TAILOR

Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization
TAILOR
Grant Agreement Number 952215

Foundational benchmarks and challenges Report

Document type (nature)	Report
Deliverable No	2.3
Work package number(s)	2
Date	Due 30 June 2022
Responsible Beneficiary	INRIA, ID 3
Author(s)	Sébastien Treguer and Marc Schoenauer
Publicity level	Public
Short description (Please insert the text in the Description of Deliverables in the Appendix 1.)	Description of challenges organized during the first 18 months of the project

History			
Revision	**Date**	**Modification**	**Author**
1	05/08/2022	Initial version	Treguer, Schoenauer

Document Review		
Reviewer	**Partner ID / Acronym**	**Date of report approval**
Fredrik Heintz	1 / LIU	22-08-2022
Joaquin Vanschoren	12 / TUE	22-08-2022

For review, the template provided in Folder B on Drive has been used.

Table of Contents

Summary of the report

This report surveys the activities of the TAILOR project related to the organization of challenges[1]. At this point in time (M22 of the project), five challenges are in the pipeline, at different stages of organization: three academic challenges

- *Inductive Links Prediction*, proposed by Fraunhofer (TAILOR partner #29),
- two MetaLearning challenges (Meta Learning from Learning Curves 2, and Cross-Domain MetaDL, proposed by Inria (TAILOR partner #3);

and two industrial challenges

- *Smarter Mobility Data Challenge*, proposed thanks to Electricité de France (EDF, TAILOR partner #48) by the AI Manifesto, a group of 16 French industries whose goal is to promote ethical AI in French Industry,
- *Learning to Run a Power Network* 22 (L2RPN *Energies of the future and carbon neutrality*), proposed by Réseau de Transport d'Electricité (RTE), a long lasting partner of INRIA, though not a TAILOR partner.

The basics of these challenges will be described in turn below, as well as their current status. As these challenges have just started, or are about to start, no results are yet available at the time of writing this report, but for each challenge, a separate document reporting their results will be published after their awards have been made public and their results have been analyzed.

[1] In this document, the word "Challenge" will be used as a synonym of competitions or benchmarks, as is now usual in the AI world, not to be misunderstood as some scientific hurdle to be tackled.

Introduction to the Deliverable

Challenges have been a strong drive in Artificial Intelligence for more than 30 years now, from the very first SAT competitions in 1992 (still on-going) to the series of Visual Recognition Challenges in the early 2010's that definitely demonstrated the incredible effectiveness of Deep Learning approaches. The introduction of Deliverable 2.2 of this project gives a more detailed historical survey of challenges in AI, that will not be repeated here.

In the absence of strong theoretical results in most AI fields, challenges and open benchmarks are the only way to test and compare algorithms on different types of situations in a fair and reproducible way. The success of the historical pioneer Kaggle challenge platform, and its 800000+ AI experts users, led Google to buy it in 2017, in order to "continue democratizing AI", as advocated by Fei-Fei Li in the official announcement. Whatever the actual motivations of Google for such a move, this shows, if at all needed, the importance of challenges in the AI world. However, many AI practitioners, in particular in Europe, have turned to other platforms to organize their challenges, to avoid disclosing their data (and expertise) to this US BigTech company. This boosted other more open and transparent Open Source platforms such as AIcrowd or the university-operated Codalab, that was chosen in the TAILOR proposal to run TAILOR challenges not only because it is a reliable and completely transparent tool, but also because its scientific coordinator is Isabelle Guyon, a pioneer in challenge design and setup, through the Chalearn organization, and a member of the TAILOR INRIA team (partner #3).

Organizing a challenge requires quite some work, and here we refer again to Deliverable 2.2 of this project, where the whole process is detailed and recommendations are given, with specifics related to Codalab. Furthermore, the challenges organized within TAILOR should address TAILOR-related topics, something that is completely problem-dependent and could not be described at the general level in the Deliverable.

The chronological history of TAILOR challenges is the following. The initial plan for TAILOR was to organize one academic and one industrial challenge per year (during the three years initially planned for the project). The academic challenges would be gathered from the 45 TAILOR academic partners, while the industrial challenges would preferably be proposed by the 10 TAILOR industrial partners, plus the analysis of the results of the Theme Development Workshops organized in the context of WP8.

We hence issued a call for challenge topics/data during the Kick-Off meeting (Sept. 29. 2020), for both types of competition, as well as during all meetings of WP8, for industrial competitions. Things started well: we rapidly received two propositions from TAILOR partners: an industrial competition from EDF (together with a consortium of large French industries), regarding Smarter Mobility (optimisation of charging stations for Electric Vehicles) and an academic competition from Fraunhofer (Prediction of Inductive Links). Unfortunately, for many reasons, including of course the Covid pandemic and the absence of physical meetings, but also the inertia of the industrial consortium around EDF, things progressed very slowly, and these challenges are still in the pipeline, hopefully to be launched next Fall for the latter. Also, the Theme Development Workshops only started in Fall 2021, i.e. AI in the Public Sector (Sept. 7 and 9 2021), Future Mobility – Value of Data & Trust in AI (Oct 28 2021), and AI for Future Healthcare (Dec. 16 2021), but no concrete challenge spontaneously emerged from them. Two other were held in Spring 2022, i.e. AI:

Mitigating Bias & Disinformation (May 18 2022), and AI for Future Manufacturing (May 10. 2022), for which the reports are still to come.

It became obvious that we would not be able to organize the promised number of challenges on our own, limited to inputs from TAILOR partners. Therefore, we identified existing challenge series, linked to TAILOR topics, that we could contribute to. We started with the activities of INRIA's TAU group on the Codalab platform, led by Isabelle Guyon, and TAILOR officially joined the organization and the lists of sponsors of the Meta-Learning challenges[2], and the Learning to Run a Power Network challenge (L2RPN). TAILOR contribution consists of human power (for all projects, Sébastien Treguer, hired part time on TAILOR budget, Marc Schoenauer, and of course Isabelle Guyon, plus interns and PhD students), advertisement over TAILOR network and affiliates, and financial contributions: to Codalab storage, with cash prizes for the winners of the Meta-Learning challenges. On the other hand, we will work toward fully organizing new challenges by being more proactive with potential industrial challenge providers, identified either from the TDW reports, or by personal relations (both academic and industrial) of some TAILOR partners (actions to be continued next Fall).

