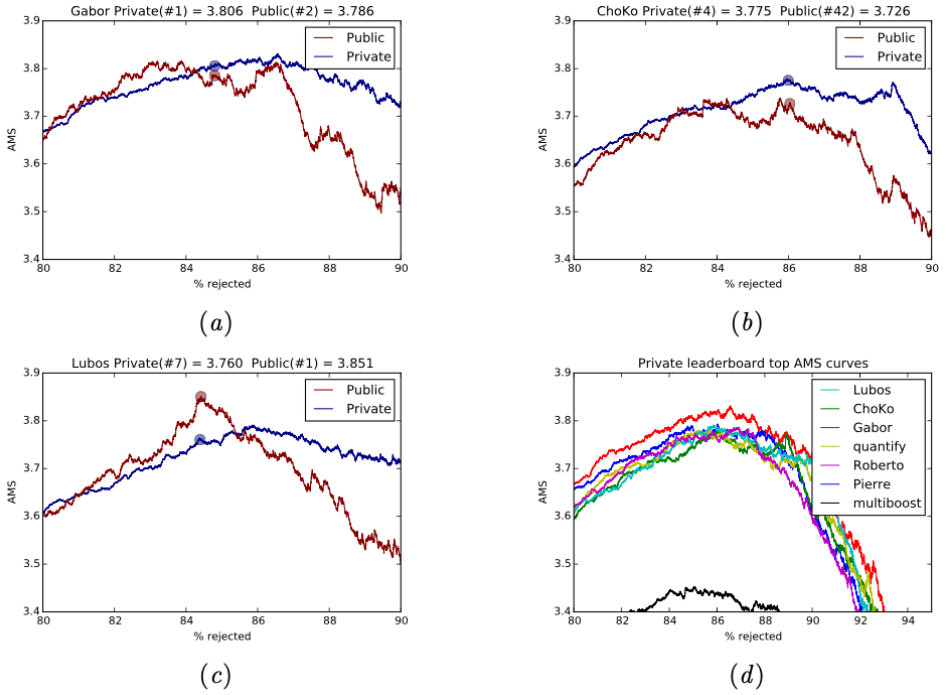


Figure 5: AMS Significance, a figure of merit of a classifier in a physics particle discovery context, as a function of decision threshold for different submissions in the HiggsML challenge (Adam-Bourdarios et al., 2014). The submission score is the maximum of the curve. (d) shows the private curves for different participants. (a),(b) and (c) compare the public and private curves from three participants, `Gabor`, `Choko` and `Lubozs`, with dots indicating their maximum values. The private curves are smoother because they are evaluated using more examples.

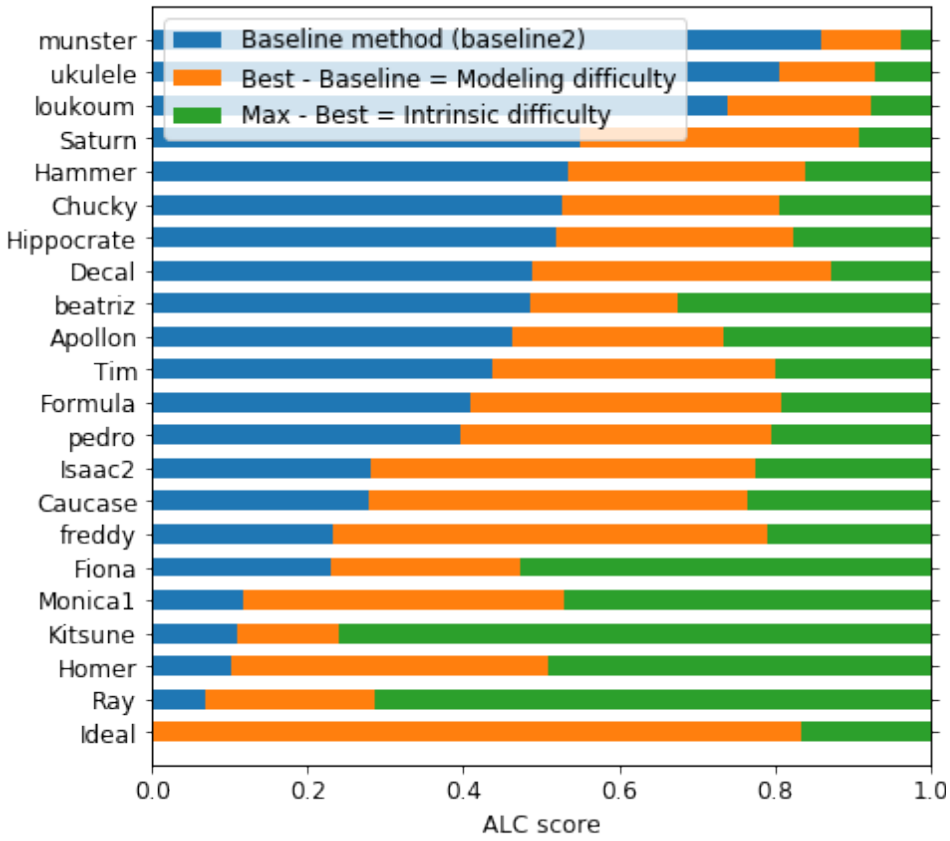


Figure 6: Measurement of task difficulty from (Liu et al., 2020). Each row corresponds to one dataset. Two difficulties are calculated here: intrinsic difficulty (maximum score minus best participant's score), shown with the green bar and modelling difficulty (best participant's score minus baseline score), shown with the orange bar.

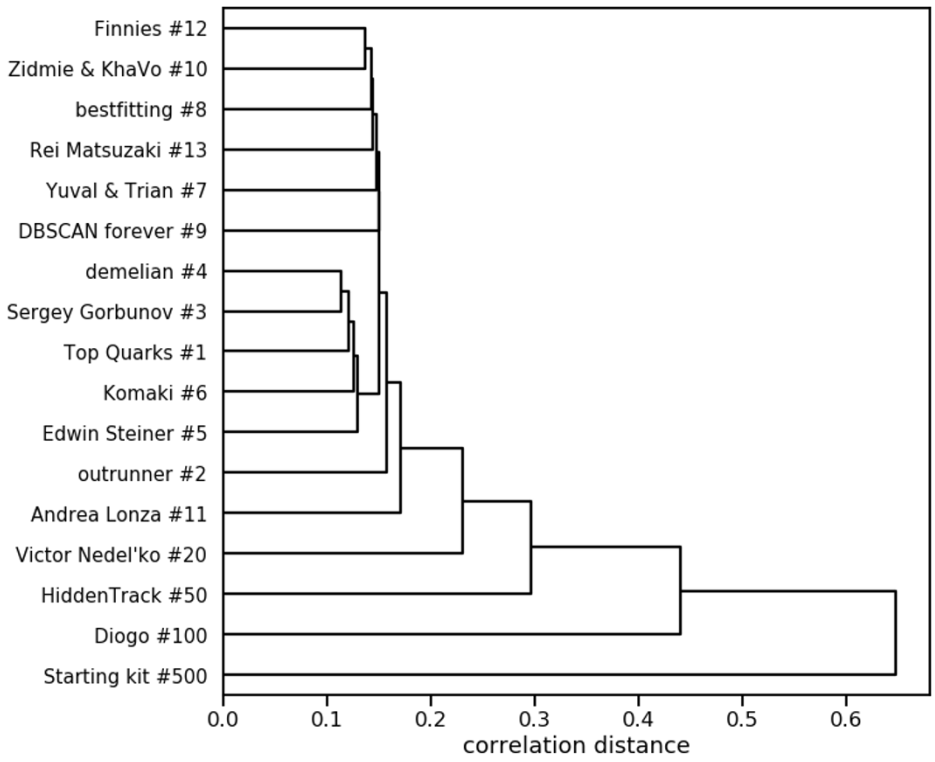


Figure 7: Dendrogram of the best solutions submitted by participants (name and final rank indicated) to the TrackML challenge (courtesy of authors of (Amrouche et al., 2019)). The diagram shows the thirteen best participants, plus participants ranked 20^{th}, 50^{th}, 100^{th}, as well as the starting kit.

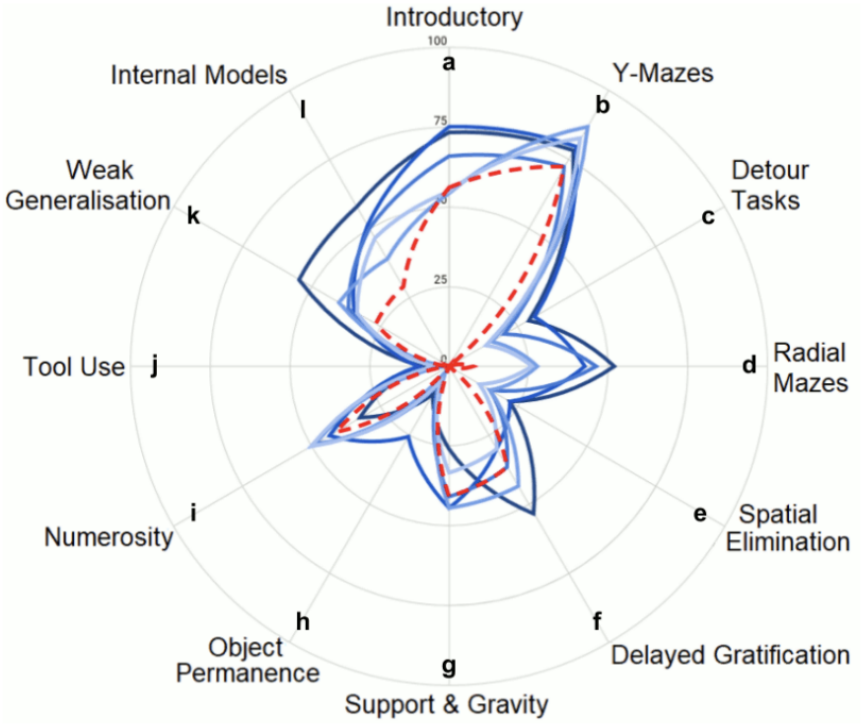


Figure 8: When a competition involves different tasks or different underlying metrics, a radar plot can provide a nice visual comparison of the submissions along all those dimensions as (Crosby et al., 2019) which tested agents on several tasks.

(MAE) for relative and absolute error analysis. Even though the organisers chose a fused version of multiple metrics, we could also evaluate each error component of the metric individually to understand the differences made by different methods.

For example, Fig. 8 shows the performance of different submissions for different tasks.

Another example, in Fig. 9 shows the performance of some of the best participants in a clustering competition (same participants as in Fig. 7). Although the participants had to optimise a single score, domain experts were satisfied to see these graphs showing that the best submissions maximise their cluster-finding algorithm's robustness (concerning ground truth parameters). One notable exception is #100, the only one showing a rising contribution in the bottom left graph. It indicated it was accidentally optimised for abnormal clusters, yielding a poor overall score, which was still interesting to domain experts.

Pipeline. A challenge solution usually consists of a pipeline of steps, e.g., data preprocessing, feature engineering, model training, hyperparameter selection, ensemble, etc. It is super interesting to investigate step by step the choices available from participants and ablation study the contribution of options. Such a study is not trivial because we need to split the steps of solutions, which is not necessarily logically clear; secondly, we need to modify many solutions to evaluate the options.

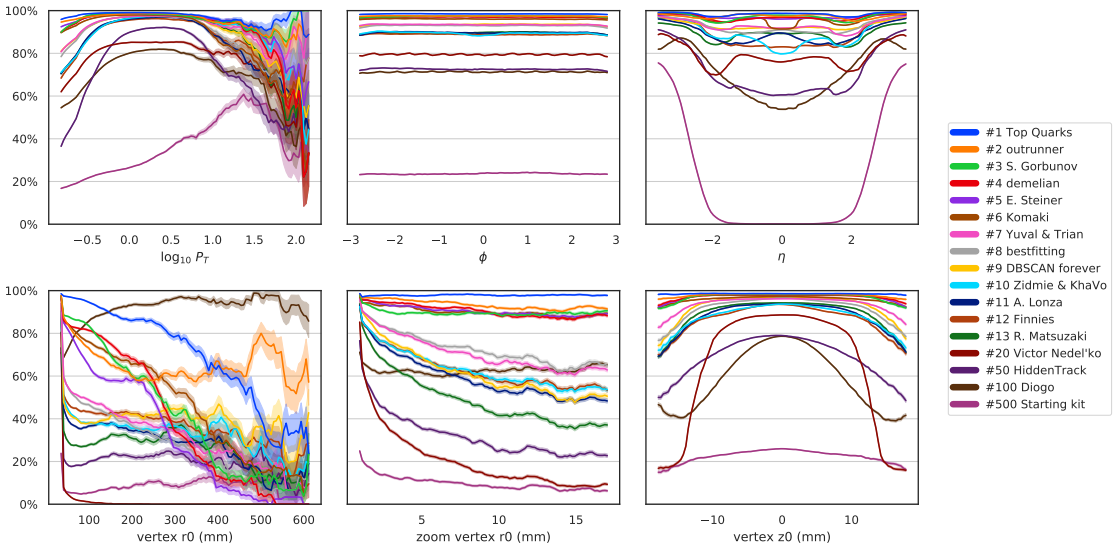


Figure 9: Cluster finding probability as a function of six ground truth cluster 3D shape and dimension parameters, from the TrackML Accuracy competition (Amrouche et al., 2019)

Once an ablation study is done, we could further ensemble the best options and output a revised version.

Ensembling. For each submission, the participant made a decision for each pipeline step. For some challenge types (i.e. classification), it should be possible to automatically ensemble different submissions to see if something could be gained from combining them (Kégl et al., 2018). In general, human ensembling (mixing and matching the various decisions taken at each step) can bring insights and improvements to even the very best submission. For example, one participant has built clever features but could have used them better.

Generalizability. In addition to the setting used in a challenge, we could also look at the generalizability of methods under different settings, including different datasets, time resource constraints, memory scalability, etc. This concept of generalizability has already been taken into account in AutoML challenges. For non-AutoML challenges, it is thus interesting to investigate.

Although these analyses are time-consuming, they can be very fruitful and worth the effort; they should be seen as the final processing step of the challenge's fruits (the submissions). They could be the theme of post-challenge workshops.

4.6 Description of the most interesting submissions

If one paper is written, which hosts contributions from the most interesting submissions (not necessarily the absolute best but also the ones deemed original), this section would hold sub-sections, each describing a submission. Participants should write these sub-sections themselves, following guidelines or a template provided by the organisers. Otherwise, the organisers can write themselves one or more subsections based on the material available to them.

If the participants have written separate papers, only summaries would be necessary here.

4.7 Scientific outcome

This section summarises the scientific outcome of the challenge based on the deeper analysis of the submissions and the individual submission descriptions.

- What new insights were brought by the challenge for data science and the domain?
- What techniques worked best in the different pipeline steps?
- What about the explainability of these techniques?
- What are the hardware/resource constraints?
- Are these originals in absolute or in this particular domain?
- What techniques did not work?
- What are the new avenues for further studies?

4.8 Lessons on the challenge organisation itself

The scientific outcome is the primary lesson from a challenge. However, another important one is the feedback on the organisation of the challenge itself.

Answers to the following questions should be sought:

- Did the participants solve the problem they were supposed to solve?
- Any fundamental flaw in the competition?
- Was the metric appropriate? How could the metric be improved?
- Was the dataset suitable? How could it be improved?
- Were the tasks of the challenge of a difficulty adapted to push the state-of-the-art in the domain considered?
- Feedback on the platform and the challenge mechanism.
- Some measurement of the popularity of the competition and comments on advertisement and dissemination.
- Participants sociology diversity (from fact sheet)

4.9 Conclusion

The paper's conclusion would summarise the scientific findings and sketch possible future actions, permanent datasets and benchmarks or future challenges.

Table 1: Comparisons of competition and benchmark (Xu et al., 2022).

	Competition	Benchmark
Purpose	Crowdsourcing problems in a short time and harvesting solutions	Continuous fair evaluation over a long period, in a unified framework
Phases	Multiple phases	Single-phase
Time period	Usually limited	Often never-ending
Cooperation & information sharing	Limited due to the competitive nature	As extensive as possible
Submissions	Usually algorithm predictions or algorithm code	Algorithm code or datasets; code or dataset name, description, documentation meta-data and fact-sheets; scoring programs for custom analyses
Outcome	leaderboard with usually a single global ranking based on one score from each team (last or best)	Table with all the submissions made; sorting with multiple scores possible; multiple analyses, graphs, figures, code sharing

5 Post challenge benchmark

Benchmarks differ from competitions in many ways, as summarised in Table 1. We organise a competition to crowdsource a task and harvest the winning solutions. This competition usually lasts a couple of months and has multiple phases (public, private, etc.). We intentionally rank participants linearly due to their competitive nature, and people are not allowed to share the code directly. However, for benchmarks, we are interested in a research task and would like to invite people worldwide to contribute continuously. The benchmarks last much longer than competitions, usually never-ending, and only one phase is associated. Another big difference is that benchmarks encourage people to share ideas, solutions, code, and findings as much as possible because the goal is to push forward this research task. Thus, rich publications, seminars, and workshops are expected for communication.

By turning a challenge into a benchmark, we gain multiple benefits for different people. Challenge organisers give people around the world more time (possibly never-ending) to join the benchmark and make submissions to push forward the research task. Participants have more time and possibly can open-source datasets for their own research, and the leaderboard of research gives credit to their methods. For the platform of benchmarks, it is always better to have high-quality benchmarks and attract more people.

Turning a challenge into a benchmark is usually labour-intensive and repetitive. We hereby develop the codabench project to host benchmarks easily and free of charge. Technically, organisers only need to prepare data, logistic code for digesting and evaluating, and a configuration file. Benchmarks will be run in separate dockers; thus, the results will be reproducible.

6 Conclusion

In the chapter, we have covered the main actions in favour of the long-lasting impact of a challenge, particularly with a post-challenge paper template and advice on how to turn the challenge into a benchmark.

The main point is that significant time and person-power resources should be allocated up-front to this activity, which will harvest and make sense of the wealth of information produced by the challenge. For example, if the challenge is funded through a call, it would be best to foresee at least one year for the post-challenge activities.

References

Claire Adam-Bourdarios, Glen Cowan, Cécile Germain, Isabelle Guyon, Balázs Kégl, and David Rousseau. The Higgs boson machine learning challenge. In *HEPML@ NIPS*, pages 19–55, 2014. URL http://www.jmlr.org/proceedings/papers/v42/cowa14.pdf.

Sabrina Amrouche, Laurent Basara, Paolo Calafiura, Victor Estrade, Steven Farrell, Diogo R. Ferreira, Liam Finnie, Nicole Finnie, Cécile Germain, Vladimir Vava Gligorov, Tobias Golling, Sergey Gorbunov, Heather Gray, Isabelle Guyon, Mikhail Hushchyn, Vincenzo Innocente, Moritz Kiehn, Edward Moyse, Jean-François Puget, Yuval Reina, David Rousseau, Andreas Salzburger, Andrey Ustyuzhanin, Jean-Roch Vlimant, Johan Sokrates Wind, Trian Xylouris, and Yetkin Yilmaz. The tracking machine learning challenge: Accuracy phase. In *The NeurIPS 2018 Competition*, pages 231–264. Springer International Publishing, November 2019. doi: 10.1007/978-3-030-29135-8_9.

Sabrina Amrouche, Laurent Basara, Paolo Calafiura, Dmitry Emeliyanov, Victor Estrade, Steven Farrell, Cécile Germain, Vladimir Vava Gligorov, Tobias Golling, Sergey Gorbunov, Heather Gray, Isabelle Guyon, Mikhail Hushchyn, Vincenzo Innocente, Moritz Kiehn, Marcel Kunze, Edward Moyse, David Rousseau, Andreas Salzburger, Andrey Ustyuzhanin, and Jean-Roch Vlimant. The tracking machine learning challenge : Throughput phase, 2021. URL https://arxiv.org/abs/2105.01160.

Matthew Crosby, Benjamin Beyret, Murray Shanahan, José Hernández-Orallo, Lucy Cheke, and Marta Halina. The animal-ai testbed and competition. 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 164–176. PMLR, 2019. URL http://proceedings.mlr.press/v123/crosby20a.html.

Balázs Kégl, Alexandre Boucaud, Mehdi Cherti, A. O. Kazakçi, Alexandre Gramfort, Guillaume Lemaître, Joris Van den Bossche, Djalel Benbouzid, and Camille Marini. The ramp framework: from reproducibility to transparency in the design and optimization of scientific workflows. In *ICML workshop on Reproducibility in Machine Learning*, 2018. URL https://openreview.net/pdf?id=Syg4NHz4eQ.

Zhengying Liu, Zhen Xu, Sergio Escalera, Isabelle Guyon, Júlio C. S. Jacques Júnior, Meysam Madadi, Adrien Pavao, Sébastien Treguer, and Wei-Wei Tu. Towards automated computer vision: analysis of the autocv challenges 2019. *Pattern Recognition Letters*, 2020.

Antoine Marot, Benjamin Donnot, Gabriel Dulac-Arnold, Adrian Kelly, Aidan O'Sullivan, Jan Viebahn, Mariette Awad, Isabelle Guyon, Patrick Panciatici, and Camilo Romero. Learning to run a power network challenge: a retrospective analysis. In *NeurIPS 2020 Competition and Demonstration Track*, pages 112–132. PMLR, 2021.

Zhen Xu, Wei-Wei Tu, and Isabelle Guyon. Automl meets time series regression design and analysis of the autoseries challenge. In *European Conference on Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track, ECML PKDD*, Lecture Notes in Computer Science. Springer, 2021.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543, 2022. doi: 10.1016/j.patter.2022.100543. URL `https://doi.org/10.1016/j.patter.2022.100543`.

Academic Competitions

Hugo Jair Escalante HUGO.JAIR@INAOEP.MX
Instituto Nacional de Astrofísica,
Óptica y Electrónica,
Tonantzintla, 72840, Puebla, Mexico

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

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

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.

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).

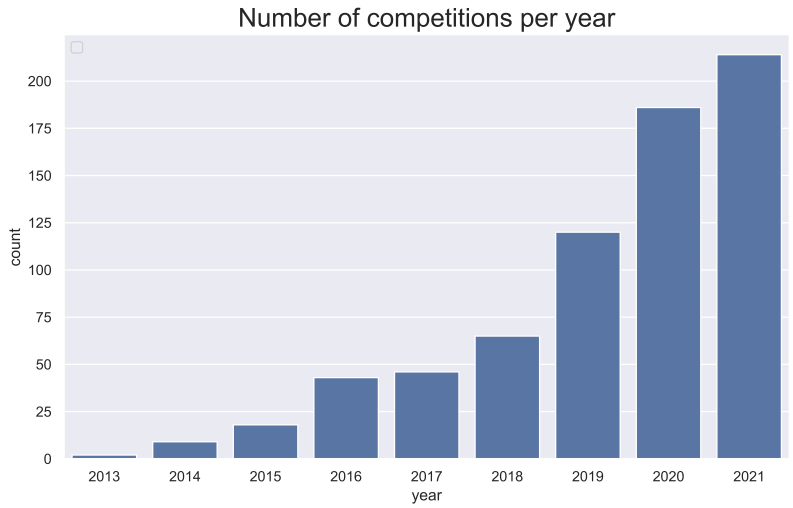


Figure 1: Evolution of the number of competitions each year. Data gathered from *CodaLab Competitions* (Pavao et al., 2023), a platform with a community focused on academic competitions.

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. Finally in Section 4 we outline some thoughts and ideas on the future of academic competitions.

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.

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

2.1 Historical review

While it is a daunting task to give 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 2022 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.

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

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/

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 MNIST is a *biased* dataset, and other versions of it 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 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).

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/
14. http://clopinet.com/isabelle/Projects/NIPS2003/

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 longlasting 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 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, 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

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 2022. 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	13	https://icdar2023.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	4	https://pan.webis.de/
TrecVid	MIR	2003	8	https://trecvid.nist.gov/
ImageCLEF†	MIP	2003	4	https://www.imageclef.org
MediaEval	MIP	2003	11	https://multimediaeval.github.io/
GECCO	EC	2004	10	https://gecco-2022.sigevo.org/HomePage
WCCI	CI	2006	13	https://wcci2022.org/accepted-competitions/
MICCAI	MI	2007	38	https://conferences.miccai.org/2022/en/
Interspeech	SP	2008	2	https://interspeech2022.org/
ICRA	RO	2008	10	https://www.icra2022.org/
ACM Multimedia	MIP	2009	10	https://2022.acmmm.org/grand-challenges/
ICPR	PR	2010	7	https://www.icpr2022.com/
SemEval	NLP	2010	12	https://semeval.github.io/
IROS	RO	2012	9	https://iros2022.org/program/competition/
ICMI	MIP	2013	1	https://icmi.acm.org/2022/
ICASSP	SP	2014	8	https://2022.ieeeicassp.org/
ICME	MIP	2015	2	https://2022.ieeeicme.org/
CIKM	DM	2017	2	https://www.cikm2022.org
ICIP	IP	2017	4	https://2022.ieeeicip.org/
NeurIPS	ML	2018	25	https://neurips.cc/Conferences
IJCAI	AI	2018	4	https://www.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/

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 sceptical 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

23. https://automl.cc/

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.

- **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

24. `https://autodl.chalearn.org/`
25. `https://metalearning.chalearn.org/`

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

As previously mentioned, 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.