The remaining of this report is the description of the five challenges that are today in the pipeline, with different levels of advancement: The two Meta-Learning Challenges, one ended and one still on-going; the on-going L2RPN challenge; the soon-to-be-launched Smarter Mobility Challenge; and the still-in-discussion Inductive Links Prediction Challenge.

[2] beyond INRIA, TUE (Technical University Eindhoven, TAILOR partner #12), and University Leiden, (TAILOR partner #7) were already participating to the organization

Participants

The following people related to TAILOR have been involved in the organization and running of the challenges:

- INRIA partner #3 is WP 2 and Task 2.3 and 2.4 leader (the tasks that address the challenges): Sébastien Treguer has been hired part time on the project, and has worked on all challenges. So did Marc Schoenauer, while Isabelle Guyon has been the main driving force for the L2RPN and Meta-learning challenges. Apart from these registered participants to TAILOR, other INRIA TAU members have contributed: Adrien Pavao, Research engineer, is the technical coordinator of Codalab, and has been of precious help whenever technical issues arose (and technical issues always arise!). Alessandro Leite (senior researcher) and Eva Boguslawski (PhD student) are working with RTE on the L2RPN; Manh Hung Nguyen and Nathan Grinsztajn (PhD students) and Lisheng Sun-Hosoya (junior researcher) are working on the Meta-Learning from Learning challenge; Dutin Carrion (PhD student) and Ihsan Ullah (intern) contributed to the Cross Domain Meta-Learning challenge.
- EDF partner #48 is part of the Manifesto, organizer of the Smart Moblity challenge, with Alzennyr Gomes Da Silva and Jean-Yves Moise at the steer; another important member of the organizing team is Jérôme Naciri, from Air Liquide (not a TAILOR partner).
- TU Eindhoven partner #12, with Joaquin Vanschoren, and U. Leiden partner #7, with Jan van Rijn, are part of the scientific organization of the Meta-Learning challenges.
- Fraunhofer partner #29 is the main organizer of the Inductive Link Prediction challenge.

Smarter Mobility Data Challenge

Context

This challenge has been spontaneously proposed by EDF (TAILOR partner #48) following the TAILOR kickoff: EDF is part of the Manifesto for AI of 8 (today 16) large French industries, and one of the tasks that the Manifesto had chosen was to design a challenge that would raise the interest of students to eventually come and work for them, demonstrating that AI in industry can be trustworthy indeed. The challenge was hence designed hand in hand with engineers from EDF and Air Liquide, the most motivated members of the Manifesto, together with Sébastien Treguer on TAILOR/INRIA side. Note that the Manifesto is not a legal entity, hence all legal documents had to be approved or signed by all members of the Manifesto, and this sometimes incurred some rather long delay in the formal decisions - hence the delay in the preparatory phases of this challenge.

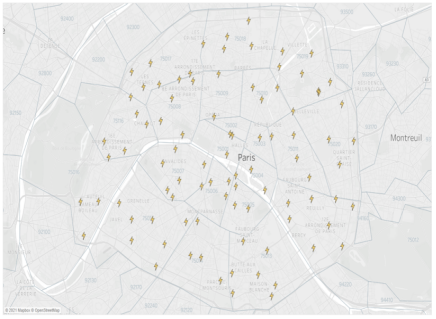
Motivations

Electric mobility development entails new needs for energy providers and consumers. Businesses and researchers are proposing solutions including pricing strategies and smart charging. A proper implementation of charging infrastructure requires a precise understanding of charging behaviors. Thus, EV load models are necessary in order to better understand the impacts of EVs on the grid. With this information, the merit of EV charging strategies can be realistically assessed.
Forecasting occupation of a charging station can thus be a crucial need for utilities to optimize their production units in accordance with charging needs. On the user side, having information about when and where a charging station will be available is of course of interest.
The main topic of the challenge is learning, though solving the charging behavior prediction problem itself will then allow optimization of maintenance, of additions of new stations to the existing network. Furthermore, improving mobility in the cities of the future is one dimension of sustainability of our society.

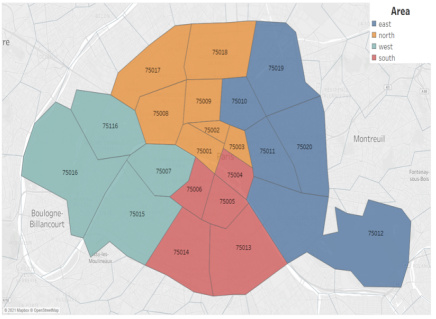
The Challenge

This challenge aims at testing statistical and machine learning forecasting models to forecast the states of a set of charging stations in the Paris area at different geographical resolutions.

This is a hierarchical forecasting problem. The data are split in 4 areas: east, north, west, south (see figure below). The objective of the challenge is to provide state forecasts at 3 different aggregation levels: individual stations, area level and global level (all stations).



Stations



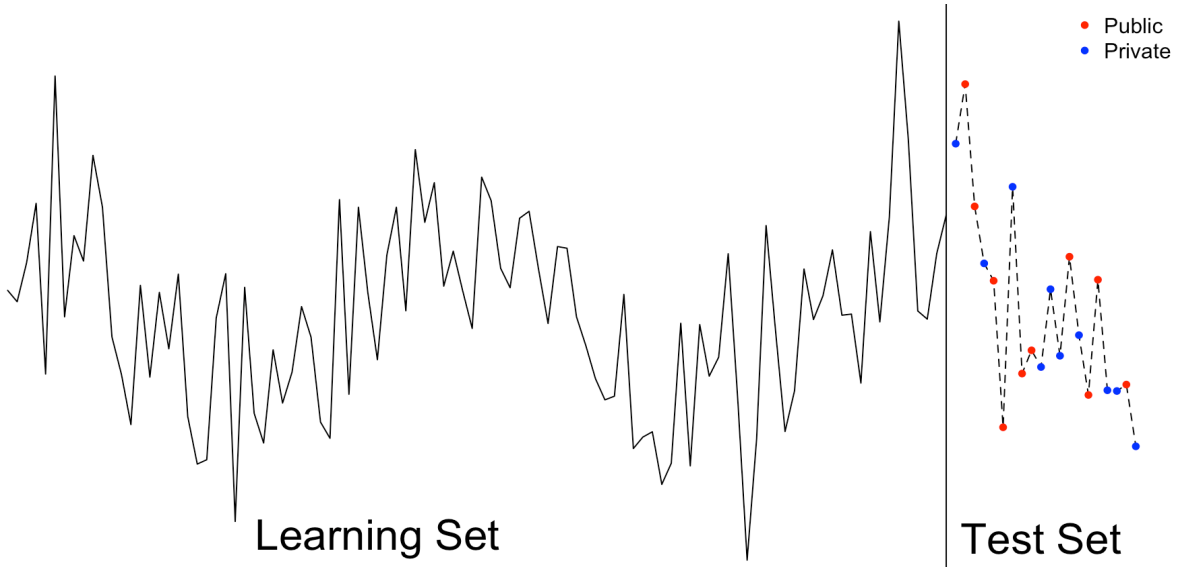
Areas

There are a total of 91 stations. For each station, at every time-step t, 4 characteristics are measured: Available (how many plugs are available), Charging (how many plugs are occupied and charging), Passive (how many plugs are occupied but not charging), Offline (how many plugs are offline). This vector of characteristics denoted, for a station k:

$$y_{t,k} = (a_{t,k}, c_{t,k}, p_{t,k}, o_{t,k})$$

where $a_{t,k}$, $c_{t,k}$, $p_{t,k}$, $o_{t,k}$ are in {0,1,2,3} and sum to 3 for each station k at time t: There are 3 charging plugs per station. For instance, state {a=0, c=2, p=1, o=0} means that no plug is available, 2 plugs are charging, one plug is occupied but not charging, and no plug is offline: the sum a+c+p+o=0. The objective of the challenge is to forecast, according to past information and side information (calendar information such as date, hour of the day, type of day etc) the future state of individual stations, areas and global level simultaneously.

During the development phase, participants will have a learning set to train their model, as well as exogenous information such as calendar information. They will also be able to evaluate the performance of their method during the development phase, with the public test set (red dots in the figure below). During the evaluation phase, their algorithm will be evaluated on the private test set (blue points below).



As the objective of the challenge is to forecast at the individual level, at the area level and at the global level simultaneously, the participants will be asked to forecast at each time t over the forecasting period the variable z_t :

$$z_t = (y_{t,1}, \ldots, y_{t,91}, y_{t,east}, y_{t,north}, y_{t,west}, y_{t,south}, y_{t,global})$$

The participants can choose whether they build a single model or a system of 3 models, one for each level of resolution.

Current status

This challenge will start in the beginning of October 2022, and a precise calendar will then be announced.
Some baselines have already been developed, the technical setup of the Codalab platform is ready, the final adjustments on the competition data and the starting kit (with a new gitlab repository) are ready since August 8th. In the meantime some last tests are ongoing with a group of data scientists from the industrial partners from the Manifesto.
Also, some final adjustments to the ¨Terms and conditions¨ are on-going to, especially to be fully compliant with EU GDPR legal regulations.
Last but not least, Cédric Villani just accepted to co-chair the jury.

The communication (see poster below) is ready to be deployed by the industrial partners to recruit participants, and by TAILOR through its complete network of supporters.

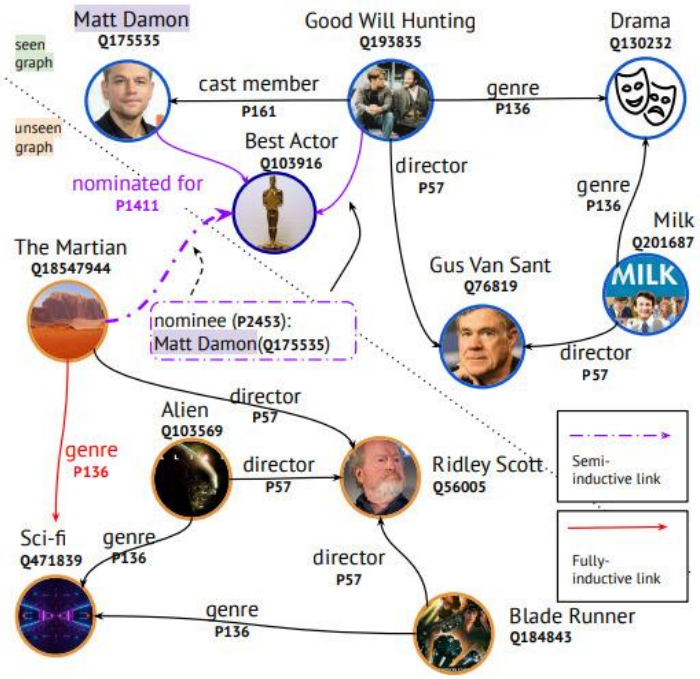
Inductive links prediction Challenge

This challenge has been spontaneously proposed by Mehdi Ali, Jens Lehman and Riccardo Usbeck, from Fraunhofer (TAILOR partner #29) as early as January 2021, based on their work on the subject[3], for which they had developed new benchmarks and datasets based on WikiData Open Database. However, some technical issues (lack of storage on Codalab) and some human factors (beside Covid, and everyone's usual overload, Jens Lehman left Fraunhofer in the meantime) have delayed the materialization of this challenge, though it is still considered "in the pipeline".