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.

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

- **Scientific rigour.** 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 2021 can be found in (Escalante and Hadsell, 2019; Escalante and Hofmann, 2020; Kiela et al., 2022).

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

27. https://www.aicrowd.com/challenges/nips-2017-learning-to-run

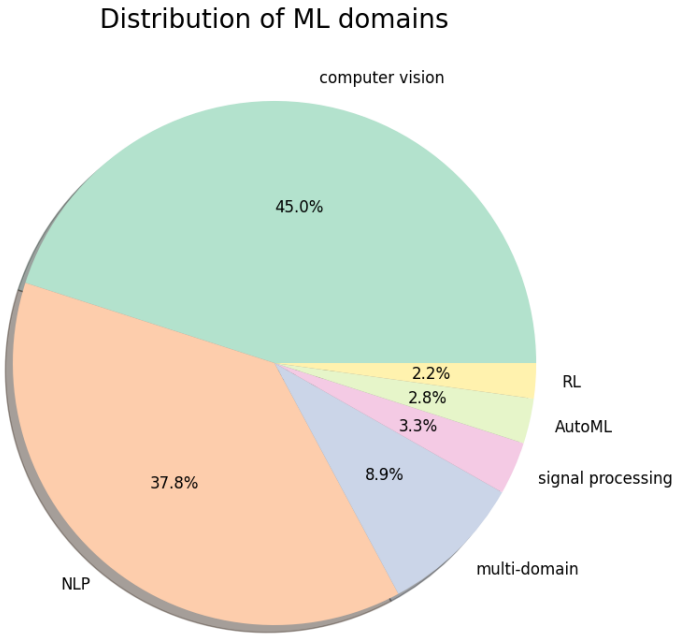


Figure 2: Distribution of competitions with different machine learning domains. Data gathered from CodaLab Competitions (Pavao et al., 2023)

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 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].

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

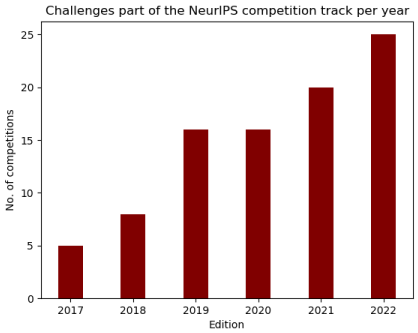


Figure 3: Number of challenges organized as part of the NeurIPS competition program.

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).
- **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

35. https://www.iarai.ac.at/traffic4cast/
36. https://www.duckietown.org/research/AI-Driving-olympics
37. https://minerl.io/

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.

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.

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/

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 scenarios 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

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/

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 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.

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/

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.

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).

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/

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 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 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:

- **Data centric competitions**[61] This is competitions where the goal is to improve a dataset by applying so called, data-centric techniques, like fixing mislabeled samples, finding prototypes, border points, summarization, data augmentation, etc.

- **Cooperative competitions.** Coopetitions is a form of crowd sourcing in which participants compete to build the best solution for a problem, but they cooperate

61. `https://https-deeplearning-ai.github.io/data-centric-comp`

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[62]). We foresee more dedicated venues will be available in the next few years.

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.

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`.

62. `https://www.springer.com/series/15602`

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-data-science-2022.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

Placeholder PDF

Competitions and challenges for education and continuous learning

Kristin P. Bennett BENNEK@RPI.EDU
Rensselaer Polytechnic Institute, USA

Isabelle Guyon GUYON@CHALEARN.ORG
Université Paris-Saclay, France, ChaLearn, USA, and Google, USA

Adrien Pavão PAVAO@MLCHALLENGES.COM
Université Paris-Saclay, France

Evelyne Viegas EVELYNEV@MICROSOFT.COM
Microsoft, USA

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

Abstract

Involving students and trainees both in organizing and in participating in competitions and challenges is a powerful pedagogical tool. From kindergarten through the last years of graduate studies, or as a way to upskill to gain new skills on the job, competitions can gamify the learning process and thus motivate people to explore by themselves, assimilate a variety of material, and expand their capabilities. Competitions also contribute to engaging students in learning about problems of societal importance. At the graduate level, designing and implementing competitions can be seen as a hands-on means of learning proper design and analysis of experiments. In the workplace, this remains an effective way to embrace continuous learning and remain up to date as technologies, tools, and approaches evolve. In this chapter, we present a synthesis of various hands-on teaching and learning experiences using competitions as a medium. We report educational efforts conducted in kindergarten, high school, university and working environment. For younger students, competitions take more the form of individual projects qualitatively evaluated by a jury, while at the university level they take the form of automatically graded homework addressing a research problem. The application domains can be very diverse: medicine, ecology, marketing, computer vision, recommendation, text processing, etc., and students enjoy being involved in creating challenges motivated by humanitarian purposes. Within a professional or working environment, competitions are used to help with upskilling, learning new approaches, and tools that can directly be used in one's job or to update one's resume to remain relevant and competitive in the market place.

Keywords: education, teaching, continuous learning

1 Introduction

Competitions and challenges (together referred to as skill-based contests) are a form of "serious games". Already in 2017, the NeurIPS workshop on "Challenges in Machine Learning" (CiML) focused on gaming and education made the connection between challenges and games. Since then, hundreds of challenges in Machine Learning and AI have been organized on platforms such a Kaggle community competitions Goldbloom and Hamner (2010) and RAMP Kégl et al. (2018), as well as Codalab competitions Pavao et al. (2022).

According to Nguyen (2021): "Research shows that using games in teaching can help increase student participation, foster social and emotional learning, and motivate students to take risks."

Engaging in games or competitions as part of an educational program provides a more collaborative and engaging experience, particularly if team work is encouraged, for students who struggle with passive learning, heavy on lectures and book reading. During Covid times, teaching via participating (or organizing) challenges helped students being connected and motivated remotely, by working on a project together. These claims are supported by research about using games in teaching Plass et al. (2015); Wu et al. (2012); Ifenthaler et al. (2012), which presumably helps increase student participation, foster social and emotional learning Hromek and Roffey (2009), and motivate students to take risks.

While the terminology "competition" and "challenge" is often used interchangeably, in this chapter, we use the following definitions. We consider only scientific evaluations running over a finite amount of time, referred to as "skill-based contests" (as opposed to benchmarks, which are run on an on-going basis and "chance contests" in which chance plays a role in winning). We call **"challenge"** those skilled-based contests in which participants are called to solve **organizer-specified tasks** evaluated by given metrics. Challenge entries are usually results of predictions made on a challenge test set or code to perform a specific task. Short challenges are sometimes referred to as **"hackathons"**. In contrast, **"competitions"** are more open-ended. The participants can be called to **define and solve their own problem**. Competition entries can include reports, prototypes, demonstrations, live performances, presentations, etc. Competitions are often judged by a jury on the basis of both quantitative and qualitative metrics, while challenge evaluations are usually automated and obtained by computing scores according to pre-defined metrics that are posted on a leaderboard.

Neither competitions nor challenges are not substitutes for other forms of learning. Like any educational tool, they need to be well-planned and integrated only when they are relevant to the learning objectives. In this chapter, we explore various aspects of including competitions or challenges as part of curricula, as participants but also as organizers, then analyze various case studies.

2 Students entering a competition

A seemingly easy way to engage students at low preparation investment is to enroll them in a competition organized by a third party. This can constitute a **project-based class** giving a lot of freedom of creativity to students.

At the K-12 level, that is between kindergarden and last year before college or university, Technovation offers programs to engage underrepresented groups, particularly young women (ages 8-18) to become leaders, creators and problem-solvers. The competitors form teams coached by parents and educators to solve real world problems, such as finding safe drinking water, identifying and removing invasive species; monitoring air quality, etc. Technovation organizes an annual competition, through which participants identify problems in their communities and use mobile and AI technologies to develop solutions. Technovation has been in operation for 14 years, reaching more than 160,000 participants through its competitions. See Section 6 for more details.

At the high-school level, robotics competitions offer an opportunity the break the ice with intimidating technology. Well known competitions include the "First Lego League" and the "Sumo Robot League". Then under-graduate level competitions include Duckietown, which offers each year several AI Driving Olympics competitions (AI-DO) for small self-driving robots, which can be purchased at a low price and self-assembled. The Duckietown project[1] was conceived in 2016

1. https://duckietown.com/ai-do-at-neurips-is-over-congratulations-to-our-winners/

as a graduate class at MIT. The Duckietown Foundation debuted AI-DO in December 2018 at the NeurIPS conference. The platform is used at several universities around the globe, including NCTU in Taiwan, Tsinghua in China, and RPI in the United States. Another well-known and well established competition is RoboCup, which offers junior level entry leagues, all the way to advanced professional leagues.

At the graduate level, students can directly enter "high profile" international competitions. This can be very motivating, as the prize, which can reach thousands to hundreds of thousands of dollars, can help funding their PhD. research and beyond. For example, the XPrize has been proposing numerous bold challenges, from sending a rocket to the moon to providing solutions to the Covid crisis. Recently, they have been attracting a lot of attention with the XPrize Carbon Removal, aimed at fighting climate change and rebalancing Earth's carbon cycle. Funded by Elon Musk and the Musk Foundation, this $100M competition is its largest incentive prize. For the first milestone, several student teams were awarded 100,000 dollars to pursue their research, including women-led U. Miami student team for their proposal of ocean-based carbon removal, and women-led U. Washington student team for their novel carbon soil gas monitoring sensors.

3 Students entering a challenge

Compare to competitions, challenges are a more contrained format of skill-based contests. They usually organized on a challenge platform, such a Kaggle community competitions, RAMP, or Codalab competitions. Many machine learning, natural language processing, and computer vision conferences organize challenges every year, so it is not difficult to find a suitable challenge to teach a class. For example, in 2022, the NeurIPS conference offered 25 high-end peer reviewed challenges.

The benefit of using challenges in teaching includes motivating students and facilitating grading (since the leaderboard scores can readily be used as part of the grade). In our experience, it is important to have a very structured class to set the student expectations, and use the competition or challenge as a medium (a means to an end), avoiding to emphasize winning as the essential goal. This last point is facilitated by grading the students on several aspects other than winning (e.g. quizzes, oral presentation, written report) and organizing students in **teams** and/or **leagues**. Teams are groups of students working together towards solving a problem making challenge entries; leagues are sets of teams, generally of a similar level or interest, competing with one another. See Section 5 for more recommendations on grading.

For project-based classes (or for the final project of a class), one may consider **letting student choose their own challenge** and just deliver a report (possibly assorted with in-class presentations or a poster session). The instructor may want to narrow down choices for multiple reasons: finding a good challenge is time consuming, and students are not necessarily the best judges of what a good challenge is. They can choose a challenge, which is either too easy or too hard, or which does not offer good learning opportunities. The choice of challenges does not necessarily need to include only on-going challenges. In fact, entering a challenge, which is already over, presents multiple benefits: lowering the barrier of entry (with available solutions already published), and diminishing the importance of winning (in favor of learning). However, the students must then highlight clearly, in their report, their personal contribution.

Another way of incorporating challenges in a class is to design homework as challenges, each assignment being delivered as an entry in a mini-challenge. In that case, the instructor(s) design the challenge. Examples of such challenges include:

- Iris data challenge, organized to a beginner undergraduate class at U. Paris-Saclay, 105 participants.
- Artificial Neural Networks and Deep Learning 2021, organized by Politecnico Milano, 486 participants.
- Linear model contest, organized by U. Washington, 322 participants.
- Black Box Optimization challenge, organized by AI master optimization class at U. Paris-Saclay, 29 participants. A simplified remake of NeurIPS 2020 BBO challenge.
- Evolutional Reinforcement Learning, organized by AI master RL class at U. Paris-Saclay, 29 participants.
- Prediction of mortality given medical records, organized repeatedly by RPI used in data science class taught for serveral years with 50 to 91 participants each semester. See case study.
- ChemsRUs, organized repeatedly by RPI for a data science class (predict biodegradability of molecules) with 50 to 91 participants.See case study 8.

Some of these challenges are re-makes of large international challenges, which have been simplified. Obviously, having good introductory material, e.g. in the form of a R-notebook or a Jupyter-notebook, is essential. Also important: one must avoid wasting student's time with complex installation instructions. Thus, it is advisable to rely on ironed-out tools, such as:

- Anaconda Python installation;
- Google Colab on-line Jupyter notebooks;
- Scikit-learn ML library Pedregosa et al. (2011);
- PyTorch Deep Learning framework Paszke et al. (2017).
- R Project for Statistical Learning R Core Team (2021).

Challenge-based classes can also be an opportunity to give students good habits, such as using revision control (Git and Github), learning about Dockers, etc.

It can also be very motivating for students to participate in challenges designed by other students. See Section 4.

4 Students organizing a challenge

Université Paris-Saclay offers each year a class on creation of AI challenges, which are then solved by other students (see a list of past challenges organized by students) as part of their class requirements. This approach is illustrated by Figure 1.

While it has become mainstream to let students enter competitions or challenges, as part of class projects or homework, little has been do so far to involve students in the **design and implementation of competitions** or challenges. Clearly this is a more difficult task. Indeed, sophisticated challenges can take several months or years of maturation, and the involvement of many researchers.

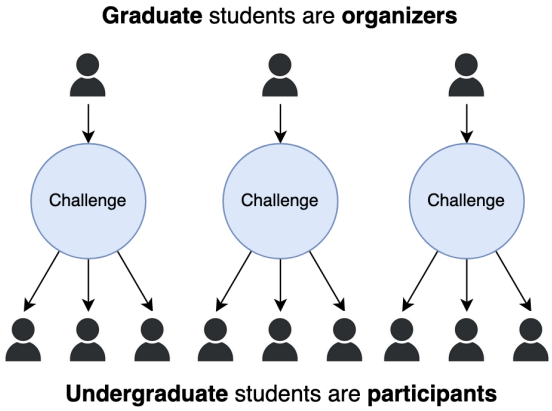


Figure 1: Graduate students organize challenges in which undergraduate students participate.

However, relatively simple challenges, of a level of difficulty that can be used to train undergraduate students, can easily be designed and implemented by graduate students, as part of class projects (typically **classification and regression problems**, but eventually recommendation or reinforcement learning problems). This allows them to gain **hands-on practice of experimental design** and harness the difficulties raised by:

- Defining well tasks and metrics,
- Collecting and preparing data,
- Ensuring that there are enough samples,
- Ensuring that no bias or data leakage is present in data,
- Preparing baseline methods.

In this process, students also learn about:

- Working in teams,
- Meeting strict deadlines,
- Acquiring good programming skills (including programming e.g. in Python or R and mastering toolkits such as scikit-learn and Keras, and learning about Github and Dockers),
- Preparing good didactic material (starting kit),
- Presenting their work (orally and in a written report),
- Promoting their challenge to engage as many participants as possible.

Emphasis is put on creating a fully working end-to-end "product" (a challenge), which will then be used by real "customers" (the undergraduate students). Quality of communication is also stressed by making the graduate students produce a short advertising video and presenting their challenge in class to the undergraduate student, who get to choose one of them for their project.

This type of educational program has been taking place since 2016 at Université Paris-Saclay Pavao et al. (2019) (see the 2021/2022 edition). Each year 30 to 40 graduate students create challenges as part of their master program in data science and about 100 second year undergraduate students solve them over a 12-week project period. This program has also used student-designed challenges as master-level projects. Already 1000 students have been trained through this program. Several alumni have become co-organizers of larger research challenges, which have been selected as part of the NeurIPS competition program, such as the TrackML particle physics challenge Calafiura et al. (2018), the AutoDL challenge series Liu et al. (2020), the Meta-Learning challenge series Carrión-Ojeda et al. (2022), and the reinforcement learning Learning to Run a Power Network (L2RPN) challenge series Marot et al. (2021, 2019).

Engaging graduate students in the design of challenges has an important far reaching impact. With the current rapid growth of AI research and applications, there are both unprecedented opportunities and legitimate worries about its potential misuses. In this context, it is important to familiarize students with good data science methodology, with respect to study design and modeling. Recognizing that there is no good data science without good data, it is important to educate students to conduct proper data collection and preparation. The objective should be to instill them good practices to reduce problems resulting from bias in data or non reproducible results due to lack of data. At RPI, the student challenge program has also been encouraging the protection of **data confidentiality or privacy** by replacing real data by realistic synthetic data Yale et al. (2020). This facilitates broadening access to undergraduate students to confidential or private data having a commercial value or the potential to harm individuals. Awareness should also be raised to issues related to **ethics and fairness**. To that end, Université Paris-Saclay offers a class on Fairness in Artificial Intelligence, in parallel with the challenge class.

From previously designed challenges constitute progressively a "library of challenge designs"[2]). Previously designed challenges can be cloned and serve as templates for new challenges. See chapter 10 for a tutorial on how to easily clone CodaLab challenges and create your own challenge or design a simplified template for your students as a starting point.

5 Evaluating teaching effectiveness and measuring impact

In this section, we provide a few tips on how to grade students and evaluate their learning experience. We note that using the raw scores on challenge leaderboards is not necessarily a good means of evaluating students. However, challenges can provide valuable and effective pedagogical tools for evaluation for busy teachers. Challenges offer a unique opportunity to measure such effectiveness, by monitoring engagement (number of submissions), collaborations (code sharing), good programming habits (e.g. good use of GitHub). Additionally, students can be asked to complete quizzes or surveys before, during, and after the course program.

Entry survey

Before the course begins, we (Isabelle Guyon and Adrien Pavão) usually ask students to complete an "entry survey", to check their level, interests, and level of motivation. The answers serve to form teams and assign students to a challenge corresponding both to their wishes and their skill level. We advocate that this survey should not be graded, but rewarded by a flat number of "bonus points" for

2. `https://saclay.chalearn.org/`

completing them. The students can complete the survey at home and are encouraged to search on the Internet to answer the questions, but answer them in their own words.

We use a similar survey regardless whether the class is about organizing a challenge or participating in a challenge. Typical questions we ask include:

- What is your level in Python programming?
- Have you already participated in a machine learning or data science challenge? If so, describe your experience and provide the URL.
- Explain the difference between supervised and unsupervised learning.
- Explain the difference between classification and regression.
- Explain what is "cross-validation".
- Explain the difference between aleatoric and epistemic uncertainty.
- Explain what "data leakage" is. Give an example.
- What machine learning toolkit do you prefer and why (Scikit-learn, Tensorflow, Keras, Pytorch, other)? Justify your choice.
- What to do think is hardest to deal with: too little or too much data? Justify your answer.
- What application domains are of interest to you?
  - Biology and medicine.
  - Ecology, energy and sustainability.
  - Internet, social media, and advertising.
  - Market analysis and financial data.
  - Image, audio, speech, video, and other sensor data.
  - Text processing, language understanding.
  - Physics and chemistry.
  - Ethics, fairness, and privacy.
  - Engineering, manufacturing, quality assurance.
  - Robotics and control.
  - Education.
  - Sports data analysis/prediction.
  - Games (Chess, Go, ...).
  - Generative adversarial networks.
- You will work as a team. Characterize your skills:
  - Good programmer.
  - Good sense of user interfaces.

- Good artistic sense.
- Well organized.
- Capable of coordinating a team.
- Good written English.
- Good oral English.
- Good practical experience of machine learning
- Good in statistics or learning theory.

It is impressive how efficient such questions are to evaluate, not only the level of the students, but their motivation and willingness to learn, and their capability to look for answers by themselves and assimilate them. Team leaders are chosen on the basis of self-declared capacity of leading a team, being well organized, and diligently trying their best to provide good answers. The rest of the students are generally grouped on the basis of topic affinity. Mixing weak and strong students in the same team is usually ineffective, because the stronger students end up doing everything and ignoring the weak students. Hence we tend to also regroup students by strength, as a secondary criterion after topic affinity.

In our challenge classes at Université Paris-Saclay, we group students in teams of five or six people. The teams are made by the professor, based on the survey answers. The team members can then elect to dispatch among themselves complementary roles and eventually work in pairs: (1) Coordinator (will be responsible to submit all the homework in time); (2) Architect (will oversee the end-to-end "product" design); (3) Domain expert (will oversee the data preparation and task definition); (4) ML expert (will oversee the preparation of the baseline software); (5) Data analyst (will oversee the production of interesting results); (6) Test engineer (will be responsible to make sure that everything works). The assignments of roles to people are flexible (more than one role per person and more than one person per role are possible).

Milestones and deliverables

For a challenge class to be successful, it is important to give the student, in advance, a clear schedule of milestones and deliverable. Depending on whether the class is a challenge design class or a challenge participation class, those are slightly different, but here are some common final deliverables:

- **Proposal:** the students must describe in a few pages what they plan to do.
- **Video:** a mini 3 minutes "promotional video", inspired by competitions of "my thesis in 180 seconds". This teaches them how to communicate succinctly and how to make a video.
- **Presentation:** a short 10-15 minute in-class presentation. This makes them practice public speaking and explaining technical contents.
- **Report:** a short 6-8 page techreport. This teaches them how to write a scientific paper with systematic comparisons of methods, graphs, and error bars.

For the challenge preparation class, other deliverable include:

- **Starting kit:** a Jupyter notebook serving as tutorial material to prepare and entry into the challenge, including sample data and a baseline method.

- **Website:** a working implementation of their challenge on a platform.

- **Tests:** a series of tests checking their final product.

For the challenge participation class, we usually intermix in-class practical work (based on Jupyter notebooks) to be finished at home as homework, quizzes, and code reviews. The ranking into the challenge weighs little in the final grade.

For details, see for example the Syllabus of the AI challenge creation class (2021/2022), and the Syllabus of the Data Science challenge class 2019/2020 at Université Paris-Saclay.

To keep the students engaged and motivated, all homework and deliverables have strict deadlines. However, if delivered on time, the instructors give only a temporary grade, which can be improved by submitting a second corrected version. This "second chance" method motivates students to work diligently and efficiently in the classes. Since they clearly know how to improve their grades from the instructor's reviews, the students usually work hard on the second versions.

Class evaluation

It is important to evaluate how much students assimilate the material and if they are encouraged to continue pursuing a career in data science or artificial intelligence. A post-program survey can help collecting such information.

For example, Rensselaer Polytechnic Institute (RPI0 formally evaluated it's low-barrier pipeline into applied undergraduate research consisting of an introductory data analytics course called Introduction to Data Mathematics Course (Case study 2 below), followed by a course-based undergraduate research experience (Case study 3 below) Bennett et al. (2022). The responses from 118 students, presented in Figure 2, show that the majority of respondents (80%) might or do plan to pursue additional courses or experiences relevant to data analytics. Almost half of the respondents plan to pursue Internships/Coops Related to Data Analytics (46%). Almost all respondents (93%) agreed or strongly agreed that as a result of taking the course they want to improve their computer skills. The great majority also recognized the value of data analytics to the healthcare field (87%), understand that data analytics will be of value regardless of their career path (94%), and aspire to use data to solve real world problems (85%). Most respondents want to learn ways to apply data analytics in other areas (90%) and the majority (67%), want to have a career in data analytics, almost double the previous response. Over three-quarters of respondents agreed or strongly agreed that they want to continue doing research projects that involve data analytics (80%).

Women are a third of students in these courses, which is proportionate to their representation at the Institute. The four most common majors of students are Mathematics (37%), Computer Science (22%), Biochemistry and Biophysics (15%), and biology (70%). A third of the undergraduates taking these courses are dual majors. The most common second major for almost half of the dual major students was mathematics (49%) followed by computer science (19%).

Measuring impact

Much remains to be done in terms of measuring impact. Nalia Kabeer presents a seminal framework for measuring empowerment, measuring increases in:

- **Resources:** Increased access to material, human, and social resources.

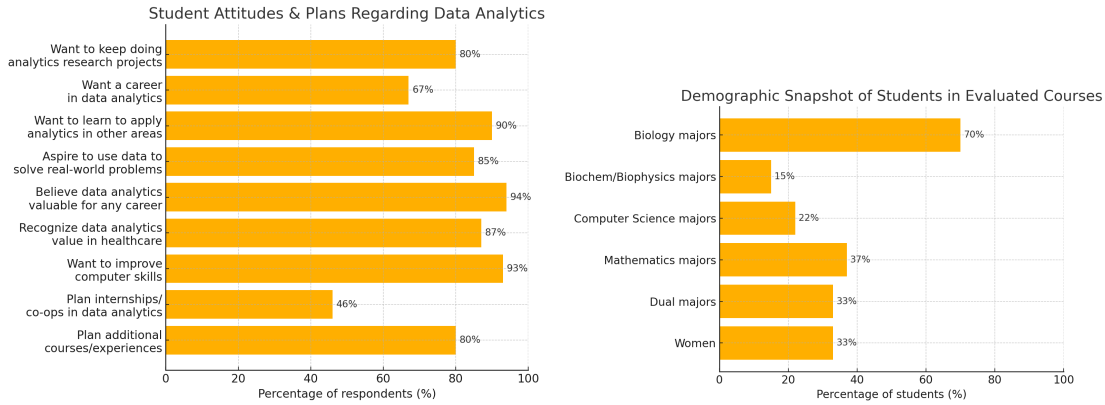


Figure 2: **Left:** Student attitudes and future plans related to data analytics after completing the evaluated courses. **Right:** Demographic snapshot of students enrolled in the evaluated courses.

- **Agency:** Increased abilities, participation, voice, and influence in the family, workplace, school, community.
- **Achievements:** Meaningful improvements in well-being and life outcomes that result from increasing agency and education (Kabeer, 1999).