Motivations

Knowledge graphs are notorious for their sparsity and incompleteness[4], so that predicting missing links has been one of the first applications of machine learning and embedding-based methods over KGs. A flurry of such algorithms has been developed over the years, and most of them share certain commonalities, i.e., they operate over triple-based KGs in the transductive setup, where all entities are known at training time. Such approaches can neither operate on unseen entities, which might emerge after updating the graph, nor on new (sub-)graphs composed of completely new entities. Those scenarios are often unified under the inductive link prediction (LP) setup. A variety of NLP tasks building upon KGs have inductive nature, for instance, entity linking or information extraction. Hence, being able to work in inductive settings becomes crucial for KG representation learning algorithms. For instance (cf. Fig. 1), the director-genre pattern from the seen graph allows to predict a missing genre link for The Martian in the unseen subgraph. This challenge clearly addresses issues related to TAILOR's moto, here combining Learning and Reasoning.

[3] Mehdi Ali, Max Berrendorf, Mikhail Galkin, Veronika Thost, Tengfei Ma, et al. Improving Inductive Link Prediction Using Hyper-Relational Facts. https://arxiv.org/abs/2107.04894 2021.
[4] Nickel, M., Tresp, V., Kriegel, H.: A three-way model for collective learning on multi-relational data. In proc. ICML, pp. 809–816. Omnipress, 2011.



The Figure above shows different types of inductive LP. Semi-inductive: the link between The Martian and Best Actor from the seen graph. Fully-inductive: the genre link between unseen entities given a new unseen subgraph at inference time. The qualifier (nominee: Matt Damon) over the original relation nominated for allows to better predict the semi-inductive link.

Current status

As of August 2022, a first hurdle has been overcome: there were some issues of data storing capacity of Codalab that prevented it from maintaining the full ontologies for all competitors. TAILOR bought (on INRIA budget) a disk server with 170 Tb of storage to allow this competition to run, and more generally to contribute to Codalab, an Open Source Open Data platform that runs on public hardware without recurrent budget. The WikiData has been downloaded, and the scripts to turn this data into “missing link input” are ready to run. The last missing steps are the choice of the datasets (extracted from WikiData) that will be the public training set, the (hidden) test set for the feedback phase, and the (hidden) test set for the final phase, with increasing difficulty, that depends on the evaluation function, that still is to be designed, too.

Learning to Run a Power Network (L2RPN)

Subtitled *Energies of the future and carbon neutrality*

Context

Power grids transport electricity across states, countries and even continents. They are the backbone of the world societies and economies, playing a pivotal economical and societal role by supplying reliable power to industry, businesses, and domestic consumers. Their importance appears even more critical today as we aim to transition towards a more sustainable world within a carbon-free economy. Problems that arise within the power grid range from transient stability issues with localized blackouts to complete system or country-wide blackouts which can create significant economic and social perturbations. Grid operators are responsible for ensuring that a secure supply of electricity is provided everywhere, at all times, and that systems are designed to be both reliable and resilient. With the advent of renewable energy, electric mobility, and limitations placed on engaging in new grid infrastructure projects, the task of controlling existing grids is becoming increasingly difficult, forcing grid operators to do “more with less”. This challenge aims at testing the potential of AI to address this important real-world problem for our future.

The main organizer and sponsor of this challenge is RTE, France’s power grid transmission system operator, in charge of the grid infrastructure made of more than 105 000 km of high and ultra-high-voltage lines spanning the whole of France, and 50 interconnections with neighboring European countries. In particular, RTE has built a grid simulator called grid2Op that will be heavily used to evaluate the submissions, and has worked to provide the test scenarios that will be used during the competition, based on real-world data from the French Grid operations in recent years.
RTE, though not a member of TAILOR network itself, is a long-lasting partner of INRIA, TAILOR partner #3: Isabelle Guyon and Marc Schoenauer have supervised 3 PhDs in cooperation with RTE – through the French *CIFRE* mechanism.

The problem

In this competition, participants are expected to develop an agent to be robust to unexpected network events and maintain reliable electricity everywhere on the network without risking power overflow, especially when the network is under stress from external events. An opponent will attack in an adversarial fashion some lines of the grid everyday at different times (as an example, you can think of either lightning strikes or cyber-attacks). Participants´ agents have to overcome the opponents' attacks by modifying the topology of the grid, and ensure the grid is operated safely and reliably, i.e., with no risk of overloads. The robustness of participants´ agents will be tested against an opponent with unknown test scenarios that are not part of the training set. The 52 test scenarios, over which we will evaluate submissions, cover a whole week and are selected among all 12 months of the year.
The task to solve is a Reinforcement Learning (RL) task, with a mix of discrete and continuous actions (see below), aiming at a more sustainable carbon-free world, hence clearly falling within TAILOR concerns.

A power grid simulator

The challenge uses Grip2Op, a python module to simulate the power grid. It is modular and can be used to train reinforcement learning agents and to assess the performance of optimal control algorithms. Using Grid2Op, participants can develop, train and evaluate performances of their RL agent that acts on a powergrid in different ways.
An exhaustive documentation of grid2op is available at https://grid2op.readthedocs.io/en/latest/

The action space

There are several types of actions allowed:

- Disconnecting/Reconnecting a power line
- Changing the topology of the grid, for instance choosing to isolate some objects [productions, loads, powerlines] from others
- Modifying the production set point with redispatching actions
- Curtailments of renewable production (not thermal production) under certain conditions especially given physical laws.

The action space contains more than 70,000 discrete actions (topology changes, either at the nodes, or for each line) and 40-dimensional continuous action space (production changes).

The evaluation score

Overall Score

The Score is the quantity that is used to compare agents. The total score is a weighted sum of two scores, the grid “operation cost” score and the attention score:

$$Score = 0.3 * Score_{Attention} + 0.7 * Score_{OperationCost}$$

Notice the different weights. More weight is still given for proper grid operation management. Yet a good amount of points are also given for sending proper alerts while managing the attention of the operator. It can even make you lose more points if not done properly.

Episodes

Each score is more specifically computed over a set of episodes. Formally, we can define an "episode" e successfully managed by an agent up to a time t_{end} (on a scenario of maximum length T_e) by:

$$e = (o_1, a_1, o_2, a_2, ..., a_{t_{end-1}}, o_{t_{end}})$$

where

o_t represents the observation at time t, and a_t the action that the agent took at time t. In particular, o_1 is the first observation and $o_{t_{end}}$ is the last one. The scenario ended at time t_{end}, either because there was a “game over” (i.e., the grid stability reached a breaking point, like a power line breakdown) or because the agent reached the end of the scenario.

The participants will indeed be tested on N hidden weekly scenarios at 5-min resolution, and on various situations that proved difficult to the baselines. This will be the way to test the agent's behavior in various representative conditions. The overall score to minimize over all the scenarios given a cost function c per episode is :

$$Score_{xxx} = \sum_{i=1}^{N} c(e_i)$$

The score metric is describe in more details in the notebook *4_Score_Agent.ipynb* of the starting kit available to download in the section ¨Participate¨ of the challenge on Codalab https://codalab.lisn.upsaclay.fr/competitions/5410#participate-get_starting_kit

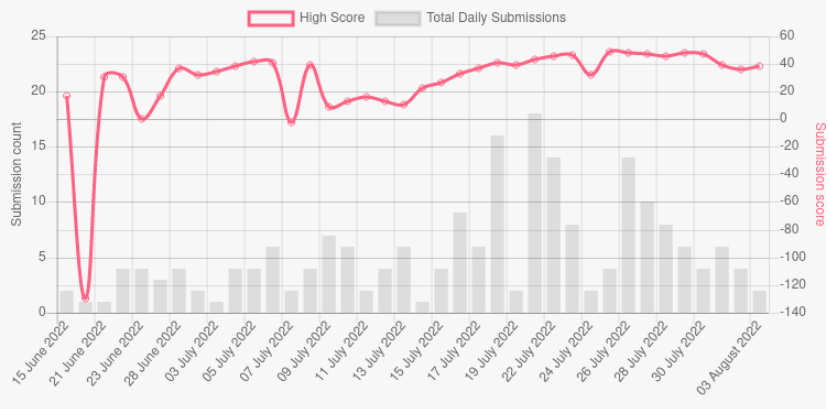
Calendar

- **June 15., 2022: Warmup Phase**: each participant can get use with the problem and the baseline results, and can start developing interesting agents and make good submissions. This phase also allows to get feedback over the clarity, ergonomy, and difficulty of the competition, allowing to improve the competition. Apart from the training data that will not change (except for major unexpected issues), everything else can marginally improve.
- **July 4., 2022: Development Phase**: this is the main phase of the competition during which participants are evaluated on a hidden problem, similar to the one they will be eventually tested on in the last phase. The participants receive feedback on their performance, and can make several submissions, regularly test how her agent is performing, and compare to others in the leaderboard.
- **Sept. 13., 2022: Test Phase**: this is an "automatic" phase under which we evaluate the last submission of the validation phase of each participant on different but similar test scenarios. This will assess against agent overfitting and will create the final leaderboard of the competition.
- **Sept. 30, 2022: Legacy Phase**: all test scenarios will be made publicly available, and the challenge will become an Open benchmark. Anyone can submit agents for experimentation purposes, and see how good they perform in the leaderboard. This phase will be unlimited in time, but there will be no prizes to win.

All details are available from L2RPN web site, and in particular in a white paper describing the challenge design. The repository with the baseline agents is also publicly available here.

The challenge has been accepted as an official competition of WCCI 2022 program and was presented there in July.

As of August 3., 8 teams have made submissions, for a total of 104 submissions, and the daily activity of submissions and scores is shown below.



evolution of the highest score and total daily submissions since competition inception.

An analysis of the submissions and results will be conducted after the end of the test phase, i.e starting from october 2022. It will be added as an annex to this Deliverable.

MetaLearn 2022

Meta-learning is the field of research that deals with learning across datasets. Among several approaches, two very popular methods are MAML (Finn et al.[5] 2017), which aims to learn an initialization for Neural Networks that works well across datasets and can be easily and rapidly fine-tuned on new datasets, and Prototypical Networks (Snell et al.[6], 2017), which builds a metric space in which prototypical examples of new classes can be built and classification done by computing distances to these prototypes. An extension of MAML with a more expressive approach is the LSTM-metalearner[7], which does not only learn the initialization, but also the optimization procedure (learning an optimizer is clearly an AutoAI task combining Learning and Optimization). Interestingly, however, Finn et al. have shown that MAML performs better than the LSTM-metalearner. Huisman et al.[8] (2022) proposed various hypotheses why this could be the case, and developed TURTLE, a novel meta-learning approach that outperforms state-of-the-art methods. Nevertheless, in this domain as in many others of AI such comparisons remain limited to a few test cases, and there is a clear lack of recognized benchmarks: Challenges are one way toward fair and reproducible comparisons in specific contexts.

Under Isabelle Guyon's scientific direction, the Chalearn organization has been organizing Challenges for many years, including the famous AutoML series of challenges that popularized AutoML and helped the rise of auto-sklearn, the state-of-the-art in AutoML on the scikit-learn platform (i.e., not concerned with Deep Learning). These were obviously followed by AutoDL, i.e., AutoAI for Deep Learning. These challenges were, in turn, naturally extended to challenges around Meta-Learning: Meta-Learning from Learning Curves (ML-LC), and MetaDL, that both directly concern TAILOR activities and involve several TAILOR partners in their organization. The first challenges of these series (ML-LC round 1, and MetaDL: a few shot learning competition) were organized too early for TAILOR to become an official partner, but this was possible for the second rounds of both ML-LC (round 2) and MetaDL (Cross-Domain MetaDL). In particular, TAILOR contributed with human-power (Sébastien Treguer, Isabelle Guyon and Marc Schoenauer, plus several other members of INRIA TAU team) and with the money prizes of both Challenges. These two challenges will now be presented in turn.