We devote the rest of this chapter to describing three case studies of uses of challenges or competitions in education.

6 Case study 1: K-12 competitions

This section covers engaging school-age children at grade levels K-12 in competitions organized by Technovation. The purpose of running competitions in K-12 is to broaden participation, and building skills, particularly debunking myths around competitions and gender. How to run successful competitions at the K-12 level that broaden participation is rooted in Bandura's Motivation Theory.

Technovation puts forward two research frameworks:

- Higher dosage and practice so learners move from "situational interest" - participating in a program because it is offered in their community - to "well-developed, individual interest" - where the activity becomes a core part of their identity Hidi and Renninger (2006).
- Adaptation of Bandura's self-efficacy theory Bandura (1997), which outlines four pillars present in every successful human behavioral intervention:
  - Exposure to mentors and stories modeling lifelong learning.
  - Multi-exposure learning experiences that are authentic, engaging and meaningful.
  - Supportive cheerleaders who hold learners to high expectations while providing necessary support.

- High-energy, dramatic, suspenseful social gatherings/competitions that help the community feel collective pride (and adrenaline) at their accomplishments.

Technovation is a technology education nonprofit with a mission to empower vulnerable groups (especially girls and women) to create technology-based solutions to problems in their communities. Over the past 14 years, Technovation has engaged 50,000 mentors and educators to support more than 250,000 participants across 100+ countries to tackle pressing problems ranging from climate change to substance abuse—and most recently, COVID-19 (Technovation impact report, 2020).

Technovation organizes yearly innovation competitions to solve community problems, with a combination of inexpensive hardware and software, typically web apps or smart phone apps. The curriculum includes coding tutorials that help students apply key programming principles to real-world problems. Such tutorials include lessons on how to access and use databases (which will help teams who develop apps that help communities better distribute resources) as well as lessons on how to integrate maps and location-based data into student projects. Most real-world problems, like COVID-19 or the climate crisis, are complex and ill-defined, operate at multiple scales across different disciplines in dynamic ways, and may not have a clear end. To prepare young people to face these challenges, Technovation created a "Solve-It" video series and an associated checklist for educators, based on Donna Meadows' primer on Thinking in Systems.

One program is called "Technovation Girls". The 2020 season of brought together **20,388 girls from 62 countries**, who, with the support of 10,491 mentors, educators, and chapter ambassadors, designed a total of 1,520 tech-based apps addressing problems in their communities. Problems ranged from environmental protection to gender-based violence, to COVID-19. For example, Technovation worked with an attorney from Hogan Lovells to apply the "Solve-It" checklist to a real example of a complex problem Technovation Girls teams frequently address—domestic violence. The video walked girls through the process of brainstorming and developing solutions that can help protect survivors of domestic abuse by exploring different potential effects of the solution on those who use it. This helped students consider the different systems that domestic violence is situated within, rather than approaching it as a stand-alone issue.

Every year, Technovation invites teams of girls around the world to learn and apply the skills needed to solve real-world problems through technology, during a 12-week program. According to Technovation, after participating in "Technovation Girls", students express a greater interest in technology and leadership, and 58% of the alumni enroll in more computer science courses. Alumni go on to start their own businesses, present at prestigious events, meet world leaders, and even return to support the next cohorts of Technovation Girls. Over the last 15 years, Technovation has trained 150,000 young women to be technology entrepreneurs and innovators, empowering them to solve problems in their communities using technology. Five or more years after participating in Technovation programs, alumnae still credit the program with influencing their interests, career, and professional pursuits. Impact indicators include that 76% of Technovation alumnae pursue a STEM degree compared to 21% of female undergraduates who received STEM-related degrees in 2012; 60% of Technovation alumnae work in STEM-related positions, compared to the 29% national average in 2013; 50% of alumnae are leading change in their communities. Mentors play a critical role: Over 90% of teams who successfully completed the program had a mentor, and 70% of the Technovation girls who completed the post survey shared that they were helped a lot by their mentors. These statistics are summarized in Figure 3

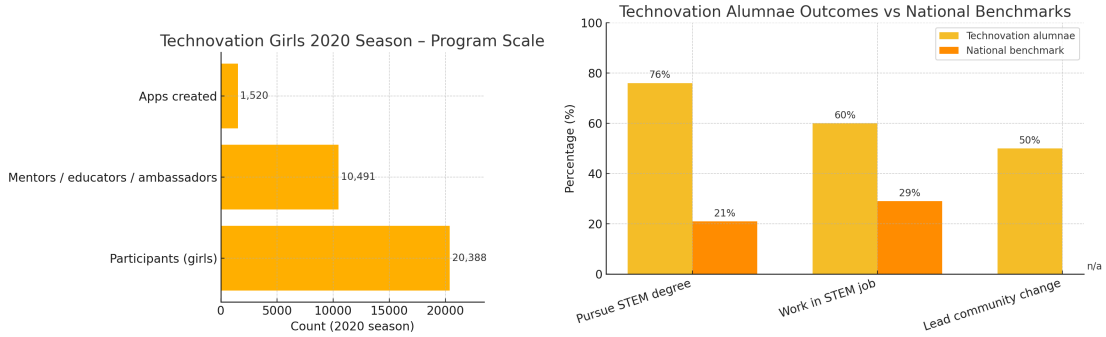


Figure 3: **Left:** Reach of the Technovation Girls 2020 season. **Right:** Long-term impact of Technovation participation.

Another program of Technovation engages families Chklovski et al. (2021). Over the course of 2 years it has engaged nearly 20,000 under-resourced $3^{rd} - 8^{th}$ grade students, parents and educators from 13 countries in a multi-week AI competition. Families worked together with the help of educators to identify meaningful problems in their communities and developed AI-prototypes to address them. The resulting projects (prototypes) were judged by a panel, using well-defined criteria. The program identified a high level of interest in underserved communities to develop and apply AI-literacy skills. Students followed 10 classes, and were tested their understanding of concepts through selected response questions on the curriculum platform. If they selected the wrong answer, they were prompted to try again. The competition element of the program provided the usual combination of pros and cons: time-based deadline that motivated families to persist and submit their prototypes, excitement of competing at a global level counterbalanced by stress, frustration, impatience, and forced deliberation. Strategies to improve retention include providing a variety of project-based learning lessons, starting from hands-on, unplugged activities and then moving onto software projects. Patience and commitment are needed: It takes 3 to 5 years to iteratively develop fun, engaging, effective curriculum, training and scalable program delivery methods. This level of patience and commitment is needed from all community and industry partners and funders.

7 Case study 2: Short-Competitions and Hackathons in the classroom

Short-term Competitions (over weeks) and Hackathons (over days) have been used in several contexts: as part of classes, student-organized events, conference tutorials, and summer schools. We provide a few hands-one experiences in this section.

Hackathons and competitions are used directly as part of **class activities**. RPI regularly uses two short-term challenges as assignments in its Intro to Data Mathematics (IDM) course. The challenges use project-based learning Anazifa and Djukri (2017) to encourage students to become creative data analysts by asking their own questions and developing their own strategies for addressing the challenge questions. This project-based learning strategy affords students the opportunity to develop DA knowledge and skills in context. We (Kristin Bennett, Adrien Pavao, and Isabelle Guyon) created two online challenges. In the first, To be or not to be: ICU Mortality, students predict whether an intensive care unit (ICU) patient lives or dies. ICU Mortality is a straight forward

classification competition. Students do the challenge in teams as a two week assignment to practice their classification skills. They learn how to enter a challenge in preparation for the second more complex challenge and get practice organizing and presenting team results. The second challenge, Chems-R-Us, serves as a three-week final project. Students perform both a classification task (which compounds are readily-biodegradable or not), and feature selection task. Students work in teams as if they are consultants hired by a hypothetical company Chems-R-Us. The lab times are spent coaching the student teams instead of traditional lectures. As a team, they are required to explore a diverse set of classification and feature selection approaches and formally present their final results and recommendations in a mini-conference.

The use of the two competitions based on real-world problem has enabled the class to incorporate compelling project-based learning experiences to a large class of 60 to 90 students. We have been reusing copies of the same competitions for over five years, but students come up with new approaches and outcomes every year. Team work and creative exploration is enhanced by having two separate tasks in one challenge and by picked the winning team for each task by the team median score. The gamification aspect encourages the many students to deeply engage with the problems and try many machine learning approaches. We vary the usage of the competitions each semester slightly by asking questions that cause them to go deeper into the analysis. Students are also asked to prepare a final report which describes the methods and results as well as their own creative analyses and visualizations to address these questions. We find the students are well prepared for the following course-based undergraduate research experiences described in the next case study 8.

Another type of hackathon taking place around the clock for 24 hours is the Data Science Game, a **world-wide competition organized by students and for students**. One of our students (Benjamin Donnot) was in the organizing team, and one of us (Isabelle Guyon) was a coach and judge. This competition is open to students from first year of master up until last year of PhD. It aims both at promoting machine learning and evaluating the level of students from the best data science programs worldwide. The event took place yearly from 2015 to 2018. A first pre-selection round is run online. The best teams are then invited on-site in a castle to compete non-stop for 24 hours. Mentors from both academia and industry are invited to coach the students. Many sponsors provide prizes and travel awards. For example, the first edition gathered 143 teams from more than 50 different prestigious universities (such as Stanford, National University of Singapore, Columbia University, Cambridge University, Franch Ecole Polytechnique, and Moscow State University), from 28 different countries from 5 different continents. The level was very high as some top ranked kagglers were participating to the event. It has been supported by various sponsors such as: Google, Microsoft, AXA, CapGemini but also some other institutions such as ChaLearn or Etalab.

Hackathons also lend themselves to be organized **in conjunction with beginner tutorials at machine learning conferences**. Université Paris-Saclay co-organized a hackathon on Malaria microscopy analysis in conjunction with Data Science Africa 2019, which attracted 254 participants. One of our students (Herilalaina Rakotoarison) made a one hour presentation: 5 min (Codalab presentation), 15 min (how to create a challenge) and 40 min (mini hackathon). See slides. The 40 min hackathon helped the students learn to run the Jupyter notebook until the end. The students could then continue making entries until the end of the conference. One difficulty was that the internet connection was not very good. The winners received support to go to the next conference. All the participants were then encouraged to enter a more complex version of the challenge.

Finally, ChaLearn also held a hackathon as part of a **summer school**, the Microsoft Machine Learning and Intelligence School, Saint Petersburg, Russia, July 29 - August 5, 2015. One of our

students (Arthur Pesah) was part of the organizing team. The school was sponsored by Microsoft Research and Yandex and is organized in cooperation with Lomonosov Moscow State University (MSU). It offered advanced undergraduates, PhD students, and young scientists and developers an opportunity to learn about the latest research in machine learning, intelligence and data science from top scientists. We offered a simplified version of the AutoML challenge, which helped students enter that competitions. Prizes were awarded, not only to the winners, but also to the team, which made the best presentation. The students were encouraged to continue on working on the larger international competition.

8 Case study 3: Course-based undergraduate research based on high-stake competition

Competitions can be a very effective way for educators to provide Course-Based Undergraduate Research Experiences (CUREs). CUREs are a pedagogical approach in which students engage in original research with unknown outcomes as part of a regular course. They offer a more inclusive entry point into research for undergraduates with proven benefits for students outcomes Bangera and Brownell (2014).

Students of Kristin Bennett at RPI participated in a high stake competition: the AHRQ Visualization of Community-Level Social Determinants of Health Challenge as research projects in two CUREs. This Challenge invited participants to develop new online tools to present social determinants of health data. The goal was ultimately to improve population health outcomes, and drive savings. The Challenge was structured in two phases. In Phase 1, which launched in March 2019, participants submitted concept abstracts and prototype designs of data visualization methods. For Phase 1, 12 semifinalists received $10,000 each based on the merits of their proposals and moved to Phase 2. The RPI students qualified. In Phase 2, semifinalists developed proofs-of-concept to be judged by the expert panel. One grand prize winner from Phase 2 won $50,000; second place was $35,000; third place won $15,000. The RPI students won third place, with Mortality Minder Bhanot et al. (2024). Mortality Minder explores mortality trends for midlife adults ages 25–64 across the United States from 2000-2017. Users can identify social and economic factors associated with increased mortality trends at the county level for the Nation and individual States. Visualizations demonstrate determinants and their impact on mortality trends.

A team of 23 RPI students developed the MortalityMinder datasets and analytics for the competition in the Health Analytics Challenges Lab course in the summer 2019 semester. A team of 22 students prepared the final Mortality Minder entry in Fall 2019. They designed Mortality Minder as an interactive, web-based dashboard that enables healthcare researchers, providers, payers, and policy makers to gain actionable insights into how, where, and why midlife mortality rates are rising in the United States. Students were advised by health industry professionals from United Health Foundation, Continuum Health, CDPHP, and NYSDOH as well as RPI professors and scientists. Students presented and interacted with advisers many times over the summer; group projects representing aspects of the Mortality Minder project culminated with poster presentations during a term-ending Mini Conference with guests invited. The student targeted a variety of issues ranging from data transformation, analytics, and interactive visualization through system design, user interface design, usability studies, and user documentation. The large number of students involved in the Mortality Minder project and the “production" nature of the coding effort was a unique chal-

lenge for the research advisers but also enhanced the student experience. The live Mortality Minder application is publicly available and the code public code repository is open-sourced.

We briefly summarize the work of the students. Mortality Minder illustrates midlife mortality rate increases reported in Woolf and Schoomaker (2019), while providing greater insight into community-level variations and their associated factors to help determine remedies. Using authoritative data from the CDC and other sources, Mortality Minder is designed to help health policy decision makers in the public and private sectors identify and address unmet healthcare needs, healthcare costs, and healthcare utilization. Innovative analysis divides counties into risk groups for visualization and correlation analysis using K-Means clustering and Kendall correlation. For each selected State and Cause of Death, Mortality Minder dynamically creates three analysis and visualization infographics, presented as pages in the app: “National View" reveals midlife mortality rates through time and compares state and national trends; “State View" categorizes counties into risk groups based on their midlife mortality rates over time. The app determines correlations of factors to risk groups and visualizes the most significant protective and destructive factors; “Factor View" enables users to explore individual factors including their relation to the selected cause at a county level for each state and the distribution of those factors within each state. Mortality Minder also allows users to perform a nationwide analysis.

9 Case study 4: University student organized challenges

We present an example of challenge organization class, which ended up as an accepted NeurIPS’22 competition: Cross-domain MetaDL Carrión-Ojeda et al. (2022).

As previously mentioned, we (Isabelle Guyon and Adrien Pavao) taught a class on challenge organization at the master level at Université Paris-Saclay (Section 4 and 5). The class, Creation of AI Challenges, has the objective of learning to create mini student challenges, which are then solved by other students, see a list of past challenges organized by students, as part of their class requirements. To raise awareness on issues related to ethics, privacy, and fairness, this class is taught in conjunction with a class on Fairness in Artificial Intelligence.

In 2021/2022, the fairness class was new, and it attracted a lot of interest. Many students chose to address problems of fairness and/or bias in their challenge. Due to the sensitive nature of data involving human subjects, some groups decided to study bias with data involving object recognition data, as a metaphor for societal bias. For example, using the background of an image as adjunct information to recognize an object can be used as a metaphor for using the dressing style of a person to evaluate their technical skills.

The challenge class is spread over 7 weeks in January and February, with 3 hours of class per week, including 1 hour of lecture, 1/2 hour of practical tips, and 1.5 hour of practical work. The students are also expected to work 1/2 day on their own each week. The master program also includes several project and internship requirements: one 60 hour academic project to be carried out in a research lab at the university, and one two to four month internship, either in a lab or in industry. Several students chose to do either or both of their project and internship on a subject related to challenge organization. This allowed us to involve them in the organization of an international challenge, leveraging the effort they put into their challenge class work.

The student work was incorporated in the creation of a large meta-dataset, called Meta-Album, consisting of 40 re-purposed public datasets, obtained from various source. All problems are image classification problems, where images are reduced to 128×128 pixels. Meta-Album datasets

were used in the Cross-domain MetaDL challenge, co-organized by one of the students (Dustin Carrión), and accepted in April 2022 in the NeurIPS'22 conference competition program (after a very selective review process).

In November and December, five students decided to work on creating datasets in Meta-Album format, as their project, knowing that such datasets would be used in their student challenge, and would possibly be incorporated in Meta-Album and the Cross-domain MetaDL challenge. They were tutored by a second year master student (Ihsan Ullah), who took the challenge class the year before, and was the leader of the Meta-Album effort. The students' work included to hunt for suitable datasets (image datasets with sufficient resolution to be reduced to 128×128 pixels, with at least 20 classes and at least 40 examples for class for these classes, and belonging to a certain chosen domain). Then they has to format the data, document the datasets, and run baseline methods. Part of their work was also to scrutinize data to identify possible biases. For example, some datasets included images that were extracted from videos, hence were not independent of one another. Other spurious dependencies between images included images cut out from the same mother image (satellite picture of microscope slice). This such images would have correlated color spectra (due to illumination or staining).

In January and February, the students created their student challenges, which are posted on this page. All challenge protocols were in the AutoML setting: code was submitted and blindly evaluated for training and testing on an unknown dataset (a different one in each phase). Sample data and a starting kit were provided to help participants make an entry. One of the student challenges, TrustAI, was not on image classification. They addressed the problem of bias in machine learning models against groups in society. Inspired by the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) biased software, this challenge deals with a classification problem that is adjudicating on the suspect based on criminal activities. The goal is to avoid relying on protected variables such as age, gender, or race. The other two challenges were on image classification. The PACHAMAMA challenge proposed to classify living species. To purpose was to help quantifying biodiversity and monitor population changes, for a better conservation of living organisms. Problems of bias, particularly due the image background, were part of the challenge. The PANACEA challenge proposed a histology classification problem. The goal was to help histologist and histopathologist make better diagnoses, using images of microscope slices. This problem is important to automate the process of medical analyses and help study problems of bias due to differences in staining, sample contamination, lighting, etc.

In March and April, other students chose to solve these challenges as their projects, under the guidance of Adrien Pavão.

Simultaneously, three students chose to combine their project and internship into a 6-month internship to continue working on the preparation of international challenges using Meta-Album Ullah et al. (2022). Dustin Carrión joined the team preparing the Cross-domain MetaDL challenge, took leadership, and decided to submit a proposal to NeurIPS'22 (which ended up being accepted). He went on to write a paper on the design of the challenge and baseline results, which was accepted to a meta-learning workshop at ECML/PKDD 2022 Carrión-Ojeda et al. (2022). The paper describes the design and baseline results for the challenge. Meta-learning aims to leverage experience gained from previous tasks to solve new tasks efficiently (i.e. with better performance, little training data and/or modest computational resources). While previous challenges in the series focused on *within-domain* few-shot learning problems, with the aim of learning efficiently *N-way k-shot* tasks (i.e. N class classification problems with k training examples), this competition challenges the participants

to solve “any-way” and “any-shot” problems drawn from various domains (healthcare, ecology, biology, manufacturing, and others) chosen for their humanitarian and societal impact. It is based on a subset of Meta-Album Ullah et al. (2022), a meta-dataset of 40 image classification datasets from 10 domains, from which we carve out tasks with any number of “ways” (within the range 2-20) and any number of “shots” (within the range 1-20). The competition is with code submission, fully blind-tested on the CodaLab challenge platform. The code of the winners will be open-sourced, enabling to deploy automated machine learning solutions for few-shot image classification across several domains.

The two other students (Gabriel Lauzzana and Romain Mussard) set on working on the preparation of a challenge on bias. Their design was submitted for publication in a junior conference Lauzzana et al. (2022). The focused on datasets, not involving human subjects, but plagued with various kinds of bias, the origin of which is not always known, and which may include confounding bias and sampling bias. They propsed to unravel causes of bias, followed by rigorous manual data curation. With the advent of fully automated machine learning (AutoML), one may wonder whether creating bias-robust learning machines is possible, to reduce the need for data curation and the possible risk of introducing further biases. In this context, they designed a bias-aware AutoML challenge, based on image classification tasks. We presented in the paper the challenge design, data preparation, and baseline methods. For reproducibility their code is provided.

In conclusion, the student competition design effort was very fruitful and resulted in the preparation of several international competitions.

10 Case study 5: Hackathons in industry - Continuous learning and upskilling

Hackathons represent an exciting way in industry to learn about newer technologies being used in the enterprise and use them to solve customer scenarios or apply acquired knowledge to new domains, or yet to remain relevant, be at the forefront of change and sometimes help advance one’s career.

As early as 2010 when ML was mature enough to being used in industry, hackathons helped bring experts and novices in industry to together learn and solve problems. Hackathons can take the form of:

- A free form track where people could work on any problem of their choice as long as they were articulate about the customer problem/scenario and that the proposed solution used AI/ML; the deliverable had to be a prototype or demo with the winners determined by a jury.
- A contest track where business groups would pitch problems to be solved and would selected the “winner” with specific goals such as ’achieving highest percentage accuracy using any tool; with the “Winners” being determined by an automated script reflected on a leaderboard.

Hackathons are today at the center stage of the culture of innovation needed to transform industries, businesses or organizations to be empowered by AI. Hackathons are becoming more than a one off tool to learn some new technologies; they are part of a larger framework within an innovation lifecycle from explorations to learning to validating ideas and building prototypes ready for scale (see `https://www.microsoft.com/en-us/garage/`).

Google DeepMind also organizes yearly hackathons, which used tom be called “DeepMind extravaganza" (now GDM-create). It is a two-week annual event focused on cross-team collabora-

tion, learning, and connection. During that time, team members have the opportunity to engage in exploratory projects that they propose or are proposed by others. It fosters a continuous learning culture within DeepMind.

11 Conclusion

In this chapter, we have shown that competitions and challenges offer a powerful pedagogical approach across all stages of learning and professional development. From the earliest levels of education, where activities often take the form of simple, juried projects, to university courses leveraging automated scoring on complex research problems, the competition format promotes engagement, self-directed learning, and creative problem-solving. Additionally, we have seen how competitions can serve as effective hands-on laboratories for teaching experimental design, data analysis, and application of cutting-edge techniques.

Beyond academia, competitions and challenges also support continuous learning in industry. They facilitate the learning of new methods, tools, and technologies, helping professionals maintain their relevance in a fast-evolving marketplace. By encouraging learners to embrace hands-on problem-solving and exploration, these activities help bridge the gap between theory and practice, and they can be adapted to a wide range of domains and skill levels.

In sum, competitions and challenges provide an adaptable and practical way to motivate learners. As educational and professional environments continue to change, their value as both a teaching strategy and a continuous learning mechanism will likely only increase.

Acknowledgements

We thank Tara Chklovski for her invaluable contributions.

References

Risqa D Anazifa and Djukri Djukri. Project-based learning and problem-based learning: Are they effective to improve student's thinking skills? *Jurnal Pendidikan IPA Indonesia*, 6(2):346–355, 2017.

Albert Bandura. *Self-Efficacy: The Exercise of Control*. W. H. Freeman, 1997.

Gita Bangera and Sara E Brownell. Course-based undergraduate research experiences can make scientific research more inclusive. *CBE—Life Sciences Education*, 13(4):602–606, 2014.

Kristin P Bennett, John S Erickson, Amy Svirsky, and Josephine C Seddon. A mathematics pipeline to student success in data analytics through course-based undergraduate research. *The Mathematics Enthusiast*, 19(3):730–750, 2022.

Karan Bhanot, John S Erickson, and Kristin P Bennett. Mortalityminder: Visualization and ai interpretations of social determinants of premature mortality in the united states. *Information*, 15 (5):254, 2024.

Paolo Calafiura, Steven Farrell, Heather M. Gray, Jean-Roch Vlimant, Vincenzo Innocente, Andreas Salzburger, Sabrina Amrouche, Tobias Golling, Moritz Kiehn, Victor Estrade, Cécile Germain,

Isabelle Guyon, Ed Moyse, David Rousseau, Yetkin Yilmaz, Vladimir Vava Gligorov, Mikhail Hushchyn, and Andrey Ustyuzhanin. Trackml: A high energy physics particle tracking challenge. In *14th IEEE International Conference on e-Science, e-Science 2018, Amsterdam, The Netherlands, October 29 - November 1, 2018*, page 344. IEEE Computer Society, 2018. doi: 10.1109/eScience.2018.00088. URL https://doi.org/10.1109/eScience.2018.00088.

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. *CoRR*, abs/2208.14686, 2022. doi: 10.48550/arXiv.2208.14686. URL https://doi.org/10.48550/arXiv.2208.14686.

Dustin Carrión-Ojeda, Hong Chen, Adrian El Baz, Sergio Escalera, Chaoyu Guan, Isabelle Guyon, Ihsan Ullah, Xin Wang, and Wenwu Zhu. In *Meta-Knowledge Transfer/Communication in Different Systems*, ECML/PKDD 2022 workshop, 2022.

Tara Chklovski, Richard Jung, Rebecca Anderson, and Kathryn Young. Comparing 2 years of empowering families to solve real-world problems with ai. *KI-Künstliche Intelligenz*, 35(2):207–219, 2021.

Anthony Goldbloom and Ben Hamner. Kaggle. 2010. URL https://www.kaggle.com/docs/competitions.

Suzanne Hidi and K. Ann Renninger. The four-phase model of interest development. *Educational Psychologist*, 41:111 – 127, 2006.

Robyn Hromek and Sue Roffey. Promoting social and emotional learning with games: "it's fun and we learn things". *Simulation & gaming*, 40(5):626–644, 2009.

Dirk Ifenthaler, Deniz Eseryel, and Xun Ge. Assessment for game-based learning. In *Assessment in game-based learning*, pages 1–8. Springer, 2012.

Balázs Kégl, Alexandre Boucaud, Mehdi Cherti, Akin Osman Kazakci, Alexandre Gramfort, Guillaume M Lemaitre, Joris Van den Bossche, Djalel Benbouzid, and Camille Marini. The RAMP framework: from reproducibility to transparency in the design and optimization of scientific workflows. International Conference On Machine Learning, July 2018. URL https://hal.archives-ouvertes.fr/hal-02072341. Poster.

Gabriel Lauzzana, Romain Mussard, Ihsan Ullah, and Isabelle Guyon. Design of a bias-aware automl challenge. In *JDSE*, In review, 2022.

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*, page 17, 2020.

Antoine Marot, Benjamin Donnot, Camilo Romero, Luca Veyrin-Forrer, Marvin Lerousseau, Balthazar Donon, and Isabelle Guyon. Learning to run a power network challenge for training

topology controllers. *CoRR*, abs/1912.04211, 2019. URL http://arxiv.org/abs/1912.04211.

Antoine Marot, Benjamin Donnot, Gabriel Dulac-Arnold, Adrian Kelly, Aïdan O'Sullivan, Jan Viebahn, Mariette Awad, Isabelle Guyon, Patrick Panciatici, and Camilo Romero. Learning to run a power network challenge: a retrospective analysis. *CoRR*, abs/2103.03104, 2021. URL https://arxiv.org/abs/2103.03104.

Hoa P. Nguyen. How to use gameplay to enhance classroom learning. *Edutopia*, 2021. URL https://www.edutopia.org/article/how-use-gameplay-enhance-classroom-learning.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Adrien Pavao, Diviyan Kalainathan, Lisheng Sun-Hosoya, Kristen Bennett, and Isabelle Guyon. Design and analysis of experiments: A challenge approach in teaching. In *NeurIPS 2019-33th Annual Conference on Neural Information Processing Systems*, 2019.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. Codalab competitions: An open source platform to organize scientific challenges. *Technical report*, 2022. URL https://hal.inria.fr/hal-03629462v1.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jan L Plass, Bruce D Homer, and Charles K Kinzer. Foundations of game-based learning. *Educational psychologist*, 50(4):258–283, 2015.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

Ihsan Ullah, Dustin Carrion, Sergio Escalera, Isabelle Guyon, Mike Huisman, Felix Mohr, Jan N. van Rijn, Haozhe Sun, Joaquin Vanschoren, and Phan Anh Vu. Meta-album: Multi-domain meta-dataset for few-shot image classification. In *Submitted to: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2022. URL https://meta-album.github.io/.

Steven H Woolf and Heidi Schoomaker. Life expectancy and mortality rates in the united states, 1959-2017. *Jama*, 322(20):1996–2016, 2019.

W-H Wu, H-C Hsiao, P-L Wu, C-H Lin, and S-H Huang. Investigating the learning-theory foundations of game-based learning: a meta-analysis. *Journal of Computer Assisted Learning*, 28(3): 265–279, 2012.

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P Bennett. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416:244–255, 2020.

Placeholder PDF

Competition platforms

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

Harald Carlens HARALD@MLCONTESTS.COM
ML Contests

Reviewed on OpenReview:

Abstract

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

Keywords: competition platform, challenge hosting services, service comparison

1 Platforms for AI contests

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2 Platform comparison criteria

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table 1: Platform overview

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

3 Platform Comparison

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4 Non-English language platforms

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

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

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

5 Domain-specific platforms

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

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

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

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

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

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

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

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

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

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

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

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

6 Alternative approaches and adjacent services

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

7 Independently hosted contests

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

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

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

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

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

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

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

8 Choosing the right platform

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

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

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

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

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

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

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

9 Conclusion

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

Acknowledgments and Disclosure of Funding

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hands-on tutorial on how to create your own challenge or benchmark

Adrien Pavão ADRIEN.PAVAO@GMAIL.COM
LISN, CNRS
Université Paris-Saclay
France

Reviewed on OpenReview: *No*

Abstract

Organizing a challenge allows you to crowd-source the most difficult machine learning problems. It is also an excellent way to learn data science. By following this short hands-on tutorial, you can create your first competition or benchmark — as early as today! In this chapter, we give you everything you need to implement, concretely, your own online competition or benchmark. We do not address other practical issues such as finding sponsors or communicating about the event; this is discussed in chapter 13.

Keywords: tutorial, CodaLab, Codabench

1 Introduction

In this chapter, you will learn how to organize, a challenge or a benchmark on the two platforms *CodaLab Competitions* and *Codabench* (Figure 1). This tutorial is divided into three parts: we first review the aspects shared by both platforms (Section 2), then the *CodaLab Competitions* platform specificity (Section 3), and finally the *Codabench* platform specificity (Section 4).

CodaLab Competitions Pavao et al. (2022) is an open-source web platform hosting data science and machine learning competitions. This means that you can set up your own instance of it, or use the main instance on codalab.lisn.fr. *CodaLab* puts an emphasis on science and each year hundreds of challenges are organized on it, pushing the limits in many areas: physics, medicine, computer vision, natural language processing or even machine learning itself. Its flexibility allows hosting challenges on a wide variety of tasks! The only limit is your imagination.

Codabench Xu et al. (2022) is another project, free and open-source as well, following the steps of *CodaLab Competitions*. It can be seen as an upgrade of it, using more recent technologies, and with an emphasis on benchmarks. This emphasis on benchmark is enforced by some features, such as the possibility of filling a leaderboard with a single user account. The public server is codabench.org.

These two platforms are well suited for a tutorial, given their flexibility and the fact that they are open-source and free to use. **Once you have an account, you can already publish your first competition or benchmark!**

Figure 1: Sources: codalab.lisn.fr, codabench.org

2 General aspects

Inside the competition bundle

To create a machine learning challenge or benchmark on these platforms, all you need to do is to upload a **competition bundle**. A competition bundle is a ZIP file containing all the pieces of your competition: the data, the documentation, the scoring program and the configuration settings. To customize your competition, you can simply change the files contained in the template bundle before uploading it. Note that every aspects of the competition (settings, data, etc.) can still be edited after the upload. Let's have a closer look at what's inside the bundle.

The competition.yaml file defines the **settings** of your challenge. The title, description, logo, dates, prizes, Docker image[1], leaderboard structure and so on. All possible settings are documented in the Wiki.

The documentation files, either *HTML* or *Markdown* files, define the various pages that participants can see when going to your competition. Use them to provide the documentation and the rules, as well as any information you find important. You can of course select your own **logo** for the competition by replacing the "logo.png" file.

Data. If you are designing a machine learning problem, it is likely that you have data. The **public data** folder is for the data fully accessible by participants, **input data** is accessible by participants' submitted code, and the **reference data** folder is for storing the ground truth information (e.g. the labels from the testing set) which is kept hidden to the participants, making it only accessible by the scoring program, as explained later in this section. You can either use a data format provided in a competition example or a new one that fits well with your competition. To ensure compatibility, you will need to update the scoring program — we will talk about it in the next section. *If your problem does not involve data, don't worry! CodaLab is flexible and allows you to define any kind of problem (e.g. reinforcement learning tasks).*

The ingestion and scoring programs are the critical pieces of your competition bundle since they define the way submissions will be executed and evaluated respectively. If you want to allow only **result submissions**, then you only need the scoring program; the ingestion program is useful for **code submissions**. Figure 2 shows the interactions between the submissions (results submissions or code submissions), the programs and the leaderboard.

- The **ingestion program** defines how to train the models and save their predictions.
- The **scoring program** defines how to compare the predictions with the ground truth and computes a score.

[1]The Docker image is the environment in which all submissions will be run, allowing to precisely control the evaluation procedure. It can be referred by its DockerHub name.

These programs evaluating submissions can be customized by the organizers to adjust them to adapt to any competition protocol. While they are written in Python in the templates provided, they can be written in any programming language.

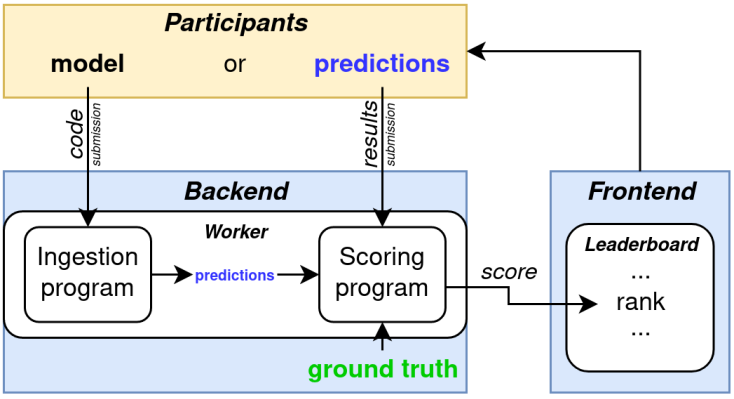


Figure 2: General competition workflow on CodaLab

The starting kit. If you have already joined a challenge as a competitor, you probably know how important it is to have a good starting kit. The goal of the starting kit is to provide participants with all the necessary resources to facilitate their dive into your competition, such as some example submissions, Jupyter notebooks, or any useful documentation and files. You can even provide the competition bundle itself (without the ground truth) inside the starting kit; this way, all the internal functioning will be perfectly transparent for the participants.

Queues and Docker

The public servers provide default compute workers. However, to run computationally demanding competitions, **organizers can create custom queues and attach their own CPU or GPU compute workers** (physical or virtual machines on any cloud service) to it. This modular architecture of *CodaLab Competitions* has been a key ingredient in growing its user base, without requiring that the institution hosting the main instance covers all computational costs. Another interesting aspect of this feature is that the training and testing of algorithms can be done on confidential data, without any leakage, by putting data directly inside the compute workers. This is especially useful for medical research, challenges organized by industries, and in other restricted domains.

The workflow of the jobs, showed in Figure 3, works as follows: each competition can be linked to only one queue, but the queue can be used by several competitions. Each worker can listen to only one queue, but the queue can be linked to several workers.

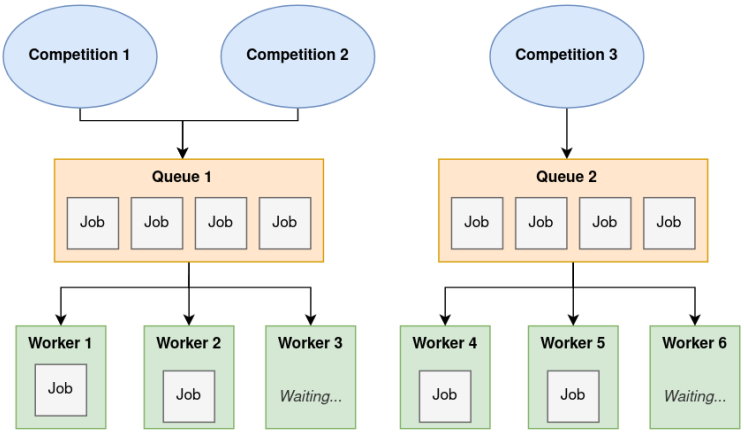


Figure 3: Diagram of the structure of workers and queues. The queues dispatch the jobs between the compute workers. Note that a queue can receive jobs (submissions) from several competitions, and can send them to several compute workers.

To setup a machine as a compute worker, you need to install Docker, note down the URL of your queue, and run a single command line[2][3].

The number of workers can be adjusted at anytime. This means that you can add more workers to the queue during the competition, they will dispatch all jobs automatically, increasing your computing power in real time in order to fit the needs of your competition.

One aspect that allows the same machine to compute jobs from many different competitions is the use of **Docker environments**. The execution of participants code and scoring are performed inside a container, which prevents to damage the servers and allows organizers to define a custom controlled environment for their competitions. Organizers can create a fully customized environment, with allowed libraries and programming languages for their participants' submissions, and simply link it to a competition by providing a DockerHub name and tag. This means that every candidate is judged in the same way, the competition does not get deprecated after some time and adding new libraries or updating the experimental environment is straightforward and transparent.

Organizer features

As a competition or benchmark creator, you have access to useful organizer-only features. These features are accessible from the grey buttons at the top of the user interface, as shown in Figure 4.

[2] `https://github.com/codalab/codalab-competitions/wiki/User_Using-your-own-compute-workers`

[3] `https://github.com/codalab/codabench/wiki/Compute-Worker-Management---Setup`

Figure 4: As an organizer, you have access to many interesting features, accessible from the grey buttons at the top of the interface of *CodaLab Competitions* (left) or *Codabench* (right).

Editor. The editor allows organizers to edit a competition that is already up and running. From this panel, you can edit every setting, and replace the data, scoring program or ingestion program.

Manage participants. You can choose to allow anybody to join your competition, or to have a registration process and validate who can join. You can accept or revoke access of the participants at any time.

Manage submissions. A panel to manage all the submissions made by the participants. From this panel, you can access details about the submissions: their date and author, the output and logs of the scoring programs and the submissions themselves (the files uploaded by the participants). The interface can be used to cancel, remove or re-run submissions. Overall, it is very useful to debug, to prevent cheating or to run post-challenge analysis.

Dumps: export your competition. A feature that you can use to download your competition as a bundle. All changes made directly through the editor will be saved and the resulting bundle can be re-uploaded on the platform or on any other instance of it. If you wish to re-upload the bundle on the same platform and want to keep the bundle as light as possible, you can use the option "use URI keys instead of files", in this case the datasets and programs will be referred by their address in the storage.

Publish your competition. By default, your competition is private. When a competition is private, it can be accessed only by its administrators, or by anyone from the "secret URL". Once you publish your competition, anyone can access it from the public URL. You can make your competition private again at any time.

Auto-migration. Automatically re-run the leaderboard's submissions from one phase to another phase.

3 CodaLab Competitions tutorial

Get started

To create your first machine learning challenge, all you need to do is to upload a **competition bundle**. A competition bundle is a ZIP file containing all the pieces of your competition: the data, the documentation, the scoring program and the configuration settings, as explained in Section 2. Let's start from an example; it's the easiest way. Here is the competition bundle of the *Iris Challenge*, based on the famous Fisher's dataset Fisher (1936): Iris Competition Bundle. Now, go to CodaLab, and upload the file named "iris_competition_bundle.zip" as shown in Figure 5. Go to "*My Competitions*", then "*Competitions I'm Running*", and finally "*Create Competition*". During the last step you will be redirected to the final menu where you can upload the competition

bundle. After uploading it, you will see the message of Figure 6 indicating that your competition is successfully created and ready to receive submissions.

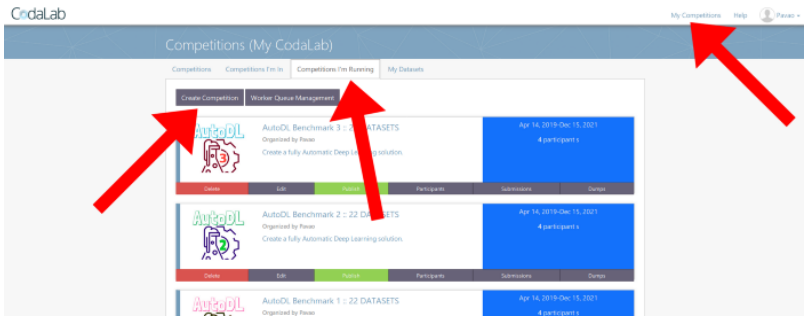


Figure 5: Go to "*My Competitions*", then "*Competitions I'm Running*" and finally "*Create Competition*".

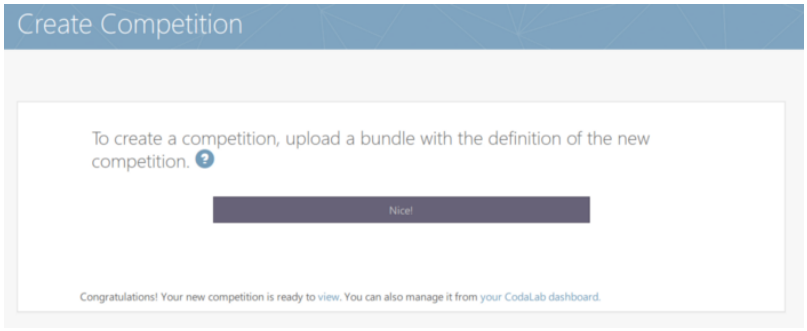


Figure 6: That's it! Your competition is ready to receive submissions.

Once the competition is uploaded, you can begin to make submissions. To do so, UnZip the downloaded bundle, go inside the starting kit and zip the content of either the "sample_result_submission" or the "sample_code_submission" folder. It is important to zip the files without directory structure and to include the "metadata" file in the case of code submission. Then, on the website (see Figure 7), go to the "*Participate*" tab, then "*Submit / View results*", click on "*Submit*" and select your zip file. The submission will process for a few moments. After that, you'll be able to see your score in the leaderboard, in the "*Results*" tab.

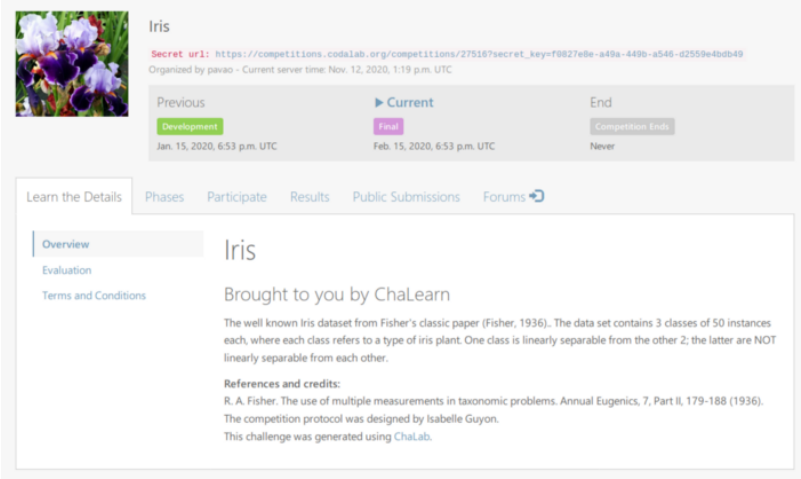


Figure 7: Main page of the Iris Challenge. The web pages are defined by the HTML files from the bundle.

Customize competition

Your competition is up and running! However, you wish to edit it. It is still possible. As an administrator of your own challenge, you have access to the "Edit" menu: a panel in which you can edit every setting. Here are some examples:

- Force submission to leaderboard: if enabled, the last submission of a participant is the one showed on the leaderboard.
- Disallow leaderboard modifying: if disabled, users can select which of their submissions appear on the leaderboard.
- Share administrator rights: you can add *CodaLab* users as administrators of your competition by giving their username or email address. They'll have access to all organizer-only features, except deleting the competition.
- Anonymous leaderboard: if enabled, the username are hidden in the leaderboard.
- Registration required: if enabled, users need to request the access to your competition. You then have to accept or reject their participation manually from the "Participants" tab.
- Select the queue.
- Specify competition docker image by its DockerHub name.
- For each phase you can chose different data and scoring programs.

If you wish to change the dataset or the scoring program, you'll first need to upload the new version from the "*My Datasets*" page, as shown in Figure 8. You will then be able to select it in the editor.

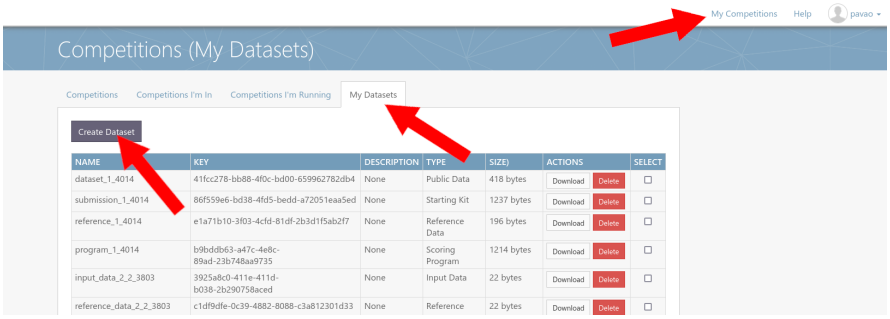


Figure 8: Go to "*My Datasets*" to upload new datasets or programs.

4 Codabench tutorial

Get started

As an evolution of *CodaLab Competitions*, *Codabench* is similar in terms of functioning and features. In this part of the tutorial, we will organize a simple benchmark on *Codabench*, in which the participants can have multiple entries displayed on the leaderboard. Note that competition bundles from *CodaLab Competitions* are compatible with *Codabench*, while not vice versa.

Let's have a look at the Mini-AutoML Bundle. To upload the file named "bundle.zip" to *Codabench*, go to "*Benchmarks*", then "*Management*" and finally "*Upload*" (Figure 9). Then click on the paper clip button to select and upload the bundle (Figure 10).

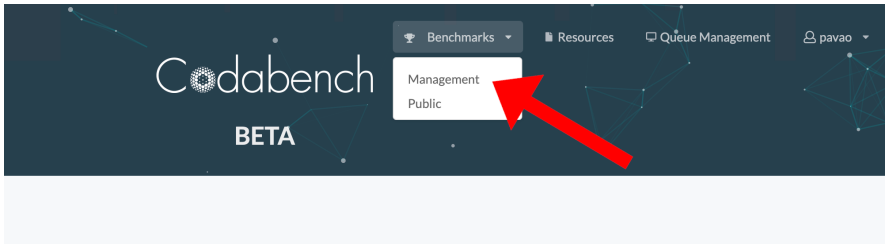


Figure 9: Go to "*Benchmarks*", then "*Management*" and finally "*Upload*".

Figure 10: Then click on the paper clip button to select and upload the bundle.

The main page of the benchmark should look like Figure 11.

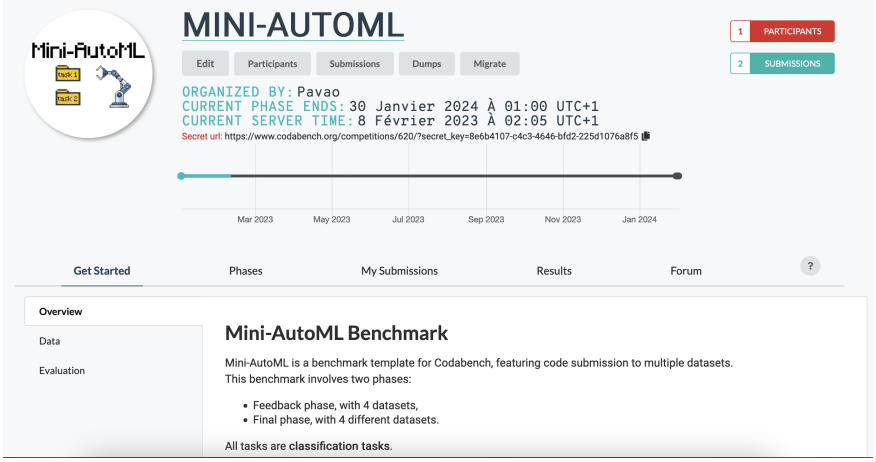


Figure 11: Overview of Mini-AutoML main page.

Let's now make a submission. To do so, download the file sample_code_submission.zip and upload it in the "*My Submissions*" tab of the benchmark. You will be able to view logs in real time during the processing of the submission. The leaderboard, available in the "*Results*" tab of the benchmark, will then be updated.

Customize benchmark

Let's customize this newly created benchmark, with the goal in mind to be able to fill in the leaderboard with many different models, in order to compare them.

Fact sheets. In the first tab of the editor, you can enable "fact sheets" to gather more information about the submissions. Enabling fact sheets means that the participants will be asked to fill in some information when making submissions. You can fully customize the information fields, making them required or not, and choosing which information appears on the leaderboard. This can be used to display the name of the methods, the URL to the source code, or a description of the method. The gathering of metadata about the methods used is crucial when conducting a benchmark, and this interface makes it simple to gather all this information in one place.

Edit data and programs. *Codabench* provides an interface to upload *Ingestion Program*, *Scoring Program*, *Starting Kit* and *Data (public data, input data and reference data)*. To upload a new dataset or program, go to "*Resources*", "*Datasets*" and click on "*Add Dataset*". Then name it, select your ZIP file, and chose the type (reference data, scoring program, etc.) (Figure 12).

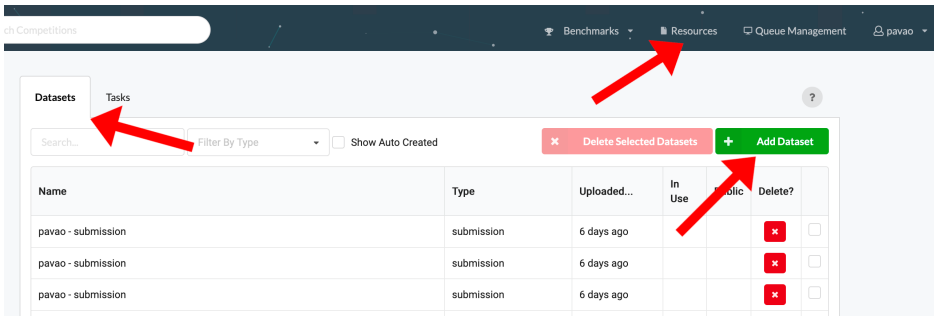


Figure 12: To upload a new dataset or program, go to "*Resources*", "*Datasets*" and click on "*Add Dataset*". Then name it, select your ZIP file, and chose the type (reference data, scoring program, etc.)