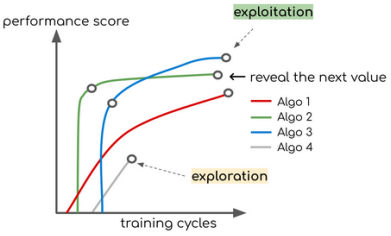
Meta-Learning from Learning Curves 2

Context

When facing a new dataset, the practitioner has to choose an algorithm and its hyperparameters to get the best possible model from the data, i.e., the model that generalizes best on unseen examples. A model is trained on some training set, and the generalization performance is measured on a yet unseen test set. When several instances of

[5] Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In Proc. ICML'17, pp. 1126-1135. PMLR.
[6] Jake Snell, Kevin Swersky, Richard Zemel. Prototypical Networks for Few-shot Learning. In NIPS 2017.
[7] Ravi, S., & Larochelle, H. (2017). Optimization as a model for few-shot learning. In ICLR'17.
[8] Huisman, M., Plaat, A. & van Rijn, J.N. Stateless neural meta-learning using second-order gradients. Mach Learn (2022). https://doi.org/10.1007/s10994-022-06210-y

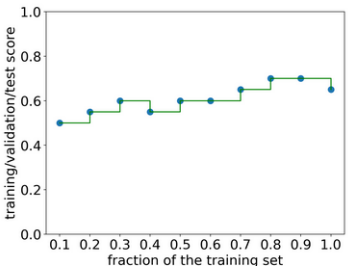
learning algorithms / hyperparameters are available (aka a portfolio of algorithms), it is possible to run in parallel several of them and to dynamically decide after every evaluation which algorithm/hyperparameters to try next, i.e., choosing between exploitation (continue with the current best performing) or exploration (try some yet untested algorithm), as shown on the figure. This can be done from scratch (e.g., using some kind of racing statistical test), or this can be (meta-)learned from sample learning curves (performance vs training time/epoch) on known datasets: Such meta-learning is the goal of these challenges.



The Challenges

The setting for Round 1 was the following: During each phase (see below), meta-data about 15 datasets are given for meta-learning (each meta-example is made of a dataset, meta-features of this dataset, hyperparameters of the algorithm used, training, validation and test learning curves obtained by this algorithm on this dataset). During meta-testing, the agent knows the meta-features of the dataset and the hyperparameters of the different algorithms it can use. The agent must then decide (in a reinforcement learning style) which algorithm to run with which hyperparameters and for how long. It then receives as feedback the learning and validation curves, and must output the next move.
The Area under the Learning Curve (ALC) of the submissions, computed on the test sets of the meta-test datasets, are used to rank them on the leaderboard.

During Round 2, during both meta-learning and meta-training, the learning curves “performance vs time” are replaced with learning curves showing the performance as a function of the dataset size (see figure).

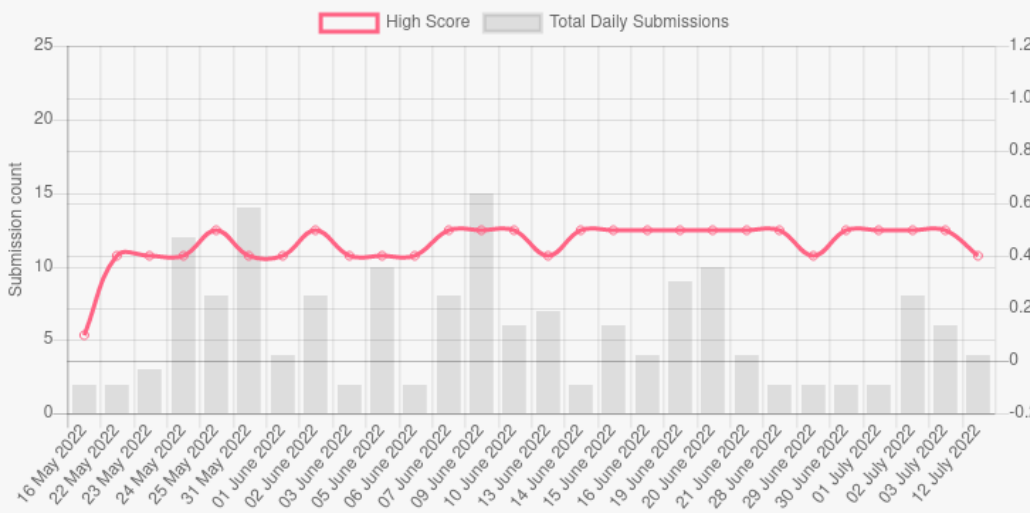


Calendar

- **May 16, 2022**: **Public phase**, using public data, for users to get used to the framework.
- **May 23, 2022**: **Development phase**, the submissions are meta-trained and meta-tested on 15 hidden datasets.
- **July 4., 2022**: **Final/test phase**, the last submission of each participant is meta-learned and meta-tested on 15 fresh hidden datasets, never seen before.
- **July 11., 2022**: **End of competition**, results to be announced at the AutoML conference on July 25. This is also the start or the **Legacy Phase**, all test datasets will be made publicly available, and the challenge will become an Open benchmark. Anyone can submit an algorithm for experimentation purposes, and see how good they perform in the leaderboard. This phase will be unlimited in time, but there will be no prizes to win.

Preliminary View of the Results

The plot below gives the daily number of submissions and best scores all along the competition, which ended on July 11, in order to allow the winners to be announced during the First AutoML conference in Baltimore on July 25.



Whereas the first round welcomed 58 participants for 763 submissions in total, the second round only gathered 44 participants, for a total of 210 submissions, even though it was not mandatory to have participated in the first round to enter the second one. All details of the results (including the links to the details about the winners and their winning approaches) are available on the Codalab web page of the challenge.

However, though the competition has ended, and the winners have been announced, the analysis of the results and the lessons to be learned are not yet available (as of August 3.), and will be published as a report to be annexed to this Deliverable.

Cross-Domain MetaDL 2022

Context and Overview

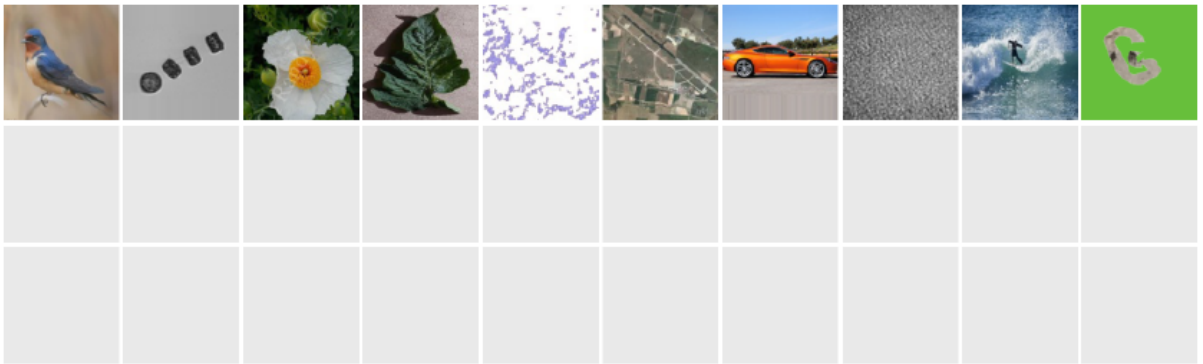
The successful application of deep neural networks often requires large amounts of data and computing resources, restricting its success to domains where such data is available. Meta-learning methods can help tackle this issue by transferring knowledge from related tasks, thus reducing the amount of data and computing resources needed to learn new tasks. The first MetaDL challenge, a NeurIPS 2021 challenge, was a competition on few-shot learning, which attracted over 15 teams that made over 100 code submissions. The lessons learned include that learning good representations is essential for effective transfer learning, and are described in a paper at NeurIPS Competition Track[9] whose co-authors include INRIA (partner #3), Leiden University (partner #7) and TU Eindhoven (partner #12).

[9] Adrian El Baz, Ihsan Ullah, Edesio Alcobaça, André C. P. L. F. Carvalho, Hong Chen, et al.. Lessons learned from the NeurIPS 2021 MetaDL challenge: Backbone fine-tuning without episodic meta-learning dominates for few-shot learning image classification. *NeurIPS 2021 Competition and Demonstration Track*, PMLR 2021.

These results were the basis for a new challenge for meta-learning, called *Cross-Domain Meta-DL*, that is co-organized by the same TAILOR partners, and run as a TAILOR challenge, within TAILOR WP2. The Cross-Domain Meta-DL has been accepted as a NeurIPS 2022 challenge, and is described in detail in a comprehensive white paper. Furthermore, detailed instructions to participants are available as a tutorial available on the challenge web site. While the previous challenge focused on *within-domain few-shot learning* problems, with the aim of learning efficiently N-way k-shot tasks (see details below) for given N and k, this second competition challenges the participants to solve “any-way” and “any-shot” problems drawn from various domains chosen for their humanitarian and societal impact (healthcare, ecology, biology, manufacturing, …).

The data: The Meta-Album

Meta-Album is a meta-dataset (or set of datasets) that has been gathered for few-shot learning and meta-learning (beyond this Challenge), and is made available through the OpenML platform. As of today, it contains 40 datasets from 10 domains (or super-classes - see Figure below), uniformly formatted as 128x128 RGB images, carefully resized with anti-aliasing, cropped manually, and annotated with various meta-data, including super-classes: these are the datasets that have been and are being used for the Meta-Learning Challenges. Also available are the codes and the results of some baseline algorithms.
However, the Meta-Album is intended to be continuously updated and augmented, both on the data side and on the algorithmic side, for Open benchmarking: Together with the data, the results of several baseline algorithms dealing with transfer learning, few-shot meta-learning, and cross-domain few-shot meta-learning tasks are made available.
All datasets and Open Source code is available at https://meta-album.github.io/. A paper giving all details about how this dataset of datasets has been built is currently submitted to NeurIPS and as such available on OpenReview.



The current content of the Meta-Album is made of 10 *domains*: Large animals, small animals, plants, plant diseases, microscopy, remote sensing, vehicles, manufacturing, human actions, optical character recognition (OCR), as can be seen in the figure above. Data sources were very varied, and mostly came from Internet searches, but we also produced our own OCR datasets and obtained novel donated data.
As described in the Cross-Domain Meta-Learning challenge calendar below, 30 out of the 40 datasets will be used for the Challenge, and hence gradually made public as the Challenge

advances. The last 10 datasets are kept hidden for possible further tests, but will be unveiled in 2023 for transparency reasons.

The Challenges

Following the AutoDL challenges, and to tackle the need for Deep Learning of huge datasets, the MetaDL challenges aimed to tackle few shot learning, or how it is possible to take advantage of the results of learning on some previous datasets to allow learning with only few examples of new unseen datasets. As described above, datasets are clustered into domains, each domain containing several distinct datasets, though all images in the datasets of the same domain pertain to a similar concept. The basic idea of the MetaDL challenges is to meta-learn a model on several datasets belonging to different domains (aka **meta-training set**), and to see how this model performs few-shot learning on datasets not seen during the meta-learning phase (aka **meta-test datasets**).
The few-shot learning problems are often referred to as N-way *k*-shot problems. This name refers to the configuration of the tasks at **meta-test time**. Each task consists of a small training set and a small test set, referred to as **support** and **query sets**, respectively. The number of **ways *N*** denotes the number of classes in a task that represents an image classification problem. The number of **shots *k*** denotes the number of examples per class in the **support set**. The final performance of a meta-learning algorithm is evaluated as follows: the algorithm is first meta-trained, fed with all the datasets of the meta-training set. The resulting **meta-learned model** is then meta-tested: for each dataset of the meta-test set, the same meta-learned model is trained on the examples of the support set, and tested on the examples of the query set. The accuracy on the latter is the basis of the final performance (eventually weighted-averaged over the different datasets of the meta-test dataset).