Once these are uploaded, a task can be created using the uploaded files. A task is a combination of *Ingestion Program*, *Scoring Program*, *Input Data*, and *Reference Data*. o create a task, go to "*Resources*", "*Tasks*" and click on "*Create Task*". Then name it and select the files previously uploaded: Input data, reference data, ingestion program and/or scoring program (Figure 13).

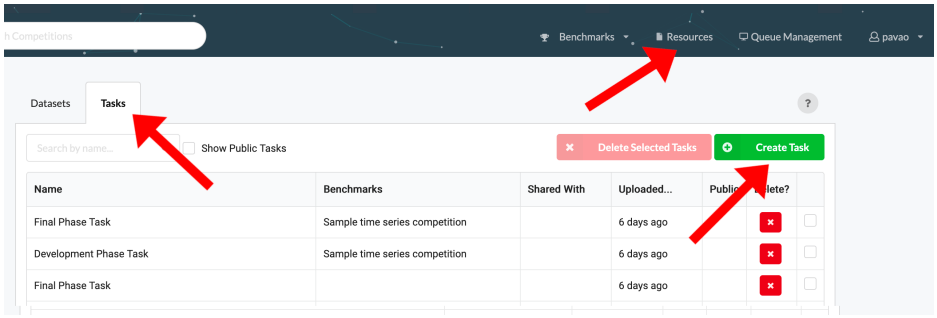


Figure 13: To create a task, go to "*Resources*", "*Tasks*" and click on "*Create Task*". Then name it and select the files previously uploaded: Input data, reference data, ingestion program and/or scoring program.

In your competition editor, you can add, update or delete a task for each phase. Unlike *CodaLab*, *Codabench* allows you to have multiple tasks for each phase. To associate a task to a phase, go to the editor, then "*Phases*", click on the edit button and select the desired task (Figures 14 and 15).

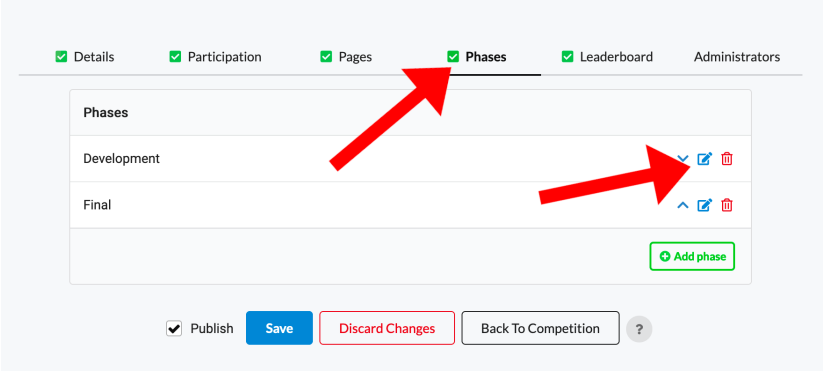


Figure 14: Go to the editor, then "*Phases*" and click on the edit button.

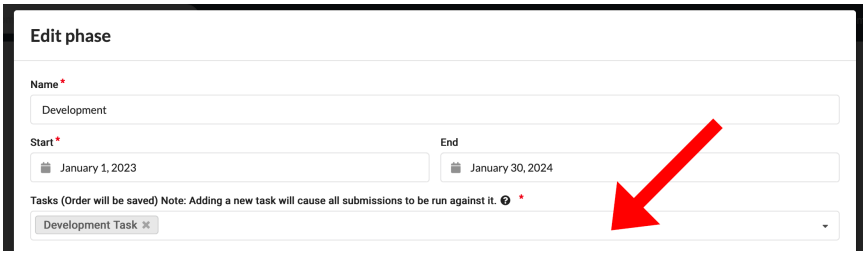


Figure 15: You can now associate the desired task to the phase.

Submission rules. The submission rule programs the behavior of the leaderboard regarding new submissions. Submissions can be forced to the leaderboard or manually selected, can be unique or multiple on the leaderboard, etc. To edit the submission rule, go to the editor, then "Leaderboard" and edit the leaderboard (Figure 16). Then change the submission rule from "Force Last" to "Add And Delete Multiple" (Figure 17). "Force Last" means that only the last submission of each participant will be shown on the leaderboard, while "Add And Delete Multiple" means that the participants will be allowed to manually select multiple submissions to show on the leaderboard.

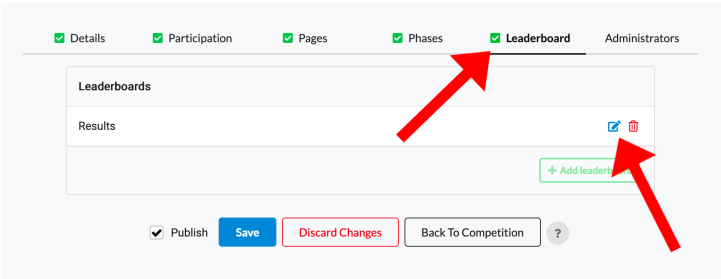


Figure 16: To edit the submission rule, go to the editor, then "Leaderboard" and edit the leaderboard.

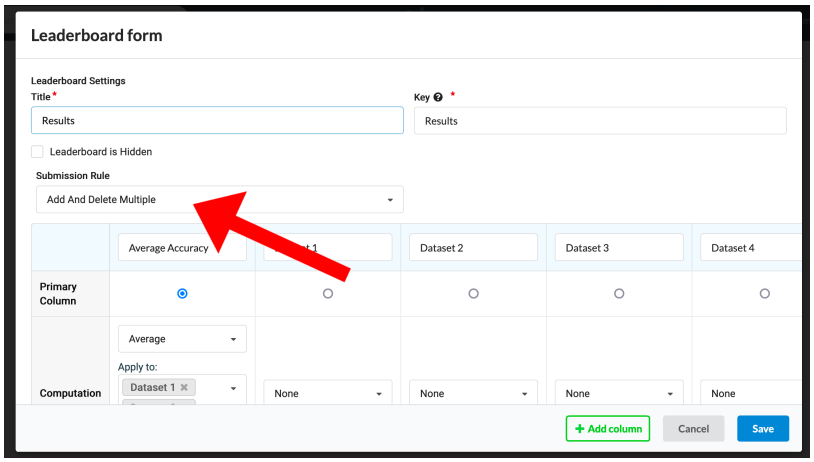


Figure 17: Change the submission rule from “Force Last” to “Add And Delete Multiple”.

The current submission rule, “Force Last”, means that only the last submission of each participant will appear on the leaderboard. This is a classical setting for competition. Changing this rule to “Add And Delete Multiple” will allow the participants to manually select which submissions will appear on the leaderboard, and multiple submissions per participant on the leaderboard are allowed.

Add submissions to the leaderboard. Now that the leaderboard is set up, let’s submit different variations of the code of the model from the sample_code_submission.zip. This example code submission simply calls a classifier from Scikit-Learn Pedregosa et al. (2011). Replace the *DecisionTreeClassifier* with the classifier of your choice. Remember to differentiate the different submissions by filling the “Method name” in the fact sheet. Once your submission is processed, click on the leaderboard button under “Actions” in the submissions table to manually add them to the leaderboard (Figure 18).

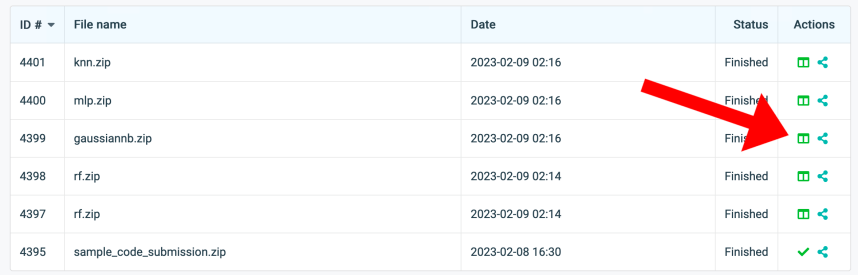


Figure 18: Once your submission is processed, click on the leaderboard button under “*Actions*” in the submissions table to manually add them to the leaderboard.

The leaderboard finally looks like Figure 19.

Results

Task:		Fact Sheet Answers	Development Task					
#	Participant	Method name	Average Accuracy	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Duration
1	Pavao	Random Forest	0.8974364460	0.8551832389	0.7678958785	0.9666666667	1.0000000000	2.9789698124
2	Pavao	K-Nearest Neighbors	0.8711441508	0.7730840101	0.7592190889	0.9555555556	0.9967179487	1.4459712505
3	Pavao	MultiLayer Perceptron	0.8709400671	0.7963898178	0.6984815618	0.9888888889	1.0000000000	3.9009172916
4	Pavao	Decision Tree Baseline	0.8622140842	0.8106189859	0.6832971800	0.9555555556	0.9993846154	0.2303986549
5	Pavao	Gaussian NB	0.8496961070	0.7948201733	0.7678958785	0.9222222222	0.9138461538	0.1322979927

Figure 19: Screenshot of the filled up leaderboard. The random forest classifier did the best job in the first phase of the benchmark!

5 Conclusion

Congratulations! You have learned the basics of *CodaLab Competitions* and *Codabench*, and now you can organize your own competitions or benchmarks! However, we barely scratched the surface of all the possibilities offered by these platforms. To learn more, you can refer to: CodaLab Competitions Documentation and *Codabench* Documentation.

From the documentation, you will learn how to link your personal compute workers (CPU, GPU), how to customize the ingestion and scoring programs, how to define complex leaderboards with multiple criteria, or even how to deploy your own instance of the platform. You can also join the effort and develop your own features!

References

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annual Eugenics*, 7(Part II):179–188, 1936. The competition protocol was designed by Isabelle Guyon. This challenge was generated using ChaLab for Codalab v1.5.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. Codalab competitions: An open source platform to organize scientific challenges. *Technical report*, 2022. URL `https://hal.inria.fr/hal-03629462v1`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Zhen Xu, Sergio Escalera, Adrien Pavao, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543, 2022. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.

2022.100543. URL https://www.sciencedirect.com/science/article/pii/S2666389922001465.

Special designs and competition protocols

Wei-Wei Tu TUWEIWEI@4PARADIGM.COM
The 4th Paradigm
China

Adrien Pavão ADRIEN.PAVAO@GMAIL.COM
LISN, CNRS
Université Paris-Saclay
France

Reviewed on OpenReview: *No*

Abstract

With the development of AI technology, many novel machine learning frameworks have been raised and applied in AI academic and industry research and business application. Organizing competitions in these areas can greatly help the research and development of related algorithms and technology. In this chapter, we explore the design of competitions in various kinds of machine learning field: supervised learning, automated machine learning, metalearning, time series analysis, reinforcement learning, adversarial learning, and using confidential data. For each of these specific competition protocol, we discuss the framework and design of the competition process. We believe this chapter can make great help to both the organizers and the participants, therefore accelerate the development of AI industry and research.

Keywords: competition, design, supervised learning, automated machine learning, metalearning, time series analysis, reinforcement learning, adversarial learning, confidential data

1 Introduction

Machine learning is an expansive field offering a rich diversity of algorithms, each developed to solve specific tasks. These algorithms are commonly grouped into three main categories: supervised learning, unsupervised learning, and reinforcement learning algorithms. Beyond the algorithms themselves, the possibilities are further augmented by the diversity of data and domains of applications. Depending on the nature of the data, its source, shape, quantity and patterns, different approaches are required. The applications of machine learning are virtually limitless, covering medicine, physics, natural language processing, economics, and more. To be able to capture this complexity and diversity in competitions and benchmarks, innovative experimental design is required.

In this chapter, we analyse the features and special designs about the challenges and benchmarks of supervised learning (Section 2), automated machine learning (Section 3), metalearning (Section 4), time series analysis (Section 5), reinforcement learning (Section 6), adversarial learning (Section 7), and using confidential data (Section 8). We also give some tips about how to perform well in these competitions as participants.

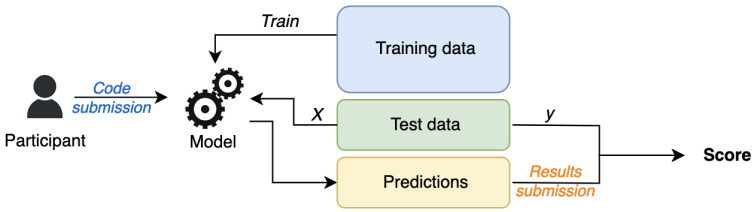


Figure 1: Supervised learning evaluation workflow. Models are trained and subsequently evaluated on the withheld test set. The two possible protocols, *code submission* and *results submission*, are illustrated. X represents the features, and y the ground truth, of the test set.

2 Supervised learning

Supervised learning is a foundational paradigm in machine learning where models are trained using labeled data. In this paradigm, for each input instance in the dataset, there is an associated correct output, commonly referred to as *label* or *ground truth*. The primary goal of supervised learning is to construct a model capable of making accurate predictions for unseen instances based on this training.

In a classic supervised learning competition, participants evaluate their models on a given task using a dataset split into training and test sets. Typically, as depicted in Figure 1, participants are provided with a training dataset to develop their models, and the evaluation is conducted on the withheld test set. Each competition phase must feature a different test set to prevent overfitting. Importantly, participants should not have access to the labels of the test set.

The choice of evaluation metrics in supervised learning challenges typically depends on the nature of the task — be it regression, classification, or others. The underlying goal is to objectively measure the performance of submitted models in terms of *accuracy*, *precision*, or other relevant metrics. Further insights into evaluation metrics are provided in the chapter 4.

3 Automated machine learning

Automated machine learning (AutoML) is a field of study that focuses on developing methods and systems that can automate the process of building machine learning models. The goal of AutoML is to make it easier to build accurate and effective machine learning models without requiring extensive human intervention. By nature, AutoML methods are built to be able to solve a wide variety of tasks. Examples of such competitions include the AutoML Challenge Series (Guyon et al., 2019), the AutoDL Challenge Series (Liu et al., 2021), the AutoML Decathlon (Roberts et al., 2022) and the AutoML Cup (Roberts et al., 2023) based on the NAS-Bench-360 benchmark (Tu et al., 2023).

The general competition protocol consists in evaluating the candidate algorithms on a set of m tasks. For each of these tasks, the model is trained from scratch and evaluated on a hold-out test set, as demonstrated by the diagram in Figure 2. The m scores that result from evaluating the algorithm across the tasks are subsequently fed into to a master scoring and ranking process. Although the scores obtained on various tasks could simply be averaged, we suggest computing the average of the ranks achieved by comparing all candidates across the given tasks. This approach ensures a more

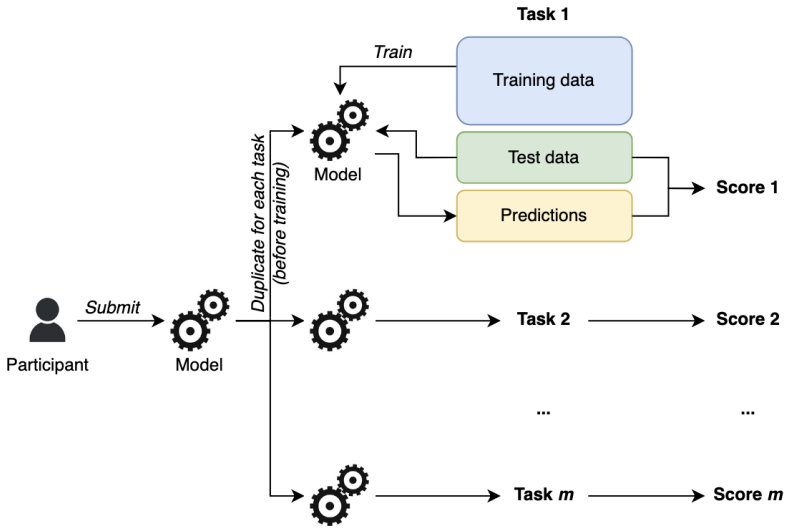


Figure 2: Automated machine learning evaluation workflow. The submitted model is trained and tested from scratch on a set of independent tasks.

robust ranking that accurately reflects the aim of a competition in automated machine learning. This point is explained in details in the chapter 4.

A key aspect of the experimental design of automated machine learning competitions and benchmarks is the **blind testing**. To accurately assess a model's capability to solve diverse and unrelated tasks, participants must not have access to the test data. While some example training datasets can be made available to help participants in developing their models, the feedback and final evaluation stages must be conducted blindly, typically through the submission of code rather than direct interaction with the test data.

The selection of datasets and evaluation metrics is flexible and intrinsically tied to the specific objectives of the challenge. The main principle is that greater diversity in datasets is likely to yield a winning solution with more general applicability. Reciprocally, using similar datasets and metrics is more likely to produce an algorithm specialized in a particular domain or task. Having a large number of different tasks, while computationally expensive, enhances the overall diversity of the model's capabilities. This topic is discussed in the chapter 4.

While most AutoML challenges focus on supervised learning tasks, classification and regression, this experimental design can be used to organize crowd-sourced competitions or benchmarks on the automation of other machine learning tasks, such as data processing, clustering, content recommendation and more. The pre-requisite is to have a scoring metric defining the objective of the problem.

4 Meta-learning

Meta-learning is a sub-problem of AutoML. In its general definition, AutoML is a process of automating the machine learning process, including tasks such as data preprocessing, feature engineering, model selection, and hyperparameter tuning. AutoML techniques use algorithms to search

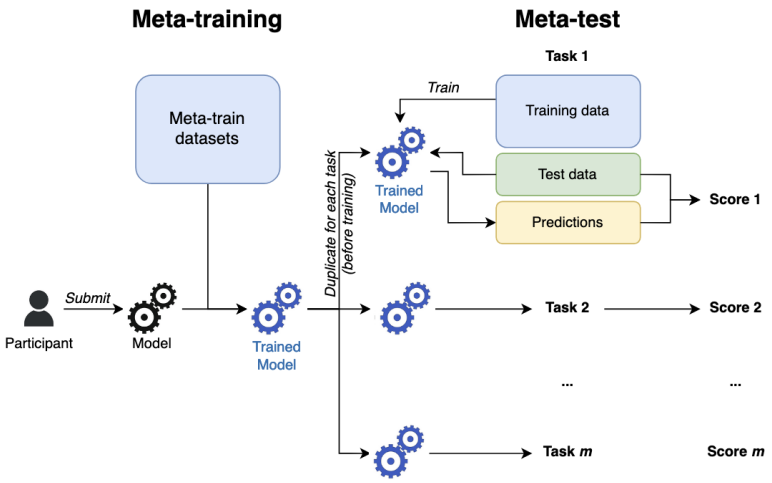


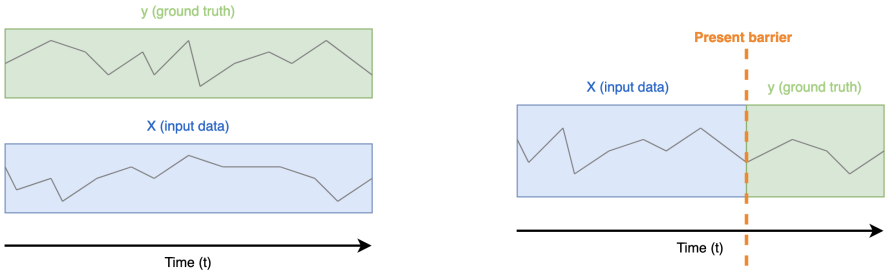
Figure 3: Meta-learning evaluation workflow. The submitted model is trained on meta-train datasets, and then it is tested on a set of meta-test tasks.

for the best machine learning pipeline automatically. On the other hand, meta-learning is focused on learning how to learn. Meta-learning algorithms learn from experience to adapt their learning strategies for different tasks and domains (Brazdil et al., 2022). In essence, while AutoML automates the process of finding the best machine learning pipeline for a specific task, meta-learning takes a step further and automates the process of improving the learning algorithm's generalization capability across multiple tasks.

In the meta-learning challenge protocol proposed by El Baz et al. (2021); Baz et al. (2021) for the Cross-domain MetaDL Challenge, the evaluation of the candidate algorithms is divided in two sequential phases: the meta-training and the meta-test. During the meta-training, the submitted algorithm is trained on a set of datasets. The trained model is then forwarded to the meta-test, where it will be trained and evaluated separately on a new set of tasks. The whole process is illustrated by the diagram in Figure 3. The set of scores produced is then used to compare the model with other candidate models. As for the AutoML Challenge, the entire process is conducted blindly, preventing the participants from adapting their approaches to the specific datasets used. The main difference with the AutoML protocol (Section 3) is the use of a controlled meta-training phase, which implies that all candidate algorithms are pre-trained on the same data.

5 Time series analysis

Time series analysis includes a wide variety of tasks, such as anomaly detection, sequence-to-sequence problems, or survival analysis, each presenting unique specificity. In the this section, our discussion centers around two central time series tasks: *time series regression* and *time series forecasting* (or prediction). While time series regression (Section 5.1) involves modeling the relationship between a dependent time-indexed variable and one or more independent variables, aiming to understand or predict the dependent variable's variations over time, time series forecasting (Section 5.2), on the other hand, is primarily concerned with predicting future values of a series based



(a) Time series regression (b) Time series forecasting

Figure 4: Schematic view of time series regression (left) and time series forecasting (right). In regression, predictors can use past, present, or future values, while in forecasting, the task is to predict future values based on historical data and trends.

on its own past values and inherent patterns. This distinction is highlighted by Figure 4. The key distinction lies in the fact that regression models are more general and can be applied to predict values at any point in time, not strictly in the future, whereas forecasting is explicitly future-oriented, leveraging the temporal order of data to make predictions. Non sequential meta-data may also be available.

5.1 Time series regression

Time series regression is essentially a supervised learning task with a temporal dimension, where the goal is to predict a continuous target variable based on historical data. While it shares similarities with classical regression in terms of learning from input-output pairs and minimizing prediction error, the time component introduces dependencies between observations, necessitating consideration of the order and timing of data points in the modeling process. Time series regression can be multivariate, meaning that multiple variables must be predicted, as it is the case in the *Paris Region AI Challenge 2020* (PRAIC) (Pavao et al., 2021). In this case, the performance can be measured using an average score (weighted or not) across the output variables, or using any ranking function. Another example of time series regression competition is the *AutoSeries Challenge* (Xu et al., 2021), which happens to be also an AutoML competition. This competition confirmed the efficiency of Gradient-Boosting Machines (GMB) to tackle time series regression tasks, as well as random search hyper-parameter tuning to tackle the AutoML part of the problem.

5.2 Time series forecasting

"*A Brief History of Time Series Forecasting Competitions*" by Hyndman (2023) traces the transformative impact of forecasting competitions from the *Makridakis Competitions* series, organized by Spyros Makridakis and spanning from 1980 to today (Makridakis et al., 1982; Makridakis and Hibon, 2000; Makridakis et al., 2018), highlighting their role in shaping forecasting methodologies across diverse data types. The paper emphasizes the consistent success of combination forecasts, encouraging the use of ensemble methods, and points out the balance needed between automated forecasting and domain-specific expertise. Other exemplary time series prediction competitions in-

clude the competitions organized at the Santa Fe Institute (Weigend and Gershenfeld, 1993) which contributed to our understanding of time series prediction in a variety of contexts.

Time series forecasting competitions can be designed in both interactive or non-interactive settings, depending on the objectives and constraints of the challenge. In a **non-interactive** format, participants are provided with a complete dataset up to a certain point in time, and they are required to make predictions for future data points. The models are then evaluated based on their accuracy in predicting these unseen data points. This format is straightforward but may not fully capture the dynamic nature of real-world time series forecasting, where new data continuously become available, and models need to be updated accordingly. On the other hand, **interactive competitions** aim to mimic these real-world conditions by releasing data in stages. Participants make predictions based on available data, and as the competition progresses, new data are released, which can be used to update and improve the models. This format encourages the development of adaptive models that can respond to changes in data patterns over time. This design was typically found in the *COVID-19 Global Forecasting*[1] challenge on Kaggle, where participants were tasked with predicting the spread of COVID-19 disease. Subsequently, the initially unknown ground truth was revealed and added to the training data on a weekly basis.

6 Reinforcement learning

Reinforcement Learning (RL) is a subset of machine learning where an agent learns to make decisions by interacting with an environment. The agent receives feedback in the form of rewards or penalties, guiding it to optimize its behavior to maximize cumulative rewards over time, as illustrated by Figure 5. RL has been successfully applied in various domains, including robotics, game playing, and autonomous vehicles. Organizers of such challenges must choose suitable problem simulations, balance environmental complexity with computational demands, and set objective evaluation criteria that consider efficiency, adaptability, and robustness of the agent's performance.

Designing challenges for RL is an inherently complex task. One of the primary difficulties is the requirement for a simulated or real-world environment where participants' algorithms can interact, learn, and be evaluated. Ensuring the stability, reliability, and realism of these environments is crucial, as inconsistencies or inaccuracies can lead to misleading results and prevent the learning process. Furthermore, RL algorithms typically require a substantial amount of interactions with the environment to learn effectively, making the computational cost a significant consideration. Additionally, there is no one-size-fits-all metric for assessing the performance of RL algorithms across various tasks and environments, necessitating the careful selection and design of evaluation criteria that accurately reflect the objectives of the specific competition.

RL challenges can be designed following different protocols, primarily distinguished by the availability of pre-collected data. In challenges **without** pre-collected data, the algorithms proposed by participants engage directly with the environment, allowing data acquisition and learning concurrently. On the other hand, challenges **with** pre-collected data enable participants to refine and train their algorithms in an offline manner. The setting without pre-collected data is often referred to as **online learning**, and typically occurs in *OpenAI Gym Competitions* such as the *Retro Contest* (Nichol et al., 2018) where agents interacts with video games environment without prior knowledge. This approached is opposed to **offline learning**, which uses pre-collected data, such as the Atari Grand Challenge dataset (Bellemare et al., 2013). This particular dataset comprises a col-

[1] `https://www.kaggle.com/c/covid19-global-forecasting-week-1`

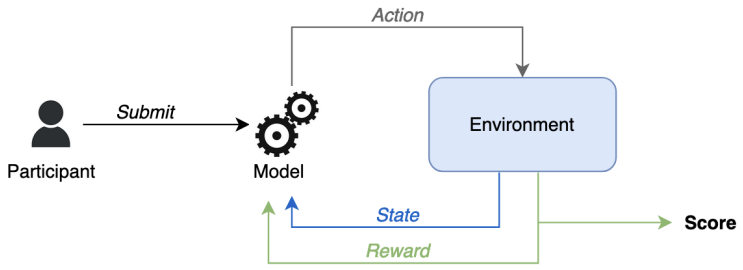


Figure 5: Reinforcement learning competition workflow. The submitted agent interacts with the environment, with the objective of maximizing the reward over time.

lection of human demonstrations across various of Atari games. In offline learning scenarios, the algorithm learns exclusively from this existing data, without the opportunity for real-time interaction or data acquisition, as seen in online learning settings. Regardless of whether algorithms are trained through online or offline learning protocols, their performance must ultimately be evaluated by interacting to live environments.

It is common to use a cumulative reward over time as a primary metric for determining the final score of participants' models. In such settings, the manner in which time is quantified plays a crucial role in the evaluation process, introducing a potential challenge in ensuring equitable and unbiased benchmarking. To mitigate inconsistencies that may arise from hardware disparities, it is advisable to standardize the measurement of total time in terms of the **number of environmental steps** taken, rather than relying on real-time duration measured in seconds. Adopting this approach ensures a consistent and equitable evaluation framework, as it remains invariant across different hardware configurations, thereby enhancing the fairness and reliability of the competition results.

Moreover, in RL challenges, having well-defined and expert-crafted metrics is crucial for evaluating the performance of participating algorithms accurately and fairly. These metrics need to capture not just the immediate rewards but also the long-term impact of decisions made by the RL agents. The design of these metrics requires a deep understanding of the specific domain, the goals of the RL task, and the potential trade-offs between different objectives.

Another interesting distinctions in RL competition design is **interacting with the environment** versus **interacting with other agents**. In the "interaction with environment" setting, the agents submitted by participants are evaluated by interact with the given environment. The related scenes includes single-player games, auto-driving, robot controlling, etc. In the "interaction between agents" setting, the agents are ranked by the performance of competition with other agents. The related scenes includes for instances multi-player games and stock market.

Figure 6 shows the process about the interacting with environment setting and the Figure 7 shows the process about the interacting with other agents setting reinforcement learning competition. The most important issue in the design of reinforcement learning competition is how to evaluate the ranking of submissions. In the setting of interacting with environment, the performance of submissions can be measured by the cumulative reward the submission gained by interacting with the environment. So it is a very important challenge for competition organizers to construct a good simulating environment. The quality of environment determines the quality of the whole competition. In the setting of interacting with other agents, the most popular way is to model the ranking

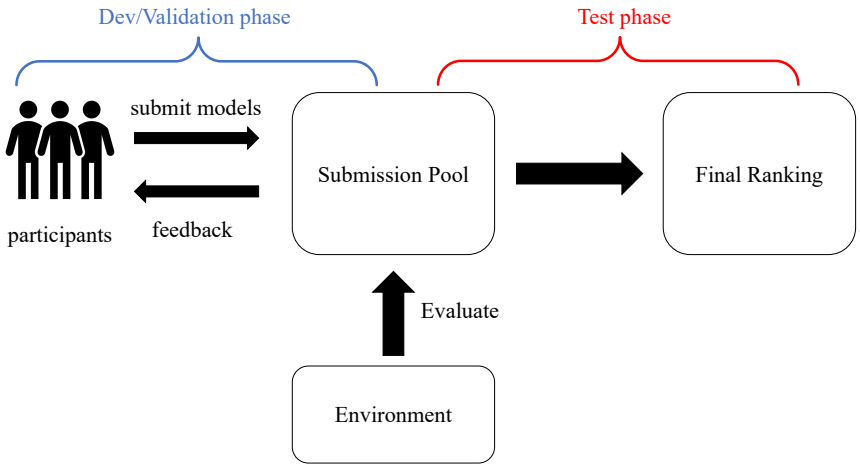


Figure 6: Process of reinforcement learning competitions of interacting with environment.

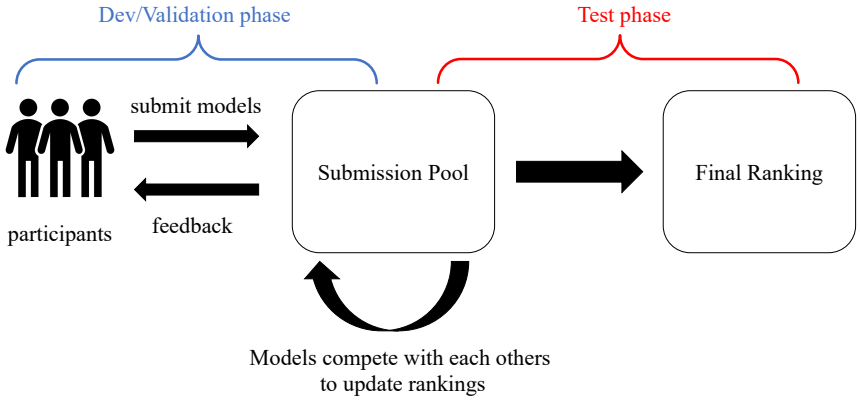


Figure 7: Process of reinforcement learning competitions of interacting with other agents.

score by a Gaussian distribution $N(\mu, \sigma^2)$. Submissions with similar skill rating would be picked to finish a match. The winner's μ will be increased while the loser's μ will be decreased. If there is a draw, the μ of two teams will be moved closer to their mean. Here the key issue is how pick submissions with similar ranking score.

Instances of RL competitions include those in the domain of biomechanics. Notable examples in this category are the *Learning to Run* challenge (Kidzinski et al., 2018) and the *AI for Prosthetics* competition (Łukasz Kidziński et al., 2018). These events enabled progress in simulating and understanding complex biomechanical processes. In addition to biomechanics, RL competitions also extend to video game environments, offering unique challenges that require agents to navigate and interact within virtual worlds. The *MineRL Competition* (Guss et al., 2019) and the *Procgen Competition* (Mohanty et al., 2021) typically test the ability of RL algorithms to adapt and perform across procedurally generated environments. Furthermore, competitions such as *Metalearning from Learning Curves* (Nguyen et al., 2022) explore general aspects of machine learning. This challenge study the meta-analysis of learning processes, encouraging the development of algorithms that can learn effectively from existing learning trajectories.

Two notably interesting examples of past reinforcement learning competitions are the *NetHack 2021 NeurIPS Challenge* and the *Google Research Football with Manchester City F.C.*

Held by Meta and DeepMind, the **NetHack 2021 NeurIPS Challenge** required participants to design an agent to play the game NetHack automatically. NetHack is a single-player video game in which the player is required to navigate the procedurally generated, ascii dungeons to find the amulet. Although it is a very complex game, it can be simulated efficiently by the NetHack Learning Environment (NLE) which is presented at NeurIPS 2020. This competition was split into a development and test phase. During the development phase, participants were able to submit their agents to the leaderboard once a day and 512 evaluation runs would be performed to calculate a preliminary place on the dev-phase leaderboard. The top 15 participants for each track were taken from the dev-phase leaderboard and invited to join the test phase. In test phase, participants were able to submit their best agents 3 to the test-phase leaderboard and 4096 evaluation runs would be performed to calculate the final ranking Net (2020). 42 teams joined this competition and submitted 632 submissions to compete a total prize of 20,000 dollars.

The **Google Research Football with Manchester City F.C** competition was held on Kaggle in 2019 by Google Research and Manchester City F.C. foo (2020). In this competition, each team was required to create AI agents to control a 11-player football team, and them compete with other teams in the simulation environment Kurach et al. (2020). To simplify this challenge, at each time step, the team only need to control one player by choosing an action from a given set of 19 actions. Each submission had an estimated skill rating modeled by a Gaussian distribution $N(\mu, \sigma^2)$, and the rating was updated by the procedure mentioned above. 1,138 teams participated this competitions to compete a total prize of 6,000 dollars.

There are some tips for participants who wants to get good performance in reinforcement learning competitions. The first tip is to focus on the design of reward function, especially in some scenes the reward function is very sparse. The second tip is to design the feature processing model and reinforcement model structure carefully, because they are the key issues to speed up the training progress and improve the model performance. The third tip is to combine with some typical reinforcement learning algorithms such as MCTS and on-policy algorithms.

In conclusion, designing challenges for RL competitions is a nuanced task that requires careful consideration of the learning environment, computational resources, and evaluation metrics.

7 Adversarial learning

Recent research shows that many machine learning classifiers, especially deep learning models, are highly vulnerable to adversarial examples Biggio et al. (2013); Szegedy et al. (2014). An adversarial example is a sample of input data which has been slightly modified to mislead the classifiers while human observers can not notice the modification at all. The existence of adversarial samples raises a huge challenge to the security of machine learning and AI systems. Adversarial learning competition is an important way to examine the adversarial attack and defense algorithms, thus plays an important role in adversarial learning researches.

The adversarial learning competitions includes two aspects: attack and defense. In the attack competition, participants are required to attack a given model. There are three types of attack settings categorized by the revealed information of victim model.

- **White-box attack setting**. In this setting, participants have the full information about the victim model, including the model structure, the value of parameters and hyper-parameters.

- **Black-box attack setting**. In this setting, participants can not directly know the details about the victim model. Instead, participants can query the victim model certain inputs and observe the prediction results.

- **Universal attack setting**. In this setting, participants can not know anything about the victim model or query the victim model. It requires to construct the universal adversarial samples which can mislead most machine learning classifiers.

The attack setting can also be divided into targeted attack and non-targeted attack. In non-targeted attack the adversary only need to cheat the victim model to give a wrong prediction, while in the target attack the adversary is required to mislead the victim model to output a special given label as prediction.

In the defense competition, the participants are required to submit classifiers trained on the given dataset. Then the accuracy of submissions are measured on the adversarial samples constructed by certain adversarial attack algorithm.

Similar with the reinforcement learning competition design, one of the most important issue in adversarial learning competition design is how to evaluate and rank the performance of submissions. For the evasion attack, there are two dimensions in the measure about the attack performance: disturb norm and the attack success rate. For simplicity, we only discuss the case of non-targeted attack, the measure of targeted attack can be designed with similar method. The performance of submissions can be evaluated by the attack success rate with the restricted disturb norm. For example, the score of submitted attack model on test sample x can be represented by

$$\text{score}(g,x) = \begin{cases} 1, \text{ if} ||x - g(x)|| \leq \varepsilon \text{ and } f(x) \neq f(g(x)); \\ 0, \text{ otherwise} \end{cases} . \tag{1}$$

Here g represents the submission, $g(x)$ represents the adversarial sample produced by the submission model, f represents the victim model, and the ε represents the threshold of disturb norm. The performance of submissions can also be evaluated by the disturb norm of adversarial samples which attack the victim model successfully. The loss can be designed as follows:

$$\ell(g,x) = \begin{cases} ||x - g(x)||, \text{ if } f(x) \neq f(g(x)); \\ A, \text{ if } f(x) = f(g(x)) \end{cases}, \tag{2}$$

in which A represents the penalty for adversarial samples failed to mislead the victim model. A must be larger than all the possible disturb norms, i.e. $A \geq \max_{x,x'} ||x - x'||$.

In the evaluation of defense model, the influence of disturb norm should be considered, too. Similar with the evasion attack case, there are two ways to measure the performance of defense model. The first way is to test the defense models with adversarial samples with restricted disturbed norm, then compare the prediction accuracy of defense models on adversarial samples. The second way is to evaluate by the disturb norm of adversarial samples the submissions can distinguish successfully. For example, the score function of the second evaluation method can be designed as follows:

$$\text{score}(f,x') = \begin{cases} 0, \text{if } f(x') \neq y; \\ ||x - x'||, \text{if } f(x') = y \end{cases}, \tag{3}$$

in which (x,y) represents the initial samples and the ground-truth label, and the x' represents the adversarial sample on x.

Interestingly, adversarial learning competitions can also be organized as interactive benchmarks, where particiants' models can attack and defend against each other. Two designs are possible: the **sequential design** where the competition unfolds in distinct stages or phases, and the **simultaneous design** where challenges run both phases concurrently. Examples of sequential adversarial challenges include the ASVSpoof Challenge (Yamagishi et al., 2021; Liu et al., 2023) and the Data Anonymization and Re-identification Challenge (DARC) (Boutet et al., 2020). Examples of sequential adversarial challenges include the Hide-and-Seek Privacy Challenge (Jordon et al., 2020) and the Privacy Workshop Cup (Murakami et al., 2023).

One of the most famous adversarial learning competition is NeurIPS 2017 Adversarial Attacks and Defences Competition organized by Google Brain. This competition is consisted with 3 tracks: 1) non-targeted black-box attack; 2) targeted black-box attack; 3) defense against adversarial attacks. In each track, the participants submitted their models, then the submitted model was given a set of images (and target classes in case of targeted attack) as an input, and had to produce either an adversarial image (for attack submission) or classification label (for defense submission) for each input image Kurakin et al. (2018). The performance of attack models were measured by the average accuracy of victim models, and the performance of defense models were measured by their average accuracy against attack models. 91 teams participated the track 1), 65 teams participated the track 2), and 107 teams participated the track 3).

In IJCAI 2019, Alibaba Group organized an adversarial learning competition including 3 tracks: targeted attack track, non-targeted attack track and the defense track ijc (2019). In this competition, 110,000 pictures of goods from 110 commodity categories are published as training and test sets. In the attack tracks, the submissions were required to attack 5 defense models, then the average disturb norms on these 5 models were used to evaluate the performance of submissions. In the defense track, the submissions were also tested by 5 different attack models, then the average disturb norms of adversarial samples were disputed as the score of submissions. 2519 teams participated this competition to compete a total prize of 39,000 dollars. The teams from USTC won the championship of

defense track, the teams from Southeast University won the championship of target attack track, the teams from Guangzhou University won the championship of non-target attack track.

In KDD 2020, biendata and zhipu.AI organized a competition about the adversarial learning on graph data bie (2020). In particular, this competition is focus on the evasion attack and defense on the citation network de Solla Price (1965). The citation network is a kind of academic graph where academic papers are the nodes and citations are the directed edge. This graph is an important tool which can help researchers to analyse the cite relation of each paper and evaluate the impact of papers. Preventing the attack against the citation network (for example, manipulating citations Chawla (2019)) This competition included 2 phases, in which 543,486 nodes were training set and 50,000 nodes were test set. In the first phase, organizers provided a graph with 593,486 nodes and 100 features on each node. The participants were required to submit a black-box attack model to mislead the organizer's classifier by adding no more than 500 nodes. The performance of attack model was evaluated by the decrease on accuracy of organizer's classifier. In the second phase, each team submitted an attack model and a defense model trained on a similar but different dataset with the one of first phase. Then, the organizer matched all attack models and all defense models. The score of defense model was disputed by the average accuracy on each match, and the score of attack model was disputed by the average error rate on each match. The final score of each team was the average score of its attack model and defense model. 608 partcipants from 511 teams joined this competition to compete a total prize of 20,000 dollars.

Here we provide some tips for participants who wants to get good performance in adversarial learning competitions. For the evasion attack competitions, it is a good idea to combine the attack strategies with the domain knowledge. It is also important to have a well-designed adversarial loss function. For the defense competitions, there are two aspects in which the defense model can be improved. In the feature aspect, feature processing technology such as feature denoising Xie et al. (2019) and feature transformation Song et al. (2020) can help to improve the adversarial robustness of submissions. Some novel models such as topology adaptive model Du et al. (2017) can also be used to construct the adversarial robust model.

8 Use of confidential data

Confidential data may include sensitive information such as personal data, financial data, or trade secrets, and it is important to ensure that this data is handled in a secure and ethical manner. There are several concerns associated with using confidential data in machine learning competitions, including the need to protect the data from unauthorized access, and the need to comply with relevant laws and regulations. Confidential data holds a great importance in many applications, both in scientific and industrial contexts. Some examples include finance, healthcare, and human resources. Using confidential data in crowd-sourced benchmarks is a challenge in itself, but can be highly beneficial by enabling innovation in critical fields.

In this section, we present two different protocols for handling confidential data: **replacing the data by synthetic data**, and **running the participants' models blindly on the real data**. These two protocols, with their advantages and drawbacks, make it possible to crowd-source research on private data without compromising confidentiality.

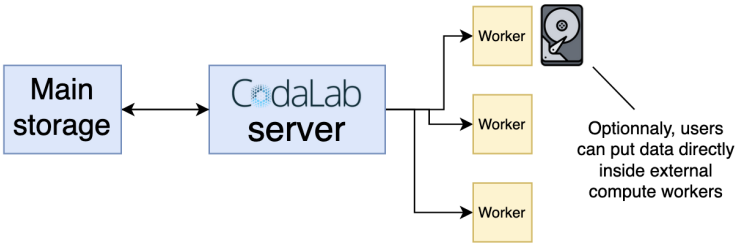


Figure 8: Confidential data can be put directly inside organizers' compute workers, externally from the main servers of the platform.

8.1 Synthetic data

In order to propose a task based on confidential data to the participants without exposing the private data, one approach is to train a generative model to replicate the dataset. Synthetic data can then be generated from the model, and used to simulate the task without disclosing the actual dataset. This approach raises two antagonistic issues: in one hand, the synthetic data must resemble the original data to ensure the problem remains relevant and connected to the real world; on the other hand, the generative model must not leak any real data points. We have developed metrics to evaluate generators *utility* and *privacy* (Yale et al., 2019, 2020) (presented in the chapter 4).

The limitation of using synthetic data is the potential trade-off between *privacy* and *utility*. The *utility* of artificial data can be evaluated by deploying it in real-world scenarios and verifying that model outcomes are consistent with those achieved using real data.

We applied this concept in "To be or not to be", referenced in Pavao et al. (2019), a challenge designed to instruct health students. The task is to predict the survival or decease of patients in intensive healthcare units, based on tabular medical records. The source of the data is the MIMIC-III dataset, which consists in both numerical and categorical variables describing thousands of patients, such as age and blood pressure. Given the inherent confidential and sensitive nature of this data, it is subject to access restrictions. We generated a synthetic dataset using a Wasserstein Generative Adversarial Network (WGAN) (Goodfellow et al., 2014; Arjovsky et al., 2017) model. The resultant challenge continues to be used in Rensselaer Polytechnique Institute to train health students[2].

8.2 Blind access to the data

The second approach for utilizing private data is to blindly execute participants' models on the real data. Two mechanisms are in play to benchmark the participants' solutions despite the private nature of the data: **code submission**, and **storing the data inside the compute workers**, as exposed in Figure 8. This way, only the uploaded models can read the data, it remains completely hidden from the participants. We implemented this feature to *CodaLab Competitions*. This is particularly interesting since, as shown in the chapter 11, organizers can link their own machines to the platform as external compute workers, ensuring a complete control over the data security. We employed this approach in the Paris Region AI Challenge 2020 (Pavao et al., 2021)

[2] `https://codalab.lisn.upsaclay.fr/competitions/3073`

Protocol	Data	Multiple tasks	Code submission	Interactive design
Supervised learning	✓			
AutoML	✓	✓	✓	
Metalearning	✓	✓	✓	
Time series	✓			*depends*
Reinforcement Learning	*depends*		✓	✓
Confidential data	✓		*depends*	
Adversarial challenges	*depends*		*depends*	✓

Table 1: Characterization of the challenge protocols presented in the chapter, indicating the specific criteria that are mandatory (✓), and highlighting those that are possible depending on the design of the challenge (*depends*).

A sample of artificial data, as well as documentation and baseline methods, should be provided to help the participants building their methods despite the constraints associated with not being able to access the dataset directly. Having extensive output logs can also helps the participants to navigate through the problem despite of the blind testing. However, it is advised to limit the size of the output logs to avoid the leakage of the sensitive data. The security of external workers is ensured because the computer workers are owned by the organizers, and *CodaLab Competitions* platform cannot read them. The main limitation of this approach is that it is harder for the participants to work without direct access to the data, making it more difficult to reach the same performance level.

9 Conclusion

In this chapter we analysed the features and special designs of various type of machine learning competitions (adversarial learning, automated machine learning, etc.). We believe that the analysis in this chapter can help both organizers and participants, and also offers reference and inspiration about competitions of novel machine learning paradigms in the future.

In this chapter, we focused on examining the design specificity inherent in competitions and benchmarks in machine learning. We illustrated various experimental designs: supervised learning, AutoML and metalearning, time series analysis, reinforcement learning, the use of confidential data and adversarial challenges. The main characteristics and differences between these designs are outlined in Table 1. A common thread of most of these protocols is the necessity for participants to submit their model's code to the platform for evaluation. This resonates with the recommendation to use code submissions, both allowing complex evaluation procedures and improving the reproducibility and the validity of the evaluation. Interactive designs are at play in reinforcement learning, where algorithms interact with a dynamic environment; in adversarial challenges, where competing algorithms engage with one another; and occasionally in time series prediction tasks, where datasets are regularly augmented with new observations, allowing previously used testing data to become part of the training set for future iterations.

It is also interesting to note that artificial data holds potential utility in certain challenge designs. For tasks where the ground truth is almost exclusively artificial data, or when emulating real data that is confidential, synthetic datasets are beneficial. The latter can also be addressed with real data, by employing blind-testing methods, ensuring participants cannot access confidential datasets. More generally, synthesizing artificial datasets can be beneficial for tasks lacking a ground truth, such as in unsupervised learning, since the synthesis rules can be precisely known by the organizers. Indeed, synthetic data in machine learning competitions can offer a controlled environment to evaluate algorithms, with the advantage of generating diverse and challenging scenarios that resem-

ble complex real-world data distributions. Moreover, using artificial data, organizers can control the task difficulty, generate large datasets, address data imbalance, and reduce data collection cost. Additionally, in scenarios like reinforcement learning, where agents must learn from interaction within an environment, synthetic data provides an endless landscape of tasks for testing the robustness and adaptability of algorithms, as seen in competitions such as the *AI Driving Olympics* (Zilly et al., 2019). The main drawback of this approach is the potential *reality gap* between artificial and real data.

Adversarial learning, focusing on attack and defense algorithms, allows to explore the boundaries of the strengths and weaknesses of existing models. Adversarial learning competition can either focus on attack, on defense, or on both using an interactive design.

Given the diverse and rapidly evolving nature of the field of machine learning, a comprehensive enumeration of all possible design features and evaluation criteria is impossible. For instance, competitions centered on one-shot learning might evaluate the ability of models to generalize from minimal data, while those focusing on fairness could prioritize unbiased predictions across diverse demographic groups. In the domain of real-time processing, the emphasis might shift to algorithmic speed and responsiveness. Tasks involving multi-modal learning demand the integration of information from varied data sources like text, images, and audio. Meanwhile, resource-constrained competitions challenge participants to optimize the model performance with tight computational or memory budgets. However, the methodologies and approaches outlined here can serve as references for future competitions, particularly those in emerging paradigms such as automated machine learning or adversarial challenges.

References

Ijcai-19 alibaba adversarial ai challenge, 2019. https://tianchi.aliyun.com/markets/tianchi/ijcai19_en, Last accessed on 2022-07-15.

Nethack 2020 neurips competition, 2020. https://www.aicrowd.com/challenges/neurips-2021-the-nethack-challenge#challenge-motivation/, Last accessed on 2022-07-15.

Kdd cup 2020: Graph adversarial attack and defense, 2020. https://www.biendata.xyz/competition/kddcup_2020/, Last accessed on 2022-07-15.

Google research football with manchester city f.c., 2020. https://www.kaggle.com/c/google-football, Last accessed on 2022-07-15.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

Adrian El Baz, Isabelle Guyon, Zhengying Liu, Jan N. van Rijn, Sébastien Treguer, and Joaquin Vanschoren. Advances in metadl: AAAI 2021 challenge and workshop. In Isabelle Guyon, Jan N. van Rijn, Sébastien Treguer, and Joaquin Vanschoren, editors, *AAAI Workshop on Meta-Learning and MetaDL Challenge, MetaDL@AAAI 2021, virtual, February 9, 2021*, volume 140 of *Proceedings of Machine Learning Research*, pages 1–16. PMLR, 2021. URL https://proceedings.mlr.press/v140/el-baz21a.html.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47: 253–279, 2013.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 387–402, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40994-3.

Antoine Boutet, Mathieu Cunche, Sébastien Gambs, Benjamin Nguyen, and Antoine Laurent. DARC : Data Anonymization and Re-identification Challenge. In *RESSI 2020 - Rendez-vous de la Recherche et de l'Enseignement de la Sécurité des Systèmes d'Information*, Nouan-le-Fuzelier, France, December 2020. URL https://inria.hal.science/hal-02512677.

Pavel Brazdil, Jan N. van Rijn, Carlos Soares, and Joaquin Vanschoren. Metalearning: Applications to automated machine learning and data mining. 2022.

Dalmeet Singh Chawla. Elsevier investigates hundreds of peer reviewers for manipulating citations. *Nature*, 573(7773):174, 2019.

Derek J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965. doi: 10.1126/science.149.3683.510. URL https://www.science.org/doi/abs/10.1126/science.149.3683.510.

Jian Du, Shanghang Zhang, Guanhang Wu, José M. F. Moura, and Soummya Kar. Topology adaptive graph convolutional networks. *CoRR*, abs/1710.10370, 2017. URL http://arxiv.org/abs/1710.10370.

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 2021. URL https://hal.science/hal-03688638.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

William H. Guss, Cayden Codel, Katja Hofmann, Brandon Houghton, Noburu Kuno, Stephanie Milani, Sharada P. Mohanty, Diego Perez Liebana, Ruslan Salakhutdinov, Nicholay Topin, Manuela Veloso, and Phillip Wang. The minerl competition on sample efficient reinforcement learning using human priors. *CoRR*, abs/1904.10079, 2019. URL http://arxiv.org/abs/1904.10079.

Isabelle Guyon, Lisheng Sun-Hosoya, Marc Boullé, Hugo Jair Escalante, Sergio Escalera, Zhengying Liu, Damir Jajetic, Bisakha Ray, Mehreen Saeed, Michèle 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.

Rob J Hyndman. Forecasting competitions, 2023. URL https://robjhyndman.com/hyndsight/forecasting-competitions/.

James Jordon, Daniel Jarrett, Evgeny Saveliev, Jinsung Yoon, Paul W. G. Elbers, Patrick Thoral, Ari Ercole, Cheng Zhang, Danielle Belgrave, and Mihaela van der Schaar. Hide-and-seek privacy challenge: Synthetic data generation vs. patient re-identification. 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 206–215. PMLR, 2020. URL http://proceedings.mlr.press/v133/jordon21a.html.

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.

Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. Google research football: A novel reinforcement learning environment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4501–4510, Apr. 2020. doi: 10.1609/aaai.v34i04.5878. URL https://ojs.aaai.org/index.php/AAAI/article/view/5878.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 195–231. Springer, 2018.

Xuechen Liu, Xin Wang, Md. Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas W. D. Evans, Andreas Nautsch, and Kong Aik Lee. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2507–2522, 2023. doi: 10.1109/TASLP.2023.3285283. URL https://doi.org/10.1109/TASLP.2023.3285283.

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, Sébastien Treguer, Jin Wang, Peng Wang, Chenglin Wu, Youcheng Xiong, Arber Zela, and Yang Zhang. Winning solutions and post-challenge analyses of the chalearn autodl challenge 2019. *IEEE*

Trans. Pattern Anal. Mach. Intell., 43(9):3108–3125, 2021. doi: 10.1109/TPAMI.2021.3075372. URL https://doi.org/10.1109/TPAMI.2021.3075372.

Spyros Makridakis and Michèle Hibon. The m3-competition: results, conclusions and implications. *International Journal of Forecasting*, 16(4):451–476, Oct 2000. doi: 10.1016/S0169-2070(00)00057-1.

Spyros Makridakis, A Andersen, R Carbone, R Fildes, Michèle Hibon, R Lewandowski, J Newton, E Parzen, and R Winkler. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1(2):111–153, Apr 1982. doi: 10.1002/for.3980010202.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808, 2018.

Sharada P. Mohanty, Jyotish Poonganam, Adrien Gaidon, Andrey Kolobov, Blake Wulfe, Dipam Chakraborty, Grazvydas Semetulskis, João Schapke, Jonas Kubilius, Jurgis Pasukonis, Linas Klimas, Matthew J. Hausknecht, Patrick MacAlpine, Quang Nhat Tran, Thomas Tumiel, Xiaocheng Tang, Xinwei Chen, Christopher Hesse, Jacob Hilton, William Hebgen Guss, Sahika Genc, John Schulman, and Karl Cobbe. Measuring sample efficiency and generalization in reinforcement learning benchmarks: Neurips 2020 procgen benchmark. *CoRR*, abs/2103.15332, 2021. URL https://arxiv.org/abs/2103.15332.

Takao Murakami, Hiromi Arai, Koki Hamada, Takuma Hatano, Makoto Iguchi, Hiroaki Kikuchi, Atsushi Kuromasa, Hiroshi Nakagawa, Yuichi Nakamura, Kenshiro Nishiyama, Ryo Nojima, Hidenobu Oguri, Chiemi Watanabe, Akira Yamada, Takayasu Yamaguchi, and Yuji Yamaoka. Designing a location trace anonymization contest. *Proc. Priv. Enhancing Technol.*, 2023(1): 225–243, 2023. doi: 10.56553/popets-2023-0014. URL https://doi.org/10.56553/popets-2023-0014.

Manh Hung Nguyen, Lisheng Sun, Nathan Grinsztajn, and Isabelle Guyon. Meta-learning from Learning Curves Challenge: Lessons learned from the First Round and Design of the Second Round. working paper or preprint, August 2022. URL https://hal.science/hal-03725313.

Alex Nichol, Vicki Pfau, Christopher Hesse, Oleg Klimov, and John Schulman. Gotta learn fast: A new benchmark for generalization in RL. *CoRR*, abs/1804.03720, 2018. URL http://arxiv.org/abs/1804.03720.

A. Pavao, D. Kalainathan, L. Sun-Hosoya, K. Bennett, and I. Guyon. Design and Analysis of Experiments: A Challenge Approach in Teaching. *NeurIPS*, December 2019. URL http://ciml.chalearn.org/ciml2019/accepted/Pavao.pdf?attredirects=0&d=1.

Adrien Pavao et al. Airplane numerical twin: A time series regression competition. *International Conference on Machine Learning and Applications (ICMLA)*, 2021.

Nicholas Roberts, Samuel Guo, Cong Xu, Ameet Talwalkar, David Lander, Lvfang Tao, Linhang Cai, Shuaicheng Niu, Jianyu Heng, Hongyang Qin, Minwen Deng, Johannes Hog, Alexander Pfefferle, Sushil Ammanaghatta Shivakumar, Arjun Krishnakumar, Yubo Wang, Rhea Sukthanker, Frank Hutter, Euxhen Hasanaj, Tien-Dung Le, Mikhail Khodak, Yuriy Nevmyvaka, Kashif Rasul, Frederic Sala, Anderson Schneider, Junhong Shen, and Evan Sparks. Automl decathlon: Diverse tasks, modern methods, and efficiency at scale. 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 151–170. PMLR, 28 Nov–09 Dec 2022. URL https://proceedings.mlr.press/v220/roberts22a.html.

Nicholas Roberts, Spencer Schoenberg, Tzu-Heng Huang, Dyah Adila, Changho Shin, Jeffrey Li, Sonia Cromp, Cong Xu, Samuel Guo, Adrien Pavao, Ameet Talwalkar, and Frederic Sala. Toward data-centric automl. In *Competition, Poster*. AutoML Conference 2023, 2023.

Chuanbiao Song, Kun He, Jiadong Lin, Liwei Wang, and John E. Hopcroft. Robust local features for improving the generalization of adversarial training. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1lZJpVFvr.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Renbo Tu, Nicholas Roberts, Mikhail Khodak, Junhong Shen, Frederic Sala, and Ameet Talwalkar. Nas-bench-360: Benchmarking neural architecture search on diverse tasks, 2023.

A.S. Weigend and N.A. Gershenfeld. Results of the time series prediction competition at the santa fe institute. In *IEEE International Conference on Neural Networks*, pages 1786–1793 vol.3, 1993. doi: 10.1109/ICNN.1993.298828.

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L. Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Zhen Xu, Wei-Wei Tu, and Isabelle Guyon. Automl meets time series regression design and analysis of the autoseries challenge. *CoRR*, abs/2107.13186, 2021.

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. Privacy preserving synthetic health data. *European Symposium on Artificial Neural Networks (ESANN)*, 2019.

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416: 244–255, 2020. doi: 10.1016/j.neucom.2019.12.136. URL https://doi.org/10.1016/j.neucom.2019.12.136.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md. Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas W. D. Evans, and Héctor Delgado. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *CoRR*, abs/2109.00537, 2021. URL https://arxiv.org/abs/2109.00537.

Julian G. Zilly, Jacopo Tani, Breandan Considine, Bhairav Mehta, Andrea F. Daniele, Manfred Diaz, Gianmarco Bernasconi, Claudio Ruch, Jan Hakenberg, Florian Golemo, A. Kirsten Bowser, Matthew R. Walter, Ruslan Hristov, Sunil Mallya, Emilio Frazzoli, Andrea Censi, and Liam Paull. The AI driving olympics at neurips 2018. *CoRR*, abs/1903.02503, 2019. URL `http://arxiv.org/abs/1903.02503`.

Łukasz Kidziński et al. Ai for prosthetics challenge, 2018. URL `https://www.crowdai.org/challenges/neurips-2018-ai-for-prosthetics-challenge`.

Practical issues: Incentives, community engagement and costs

Magali Richard MAGALI.RICHARD@UNIV-GRENOBLE-ALPES.FR
TIMC
UMR 5525, Univ. Grenoble Alpes, CNRS
F-38700, Grenoble, France

Yuna Blum YUNA.BLUM@UNIV-RENNES1.FR
IGDR
UMR 6290, ERL U1305, Equipe Labellisée Ligue Nationale contre le Cancer, Univ Rennes, CNRS, INSERM
Rennes, France

Justin Guinney JGUINNEY@UW.EDU
Tempus AI, Inc.
Chicago, IL 60654, USA

Gustavo Stolovitsky GUSTAVO.STOLO@GMAIL.COM
DREAM Challenges
New York, NY, USA

Adrien Pavão ADRIEN.PAVAO@GMAIL.COM
Université Paris-Saclay, France

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

Abstract

Each organization of competitions and benchmarks involves a large number of practical problems, such as obtaining sufficient financial support or recruiting participants through appropriate incentives and community engagement. In addition to defining scientific tasks, preparing data and creating challenges, a very important practical administrative organization remains to be achieved. Indeed, cost assessment, corresponding requests for financial support and adequate publicity are key factors for successful organization of the competition. In addition, a good understanding of the incentives that lead participants to engage in a given challenge is fundamental for effective practical organization success. In this chapter, we will cover these topics and give some practical tips and examples for overcoming the "challenge" of organizing the challenges.

Keywords: practical issue, cost, publicity, management

This chapter provides a comprehensive guide to organizing successful scientific competitions, addressing both strategic and practical aspects of challenge organization. We begin by exploring participant motivations and incentives, offering insights into what drives researchers, students, and professionals to engage in scientific challenges. The chapter then delves into community building and outreach strategies, detailing effective methods for recruiting participants and disseminating challenge results within the scientific community. The final sections address the practical aspects of challenge management, including detailed breakdowns of financial and human resource requirements, along with guidance on securing funding sources. Throughout the chapter, we provide concrete examples and actionable recommendations drawn from successful competitions across various scientific domains.

The recommendations presented in this chapter stem from a multi-faceted approach to understanding competition organization. While our primary insights derive from extensive practical ex-

Figure 1: The incentives for participating in a challenge.

perience in organizing scientific competitions, we have strengthened these empirical observations through systematic analysis of documented outcomes from past competitions across various scientific domains. This analysis is complemented by structured feedback collected from both previous participants and experienced organizers, providing valuable perspectives on what contributes to competition success. Furthermore, we have aligned our practical recommendations with current research in the field, particularly regarding best practices in data handling, participant engagement, and competition design. This combination of practical experience, documented evidence, and academic research provides a robust foundation for the guidelines presented throughout this chapter.

1 Incentivizing participation

How to incentivize participants to work on complex problems is a key feature of challenge organization. In this section, we review various types of motivations (Figure 1), from a participant perspective.

1.1 Skills: Knowledge acquisition, communication, education

Traditional university programs in Artificial Intelligence are evolving rapidly, trying to meet the new needs of students, especially on their ability to work collaboratively while improving their scientific knowledge on data mining. Data challenges are mainly based on a coopetitive model, which has the advantage of responding to this dual motivation. Coopetition ((Brandenburger and Nalebuff, 2011) is an active learning pedagogical approach based on the combination of a strategy of competition, where students compete for the best result, and cooperation, where students collaborate for a mutual benefit. Coopetition-based data challenges have the advantage of simultaneously offering two types of learning. On the one hand, this gives a participant a solid methodological training on the scientific question addressed, thanks to the sharing of knowledge between professors and students, but also between the students themselves. On the other hand, these approaches allow students to acquire new skills in collaboration, communication and networking. For more details, please refer to chapter 9: Competitions and challenges in education.

Educational data challenges can be organized into teamwork, recruiting participants from different backgrounds (academic and cultural), with a scientific preparation that can range from minimal information about the challenge before starting to full preparation through a series of dedicated conferences. To meet the expectations of the students, a key factor is the will of the organizer to build a "friendly environment" which will help to boost the motivation of the students and their self-esteem, and to focus more on the process itself than on the results and objectives. Building multidisciplinary teams with different scientific expertise and focusing on real problems are important aspects in the organization of educational challenges. It is also important to provide an environment where participants can communicate with their team members, other teams, and teachers. Ultimately, setting the right reward and price is a major motivator for winning student buy-in (Abernathy and Vineyard, 2001).

Finally, organization of competitions itself can be used as a pedagogical tool. Designing such task is complex and can be, in some regards, more interesting than solving it (Pavao et al., 2019).

1.2 Hot topics: Scientific crowdsourced benchmarking

The quintessential challenge revolves around an existing quantitative standard or benchmark, and seeks to improve upon state-of-the-art. One of the more longstanding benchmark initiatives is the Critical Assessment for Structural Proteins (CASP) that asks participants to predict protein structure (folding) from protein sequence. Groups who specialize in this domain are naturally incentivized to compare their approach in the structured and objective format of a data challenge in the hope that their method out-competes other approaches and can therefore become a new standard in the field (Bender, 2016). CASP is now recognized within the protein structure community as the *de facto* forum for assessing algorithms, and is therefore as much an incentive as a mandate for formal recognition with the community. This incentive generalizes to all specialties, including image recognition (e.g. MNIST (Madry et al., 2019), ImageNet (Russakovsky et al., 2015)), gene identifi-

cation and function prediction (e.g. RGASP (Steijger et al., 2013), CAFA (Radivojac et al., 2013)) or translational research in biomedicine (Saez-Rodriguez et al., 2016).

Any published AI algorithm is expected to include a formal performance comparison against state-of-the-art methods. No good data-driven approach could emerge without good quality, well curated data. This task can be cumbersome and require a great deal of work to assemble and prepare benchmark datasets. Depending on the type of data, data acquisition and/or generation can be very time-consuming and costly (see cost section below). Consequently, a natural perk of a scientific data challenge is that the work involved to generate and prepare a benchmarking dataset is managed by the challenge organizers. Therefore, AI competitions offer a playground with data that are usually costly and complicated to generate. Access to high-quality datasets in machine learning remains an ongoing challenge (10.1007/s00778-022-00775-9). We believe that providing access to such datasets serves as a strong motivation for participants seeking to develop cutting-edge methodological approaches to address complex scientific problems..

Recurrent challenges also present the advantage or keeping people on a regular schedule, as they expect the challenge to come and reserve time for it. As foor a classic scientific event, it provides participants the opportunity to expand their professional network and to start new collaborations with people working in the same research field or people from different disciplines gravitating around the same topic. Finally, data challenges remain the best functioning way of implementing coopetitions: people compete and get credit for winning, then they share their solution publicly and the community can move together to the next step.

1.3 Environment and awards

One appealing aspect of the challenges is the spirit of games. This translates into a friendly yet competitive environment along with rewards. It is not unusual to gather common participants on different challenges. A passion to participate in this type of competitions can develop, along with the excitement of witnessing the evolution of the social community, particularly on commercial platforms like Kaggle. The rewards can be of various nature going from small prizes (e.g. book) to high amounts of money (e.g. 1 million dollars, Salesforce 1 Hackathon) or even positions in companies. Large awards may naturally attract more participants, but this must be balanced with the context of the challenge and the scientific problem being addressed. In other words, factors such as feedback, non-monetary recognition, and opportunities for knowledge advancement should also be considered.

1.4 Visibility, career and recruitment

Challenges are opportunities for participants to showcase their various skills to recruiters and even get a position at the end. A growing number of organizations are adopting modern hiring practices such as challenges to find best candidates. Recruiters use this tool to assess candidates' technical and behavior skills. Challenges have indeed the great advantage of evaluating many different criteria at the same time. Companies can assess technical competencies such as problem solving skills, time management and innovations. They can also assess the behavioral skills they value, such as communication, openness to diversity and leadership.

The implementation of a challenge allows recruiters to define certain expectations towards the evaluated candidates (candidates gain insight into the work culture of their future employer), while verifying if their personality corresponds to the company's fundamental values. One of the diffi-

culties in recruitment is that many companies still follow long selection processes that waste time and interest for both candidates and recruiters. To overcome this problem, challenges can be used to evaluate candidates in a short period of time and a friendly environment, where they can demonstrate real-time expertise. It can also serve as a pre-selection process that will also save time for recruiters.

Interestingly, challenges can bring together a larger number of candidates from more diverse backgrounds than traditional recruiting. Organizers can build a portfolio of interesting candidates for present and future positions, without necessarily limiting themselves to the winners of the challenge. For instance, Kaggle, one of the leading challenges platform acting as a recruiting tool, usis a performance tracking system to evaluate participants[1]. Some companies even sell expertise from Kaggle Grandmasters[2]. Besides, challenges are also an excellent way to increase brand awareness. They can be used as a marketing tactic for big companies to reinforce their leadership in their field. Smaller companies can also increase their visibility though challenges and attract more applicants for a position.

Finally, in addition to recruiting new talent, challenges allow companies to bring innovative solutions and ideas to technical problems. Based on the clear success of challenges in the recruitment process, we can easily expect their increase in the upcoming years.

1. https://www.kaggle.com/progression
2. https://h2o.ai/company/team/kaggle-grandmasters/

Practical tips and resources to optimize incentivization

- Define your working plan and your objectives[a]
- Carefully prepare benchmarking datasets (see Chapter 3 on data preparation).
- Set up a website to collect a list of interested people[b].
- Bring together an expert steering committee
- Provide good educational material together with the challenge (i.e. a good starting kit, white paper).
- Make yourself available during the challenge to answer questions.
- Be responsive to questions on the forum.
- For a recurrent challenge, provide open-source previous winning solutions.
- Organize good publication venues (see details and examples in section 12.2 Community engagement)
- Associate with established conferences (see details and examples in section 12.2 Community engagement)
- For education challenges, you can find inspiration on existing education challenges on open-source platforms such as RAMP[c], or Codalab[d]

a. 10 tips here: http://www.chalearn.org/tips.html
b. see e.g. https://l2rpn.chalearn.org/
c. https://ramp.studio
d. https://codalab.lisn.upsaclay.fr/

2 Community engagement

Mechanisms for engaging and disseminating a competition towards a targeted community are complex and highly dependant on the scientific field. In this section, we try to review general aspects of community engagement that could help challenge organizers to properly define their strategy. See Figure 2 and Table 1 for a review of community engagement strategies and examples of recent competitions.

2.1 Organization of the challenge

The community that will engage in a specific competition will depend on several key aspects of defining the challenge. First, the organizers should define an optimal number of participants and implement the maximum number of participant (if any). Large open competitions have the advantage of ensuring visibility and optimizing scientific production (in the case of crowdsourced benchmarking for example) while smaller competitions will promote communication between participants (more adapted to challenges aiming at educational results). Then they have to determine

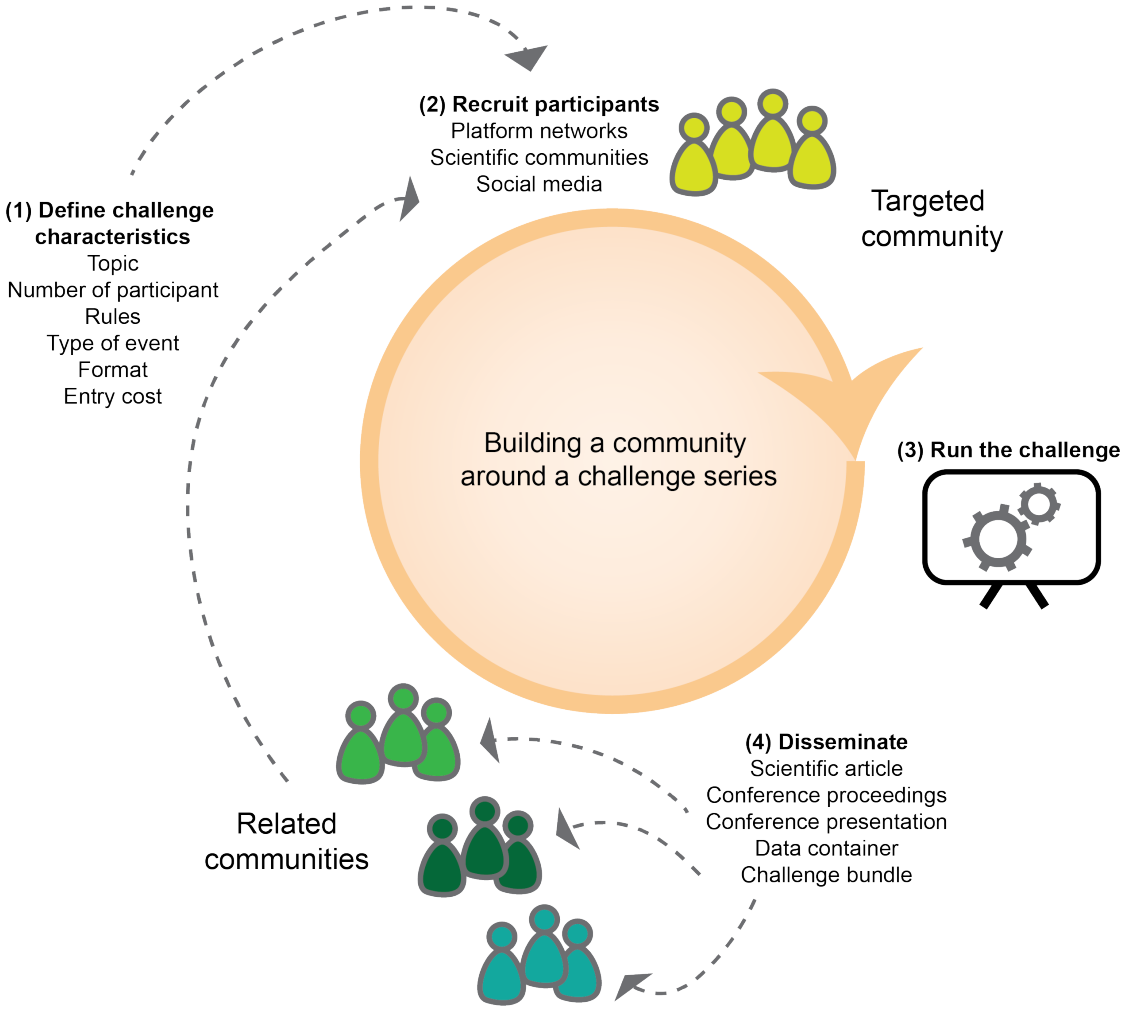


Figure 2: The process of engaging a community

an entry cost: is it easy to participate in the competition? The entry cost depends on several factors: clarity of the rules, specificity of the tasks, size of the dataset, computational resources required to run the methods... All of this will have an impact on the participants who will enter the competition and indirectly define the target audience. Finally, the organizers should established what the format of the competition will be: online events will increase the chances of getting a large pool of participants while in-person events (e.g. at dedicated schools or at scientific conferences) are more suitable for collaborative team work. Once all these parameters are specified, organizers can adapt their communication strategy accordingly and start communicating through dedicated channels, such as the scientific communities mailing list, the digital challenge platform networks and the social media.

2.2 Ensuring diversity and inclusion

A crucial aspect of organizing scientific competitions is ensuring Equity, Diversity, and Inclusion (EDI). Challenge organizers must proactively work to make their competitions accessible and attractive to participants from diverse backgrounds. This includes considering participants who may be traditionally underrepresented due to their gender, race, socioeconomic status, or neurodivergence. Practical measures include offering flexible participation options, such as remote participation possibilities and adjustable deadlines to accommodate different time zones and work constraints. Financial accessibility should be addressed through measures like reduced registration fees for students and participants from low-income countries, or travel grants for in-person events. Ideally, the competition platform and documentation should be designed with accessibility in mind, ensuring compatibility with screen readers and providing materials in multiple formats. Additionally, organizers should establish clear codes of conduct and communication guidelines that promote respectful interactions and create an inclusive environment. The selection of challenge topics, datasets, and evaluation metrics should also be examined for potential biases that might disadvantage certain groups. Building a diverse organizing committee can help identify and address potential barriers to participation early in the planning process. Regular feedback from participants about accessibility and inclusion can help refine these measures over time.

2.3 Challenge output dissemination

The dissemination of the data challenge can take several formats (complementary and not exhaustive) and should match the following question: how would it serve the targeted community?

Participatory benchmarking competitions generally result in scientific publications (see examples (Creason et al., 2021; HADACA consortium et al., 2020; Marbach et al., 2012; Eicher et al., 2019; Marot et al., 2021; Le et al., 2019)) which will be of use to the community. Offering authorship to competing teams, along with participation in manuscript design and writing, is also a strong incentive that will provide international visibility and recognition to participants. Organizers might try to connect with high-profile journal editors ahead of the challenge organization to discuss the possibility of publishing the competition outcome. Depending on the scientific field of the competition, publications can take various form, such as scientific articles, contributions to special issues, conference proceedings, or even books. Best performing teams can also be offered the ability to present their solution in international scientific conferences (e.g., since 2008 all best performing teams in the DREAM Challenges present at the yearly "RECOMB/ISCB RSGDREAM" conference). In addition to an article describing the results of the competition, a challenge built on the data to modeler model (Guinney and Saez-Rodriguez, 2018) could also result in publishing the benchmark dataset along with a container providing a reproducible and continuous benchmark (e.g. a dedicated docker container). Competition data can then be re-used by research scientists as gold standard for new computational methods that will be developed in the future. Challenge organizers may also consider giving open access to their challenge design and templates, especially regarding educational challenges, so that these competition can be massively disseminated to various universities at no cost.

Challenge output and dissemination strategy differ a lot according to the competition organizers and environments. Academic competitions massively rely on the open science framework, encouraging participants to submit their code under an open source license (ex: L2RPN, DREAM challenges). On the opposite, private companies are often motivated by solving an theoretical and

methodological obstacle in order to further develop private commercial solutions that will be put on the market. Such organizers may be more inclined to follow a 'private output' model where participant surrender intellectual property of their findings in exchange for earning money prizes.

COMMUNITY ENGAGEMENT					
Name	Field	Year	Platform	Number of participants	Dissemination
TrackML Particle Tracking Challenge	Physics	2018	Kaggle	739 participants	IEEE WCCI competition (Rio de Janeiro, Jul 2018) and NIPS competition (Montreal, Dec 2018)
LAP series	Computer Vision	2013-22	CodaLab	more than 300 teams	Springer Series on Challenges in Machine Learning, ECCV, IEEE TPAMI, JMLR, IJCV, PAA, CVPR
Tumor Deconvolution	Health	2019-20	DREAM	38 teams	2019 RECOMB/ISCB Regulatory and Systems Genomics, Nature Communications
AutoDL series (6 competitions so far)	Automated ML	2019-21	CodaLab	more than 300 teams	ECML/PKDD, ACML, NeurIPS, IJCNN, WAIC, IEEE TPAMI
Digital Mammography	Health	2017	DREAM	126 teams	RECOMB/ISCB Regulatory and Systems Genomics, JAMA Netw Open.
L2RPN	Energy	2020	CodaLab	more than 300 participants	NeurIPS, ArXiv
Challenge AI for industry	Aeronotic	2021	CodaLab	10 teams	
HADACA series (3 competitions so far)	Life sciences	2018-24	CodaLab	150 participants	BMC bioinformatics, JOBIM

Table 1: Table of communities engagement

As a complement, a non-exhaustive list of conferences that have call for competitions, or can offer workshops and/or proceedings, as well as journals that can welcome competition result publication :

- *Conferences and workshops*: ESANN, ICMLA, WCCI (IJCNN, CEC), ECML/PKDD (Discovery challenges), KDD (KDD cup), CVPR, ECCV, ECML/PKDD, ICPR, ICDAR, IEEE international conference on big data, IEEE International Conference on Automatic Face and Gesture Recognition (FG), ACM SIGIR Forum, NeurIPS dataset and benchmark track, NeurIPS competition track, Workshops @ NeurIPS, ICML, AAAI, CVPR, ICCV, Workshop on Semantic Evaluation, etc.
- *Book series*: CiML Springer series, etc.
- *Journals and pre-prints*: International Journal of Forecasting, International Journal of Information Retrieval Research (IJIRR), IEEE Journal of Biomedical and Health Informatics, IEEE Access, Machine Vision and Applications, IEEE TPAMI, Nature methods, Nature com-

3 Costs, human labor and resources

Depending on the model chosen by the organizers, various costs will be associated with a competition organization. To mitigate the problem of financing a competition, diverse sponsors, private companies or academic institutions can be involved. See Figure 3 and Table 3 for a review of costs and resources associated with recent competitions. Complementary to this section, “Chapter 2: Challenge Design Roadmap" offers guidelines and case studies for developing a robust plan for challenge design.

An example of challenges costs: the L2RPN challenge / NeurIPS 2020

- **Research field**: Energy and environment.
- **Challenge platform**: Codalab[a].
- **Duration of the challenge**: 4 months.
- **Number of participants**: 300.
- **Data generation, access and curation: costs and resources description** : 70,000 euros.
- **Challenge engineering: costs and resources description**: 120,000 euros.
- **Challenge design, scientific expertise: costs and resources description**: 170,000 euros.
- **Prices, travel, conference organization (approximate evaluation of costs)**: 30,000 euros.
- **Challenge governance (cost evaluation of legal, ethics and data privacy costs)**: none.
- **Dissemination**: RTE, Google Research, University College of London, EPRI, IQT Labs. Chalearn.
- **Sponsors**: PMLR[b] & ChaLearn[c]

a. https://competitions.codalab.org/competitions/25426
b. https://arxiv.org/abs/2103.03104
c. https://l2rpn.chalearn.org/

3.1 On overview of the requirements and associated costs

PLATFORM AND REGISTRATION SYSTEM

Several digital platforms can support challenge organization (see chapter 5 for different models of challenge platforms). Defining the platform should be a starting point of challenge organization, as open-source projects such as CodaLab or commercial challenge platforms such as Kaggle will provide very different resources (technical support, engineering manpower dedicated to the compe-

Figure 3: Costs of data challenge organization. Pictures adapted from open work on freepik: macrovector, alvaro_cabrera, visnezh & vectorjuice.

tition...) and associated costs. Please refer to Chapter 5 for more details on the different services provided by each platform.

Data generation, access and curation

High-quality, well-curated data is fundamental to competition success. Recent research, particularly the work of Mougan et al. (2023), has provided comprehensive frameworks for data handling in scientific competitions. A high-quality dataset requires careful planning across multiple dimensions: from initial requirement analysis that clearly establishes the dataset's purpose, through implementation considerations such as sample size and data balance, to thorough documentation and annotation. Additionally, a robust data management plan is essential to ensure data integrity and accessibility throughout the competition (for detailed guidelines on dataset development, see Chapter 4). This

structured approach to data preparation helps ensure that the competition's scientific objectives can be effectively addressed while providing participants with reliable resources for developing their solutions. General cost evaluation of data generation is complicated because it is highly variable depending on the scientific discipline involved. Data generation has always a cost, but this cost can bee supported by different players of the competition (sponsors, private companies, organization committee, care providers, health insurance, etc). This costs also depends on the data type, size and accessibility. Good quality data also relies on the willing of organizers to work in synchronisation with the global efforts for technical standardization and ethic responsible data sharing, e.g. Global Alliance for Genomics and Health or FAIR principles for data management and stewardship (Wilkinson et al., 2016; Cabili et al., 2018).

Governance and legal costs

Competition governance strategy should also include legal counseling costs, that will ensure that the data storage and sharing concept complies with national and international legal requirements. In particular, usage of identifiable personal data (such as patient clinical data) is a complex and significant legal and data protection challenge (Nicol et al., 2019). Moreover, rules for awarding prizes and travel grants should be clearly defined, this includes definitions of:

- jury's composition (committee of experts)
- criteria of evaluation (e.g. relevance, usefulness, novelty, etc.)
- challenge submission process
- intellectual properties
- exclusion and appeal procedures
- control of the use of funds and goods, including prices
- privacy policy
- errors, frauds and breaches of rules mitigation plan

Computation and storage

The digital data challenge platforms rely on cloud computing services to run and evaluate models. Access to these services can be externalized (such as Google Could Platform, Openstack, IBM Cloud or Amazon web services) or provided internally using the computing infrastructure of the challenge organizers. Depending on the competitions, the problem to solve and the type of data, the required resources vary a lot. For instance, in the case of code submission, it is important to estimate well the number of participants, and sometimes to limit the entries by setting a hard threshold. Indeed, code submission offers many advantages (controlled environments, confidential data, good sharing of the resources among participants, etc.) but is computationally very demanding. Thus, the organizers must accurately estimate the computation time of the expected methods as well as the type of computing units to use ((Ellrott et al., 2019)), knowing that donation of cloud units from Google, Azure and Amazon are relatively easy to obtain. Some platform such as Codalab can be coupled with such cloud services, via the use of compute workers. Finally, they need to decide accordingly whether they wish to offer computational services (allowing code submission) or ask participants to provide their own computational resources (only allowing the submission of results).

SCIENTIFIC EXPERTISE, CHALLENGE DESIGN AND ENGINEERING

Bringing together an expert steering committee is a key factor to ensure that the issue raised by the competition corresponds to the needs of the community, and that the data will be used correctly to ask the right question. These two points are essential to ensure community engagement and the quality of the competition. Code development is also an important factor to consider. In certain specific situations, building a dedicated application or a realistic environment to simulate the various tasks of a competition can demand significant effort, including extensive research and substantial engineering manpower prior to the competition. For instance, L2RPN competition series required the generation of a dedicated framework and the generation of synthetic data with several people working on the project for over a year (cost of $\sim$200k€). Once the competition is completed, manpower is also needed to analyse the results, summarize, and disseminate the challenge outcomes.

PRIZES, TRAVEL AND CONFERENCE ORGANIZATION

Reward costs should be included in the challenge budget. Prizes can be an important incentive to recruit participant (see section 1). In case of in-person events, travel and conference organization costs should be considered. This can include speakers invitations, participation to the venue costs and travel grants for students. Competitions can be short (one week) or long (over several months), held remotely or in person, and may or may not be associated with an international conference (see “Part II : The best of challenges and benchmarks" for more examples of academic and industry competitions). All these elements must be taken into account when preparing the budget. For example, the HADACA challenges (Health Data Challenges) take place in the form of a one-week winter school in the French Alps, with around fifty participants. The total cost of organizing the event (including accommodation and meals) was €30,000 for the 2024 edition (HADACA3[3]). Example of costs to organized a one day workshop can be found in Table 2).

	Expense type	Estimated cost (EUR)
1	Invited speakers registration (4x$250)	1,000
2	Organizer travel expenses (3x$2000)	6,000
3	Lunch (catering) for 40*$50	2,000
4	Dinner for invited speakers, winners, organizers (20*$50)	1,000

Table 2: Conference or workshop organization for a total budget of 10,000 euros.

3.2 Person power

Person power is crucial in competition organization and should not be underestimated. While 3 provides an average estimation of person power required to organize a challenge, accurately estimating human resource needs remains one of the most challenging aspects of competition planning. These requirements often evolve throughout the competition lifecycle, with varying demands across different phases - from initial planning to final evaluation. Resource needs can fluctuate based on

3. HADACA3 website : https://hadaca3.sciencesconf.org/

unexpected technical challenges, participant engagement levels, or administrative complexities. A proven strategy to address this uncertainty is to establish a robust technical committee from the outset, comprising members with diverse expertise. This committee should include not only scientific experts but also professionals skilled in administrative tasks, accounting, publicity and communication, software development, data analysis, and reporting. Such diversity in expertise helps ensure that the competition can adapt to evolving demands while maintaining high standards across all aspects of organization. This distributed approach to human resources also provides redundancy and flexibility, allowing the organizing team to better handle peak workloads and unexpected challenges.

3.3 Resources: sponsors and grant agencies

As the global cost of competition organization grows along with the complexity of the data and tasks, proposal and grant writing to find money is essential. By leveraging institutional support and sponsors, organizers will achieve good quality challenges and ensure community participation. More and more universities and national funding agencies[4] or scientific societies[5] support competition organization. Building partnership with private companies[6] and involving collaborators in scientific consortium is also likely to be very helpful to reduce the financial barriers in organizing challenges.

4 Conclusion

Organizing a competition necessitates the dedication of a scientific committee, substantial time, and financial resources. It is imperative not to underestimate the level of commitment required for the successful execution of such events. However, potential organizers should not be discouraged. On the contrary, organizing a competition is a highly rewarding experience, and we encourage any aspiring organizer to undertake it. It's worth noting that competitions represent just one approach to collaborative science. Recent initiatives demonstrate the diversity of possible formats: from large-scale collaborative projects like BLOOM by BigScience [7], which brought together hundreds of researchers to create an open multilingual language model, to the development of innovative evaluation frameworks for language models. Furthermore, while this chapter has primarily focused on traditional competition formats, emerging approaches such as dynamic benchmarking offer promising alternatives to static competitions. These dynamic formats enable iterative data collection and model development, though they require specific design considerations to be implemented effectively."

This chapter is designed as a practical guide, and given the large number of competitions already held, newcomers to the field will find abundant examples to draw inspiration and ideas from. The recommendations and guidelines presented in this chapter are intended to serve as a theoretical framework, not as rigid constraints. The innovative nature of this field extends to the format and design of the competitions themselves, fostering a continuous environment of creativity and development.

4. For instance the University College of London, the National Research Agency in France, the ETH in Switzerland, or the EIT Health in Europe
5. National Science Foundation in the United States, the IEEE Computational Intelligenece Society, or the International Neural Network Society
6. Non-exhaustive list of potential sponsors: Google, Microsoft, Orange, Kaggle, Health discovery corporation
7. BLOOM: https://huggingface.co/bigscience/bloom

H]

Task	Description	Hours
1	**Finding/reviewing data.**	50
2	**Formatting data.** Preprocess and format the data to simplify the task of participants, obfuscate the origin, anonymize.	100
3	**Assessment.** Define a task and evaluation metrics. Define and implement methods of scoring the results and comparing them.	50
4	**Baseline software; starting kit.** Implement a simple example performing the tasks of the challenge. Prepare useful software libraries, make examples.	100
5	**Result formats and software interfaces.** Define the formats in which the results should be returned by the systems and how experimentation will be conducted during the challenge.	50
6	**Benchmark protocol.** Define the rules of the competition and determine the sequence of events.	50
7	**Web portal.** Implement on challenge platform the benchmark protocol allowing on-line submissions and displaying results on a leaderboard.	25
8	**Guidelines to participants.** Write the competition rules, document the formats and the scoring methods, write FAQs.	50
9	**Beta testing.** Organize and conduct tests of the challenge.	25
10	**Run the challenge.** Answer participants, attend to the platform (2h/week).	100
11	**Prepare the workshops.** Write proposals. Look for tutorial speakers. Select speakers. Create a schedule. Advertise.	50
12	**Competition result analysis.** Compile the results. Produce graphs. Derive conclusions.	50
13	**Reports.** Write reports on the benchmark design, the datasets, and the results of the competition.	100
14	**On-line result dissemination.** Make available on-line the competition result analyses, fact sheets of the competitors's methods, and the workshop slides.	50
15	**Prepare workshop proceedings.** Solicit papers, organize the review process, and edit the papers.	100
16	**Distribute prizes and awards.**	10

Table 3: Evaluation of person power to organize a challenge (varies from challenge to challenge, should be estimated by the organizing team)

References

Tammy V. Abernathy and Richard N. Vineyard. Academic Competitions in Science: What Are the Rewards for Students? *The Clearing House*, 74(5):269–276, 2001. ISSN 0009-8655. URL https://www.jstor.org/stable/30189679. Publisher: Taylor & Francis, Ltd.

Eric Bender. Challenges: Crowdsourced solutions. *Nature*, 533(7602):S62–S64, May 2016. ISSN 1476-4687. doi: 10.1038/533S62a. URL https://www.nature.com/articles/533S62a. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7602 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Drug discovery and development Subject_term_id: drug-discovery-and-development.

Adam M. Brandenburger and Barry J. Nalebuff. *Co-Opetition*. Crown, July 2011. ISBN 978-0-307-79054-5.

Moran N. Cabili, Knox Carey, Stephanie O. M. Dyke, Anthony J. Brookes, Marc Fiume, Francis Jeanson, Giselle Kerry, Alex Lash, Heidi Sofia, Dylan Spalding, Anne-Marie Tasse, Susheel Varma, and Ravi Pandya. Simplifying research access to genomics and health data with Library Cards. *Scientific Data*, 5(1):180039, March 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.39. URL https://www.nature.com/articles/sdata201839. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Medical research;Research data Subject_term_id: medical-research;research-data.

Allison Creason, David Haan, Kristen Dang, Kami E. Chiotti, Matthew Inkman, Andrew Lamb, Thomas Yu, Yin Hu, Thea C. Norman, Alex Buchanan, Marijke J. van Baren, Ryan Spangler, M. Rick Rollins, Paul T. Spellman, Dmitri Rozanov, Jin Zhang, Christopher A. Maher, Cristian Caloian, John D. Watson, Sebastian Uhrig, Brian J. Haas, Miten Jain, Mark Akeson, Mehmet Eren Ahsen, Gustavo Stolovitzky, Justin Guinney, Paul C. Boutros, Joshua M. Stuart, Kyle Ellrott, Hongjiu Zhang, Yifan Wang, Yuanfang Guan, Cu Nguyen, Christopher Sugai, Alokkumar Jha, Jing Woei Li, and Alexander Dobin. A community challenge to evaluate RNA-seq, fusion detection, and isoform quantification methods for cancer discovery. *Cell Systems*, page S2405471221002076, June 2021. ISSN 24054712. doi: 10.1016/j.cels.2021.05.021. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471221002076.

Tara Eicher, Andrew Patt, Esko Kautto, Raghu Machiraju, Ewy Mathé, and Yan Zhang. Challenges in proteogenomics: a comparison of analysis methods with the case study of the DREAM proteogenomics sub-challenge. *BMC bioinformatics*, 20(Suppl 24):669, December 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-3253-z.

Kyle Ellrott, Alex Buchanan, Allison Creason, Michael Mason, Thomas Schaffter, Bruce Hoff, James Eddy, John M. Chilton, Thomas Yu, Joshua M. Stuart, Julio Saez-Rodriguez, Gustavo Stolovitzky, Paul C. Boutros, and Justin Guinney. Reproducible biomedical benchmarking in the cloud: lessons from crowd-sourced data challenges. *Genome Biology*, 20(1):195, September 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1794-0. URL https://doi.org/10.1186/s13059-019-1794-0.

Justin Guinney and Julio Saez-Rodriguez. Alternative models for sharing confidential biomedical data. *Nature Biotechnology*, 36(5):391–392, May 2018. ISSN 1546-1696. doi: 10.1038/nbt.4128.

URL http://www.nature.com/articles/nbt.4128. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5 Primary_atype: Correspondence Publisher: Nature Publishing Group Subject_term: Policy;Research data Subject_term_id: policy;research-data.

HADACA consortium, Clémentine Decamps, Florian Privé, Raphael Bacher, Daniel Jost, Arthur Waguet, Eugene Andres Houseman, Eugene Lurie, Pavlo Lutsik, Aleksandar Milosavljevic, Michael Scherer, Michael G. B. Blum, and Magali Richard. Guidelines for cell-type heterogeneity quantification based on a comparative analysis of reference-free DNA methylation deconvolution software. *BMC Bioinformatics*, 21(1):16, December 2020. ISSN 1471-2105. doi: 10.1186/s12859-019-3307-2. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3307-2.

E. P. V. Le, Y. Wang, Y. Huang, S. Hickman, and F. J. Gilbert. Artificial intelligence in breast imaging. *Clinical Radiology*, 74(5):357–366, May 2019. ISSN 0009-9260, 1365-229X. doi: 10.1016/j.crad.2019.02.006. URL https://www.clinicalradiologyonline.net/article/S0009-9260(19)30116-3/abstract. Publisher: Elsevier.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv:1706.06083 [cs, stat]*, September 2019. URL http://arxiv.org/abs/1706.06083. arXiv: 1706.06083.

Daniel Marbach, James C. Costello, Robert Küffner, Nicole M. Vega, Robert J. Prill, Diogo M. Camacho, Kyle R. Allison, DREAM5 Consortium, Manolis Kellis, James J. Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804, July 2012. ISSN 1548-7105. doi: 10.1038/nmeth.2016.

Antoine Marot, Benjamin Donnot, Gabriel Dulac-Arnold, Adrian Kelly, Aïdan O'Sullivan, Jan Viebahn, Mariette Awad, Isabelle Guyon, Patrick Panciatici, and Camilo Romero. Learning to run a Power Network Challenge: a Retrospective Analysis. *arXiv:2103.03104 [cs, eess]*, March 2021. URL http://arxiv.org/abs/2103.03104. arXiv: 2103.03104.

Dianne Nicol, Lisa Eckstein, Heidi Beate Bentzen, Pascal Borry, Mike Burgess, Wylie Burke, Don Chalmers, Mildred Cho, Edward Dove, Stephanie Fullerton, Ryuchi Ida, Kazuto Kato, Jane Kaye, Barbara Koenig, Spero Manson, Kimberlyn McGrail, Eric Meslin, Kieran O'Doherty, Barbara Prainsack, Mahsa Shabani, Holly Tabor, Adrian Thorogood, and Jantina de Vries. Consent insufficient for data release. *Science*, 364(6439):445–446, May 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aax0892. URL https://science-sciencemag-org.insb.bib.cnrs.fr/content/364/6439/445. Publisher: American Association for the Advancement of Science Section: Letters.

Adrien Pavao, Diviyan Kalainathan, Lisheng Sun-Hosoya, Kristen Bennett, and Isabelle Guyon. Design and Analysis of Experiments: A Challenge Approach in Teaching. *CiML Workshop, NeurIPS*, page 3, December 2019. URL http://ciml.chalearn.org/ciml2019/accepted/Pavao.pdf?attredirects=0&d=1.

Predrag Radivojac, Wyatt T. Clark, Tal Ronnen Oron, Alexandra M. Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, Gaurav Pandey, Jeffrey M. Yunes, Ameet S. Talwalkar, Susanna Repo, Michael L. Souza, Damiano Piovesan,

Rita Casadio, Zheng Wang, Jianlin Cheng, Hai Fang, Julian Gough, Patrik Koskinen, Petri Törönen, Jussi Nokso-Koivisto, Liisa Holm, Domenico Cozzetto, Daniel W. A. Buchan, Kevin Bryson, David T. Jones, Bhakti Limaye, Harshal Inamdar, Avik Datta, Sunitha K. Manjari, Rajendra Joshi, Meghana Chitale, Daisuke Kihara, Andreas M. Lisewski, Serkan Erdin, Eric Venner, Olivier Lichtarge, Robert Rentzsch, Haixuan Yang, Alfonso E. Romero, Prajwal Bhat, Alberto Paccanaro, Tobias Hamp, Rebecca Kaßner, Stefan Seemayer, Esmeralda Vicedo, Christian Schaefer, Dominik Achten, Florian Auer, Ariane Boehm, Tatjana Braun, Maximilian Hecht, Mark Heron, Peter Hönigschmid, Thomas A. Hopf, Stefanie Kaufmann, Michael Kiening, Denis Krompass, Cedric Landerer, Yannick Mahlich, Manfred Roos, Jari Björne, Tapio Salakoski, Andrew Wong, Hagit Shatkay, Fanny Gatzmann, Ingolf Sommer, Mark N. Wass, Michael J. E. Sternberg, Nives Škunca, Fran Supek, Matko Bošnjak, Panče Panov, Sašo Džeroski, Tomislav Šmuc, Yiannis A. I. Kourmpetis, Aalt D. J. van Dijk, Cajo J. F. ter Braak, Yuanpeng Zhou, Qingtian Gong, Xinran Dong, Weidong Tian, Marco Falda, Paolo Fontana, Enrico Lavezzo, Barbara Di Camillo, Stefano Toppo, Liang Lan, Nemanja Djuric, Yuhong Guo, Slobodan Vucetic, Amos Bairoch, Michal Linial, Patricia C. Babbitt, Steven E. Brenner, Christine Orengo, Burkhard Rost, Sean D. Mooney, and Iddo Friedberg. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3):221–227, March 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2340. URL `https://www.nature.com/articles/nmeth.2340`. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 3 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Bioinformatics;Protein function predictions Subject_term_id: bioinformatics;protein-function-predictions.

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, December 2015. ISSN 1573-1405. doi: 10.1007/s11263-015-0816-y. URL `https://doi.org/10.1007/s11263-015-0816-y`.

Julio Saez-Rodriguez, James C Costello, Stephen H Friend, Michael R Kellen, Lara Mangravite, Pablo Meyer, Thea Norman, and Gustavo Stolovitzky. Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nature Reviews Genetics*, 17(8):470–486, 2016.

Tamara Steijger, Josep F. Abril, Pär G. Engström, Felix Kokocinski, Tim J. Hubbard, Roderic Guigó, Jennifer Harrow, and Paul Bertone. Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, 10(12):1177–1184, December 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2714. URL `https://www.nature.com/articles/nmeth.2714`. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 12 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genome informatics Subject_term_id: genome-informatics.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George

Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, March 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.18. URL http://www.nature.com/articles/sdata201618. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Publication characteristics;Research data Subject_term_id: publication-characteristics;research-data.