The NeurIPS 2021 MetaDL Challenge was a “within domain” competition, with fixed 5-ways 5 shots meta-testing: each dataset was a multi-class dataset, and half of the classes were used for meta-training, the other half (from the same domain) for meta-testing, and for each class in the meta-test set, 5 examples were used in the support set, and 20 in the query set. However, the winners obtained over 92% accuracy on all 5 meta-test datasets, which means that the problem was probably too easy.
Going further, the NeurIPS 2022 Cross-Domain MetaDL Challenge uses different domains in the meta-learning and the meta-testing phases, and variable N and k at meta-test time (N is randomly chosen in [2,20] and k in [1,20] – they are the same on each meta-test dataset for all participants of course!).
The performance on each meta-test dataset is then the sum of the normalized accuracies for each meta-test dataset:

$$\text{Normalized Accuracy} = \frac{bac - bac_{RG}}{1 - bac_{RG}}$$

where bac is the macro-averaging recall (or average accuracy per class)

$$bac = \frac{1}{\text{num ways}} \sum_{i=1}^{\text{num ways}} \frac{\text{correctly classified examples of class } i}{\text{total examples of class } i}$$

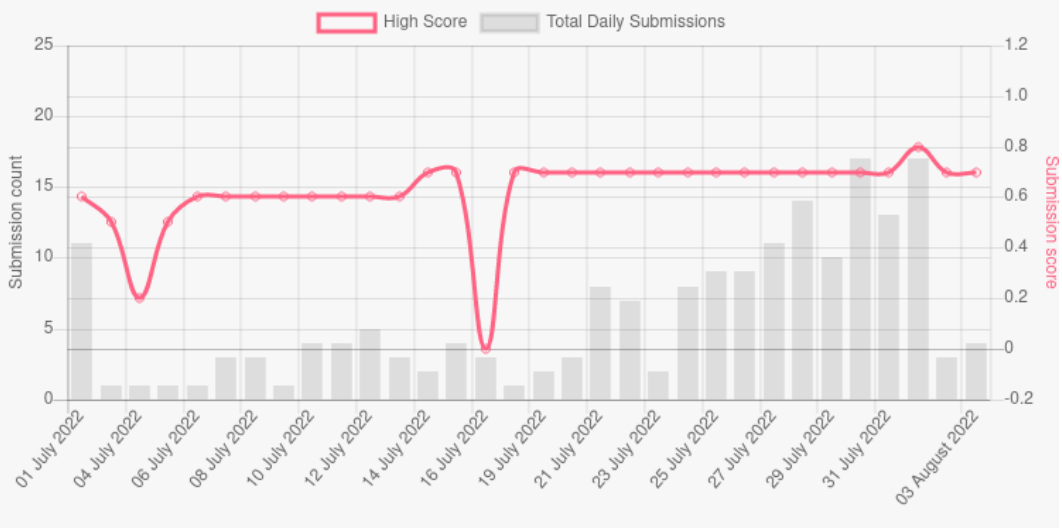
and bac_{RG} is the accuracy of random guessing

$$bac_{RG} = \frac{1}{\text{num ways}}$$

Calendar

- **June 15., 2022: Public phase** (ended) using 10 public datasets, for users to get used to the framework
- **July 1., 2022: Feedback phase** (on-going), using 10 other hidden datasets. Only the performances of the participants are unveiled and published in the leaderboard
- **Sept. 1., 2022: Legal phase**, during which the last submissions of each participant from the feedback phase are blind-tested on 10 new hidden datasets to rank the participants.
- **Oct. 1., 2022: End of competition**, the winners are announced, and invited to publish their approach at NeurIPS Competitions workshop. This is also the start or the **Legacy Phase**, all test datasets will be made publicly available, and the challenge will become an Open benchmark. Anyone can submit an algorithm for experimentation purposes, and see how good they perform in the leaderboard. This phase will be unlimited in time, but there will be no prizes to win.

The plot below gives the daily number of submissions and best scores as of August 3 - the competition is still on-going.



To date, there are 89 registered participants. From these, 34 have submitted at least once a valid submission, so they appear in the leaderboard. There have been 361 submissions in total, but from these, only 187 are the valid submissions, the remaining submissions failed either by an error in the code or by exceeding the maximum allowed running time (5 hours). After the competition has ended and the winners have been announced, a report about the lessons to remember from this challenge will be written, and added as yet another annex to this Deliverable.

Lessons learned and future plans

The main take-home message to date from these challenges is that it is difficult to motivate people to provide use-case and data that will allow us to design a meaningful challenge. And in spite of all recommendations made in Deliverable 2.2 (that most probably even the TAILOR partners who might be motivated to propose a challenge haven't read), people are not aware of the amount of work that is required after providing the basic idea and the data.

Another lesson from the existing challenges concerns the legal and commercial aspects of organizing such events with industry partners. The L2RN challenge, co-organized with RTE, might give a wrong impression of easiness. But keep in mind that this is not the first challenge that RTE is organizing with us, and legal aspects regarding GDPR related to the list of participants for instance had been solved before the organization of the current challenge – this was not the case with the "Smart Mobility" challenge, as pointed out above. Another aspect concerns the industrial property of the data that will be used for the challenges. The use of public data of course solves all problems, but some challenges require very specific data that are part of the industrial know-how of the organizing company, that is reluctant to release them, even if protected by licenses that do not allow their use outside the challenge: they can contain commercial "secrets" that their competitors can guess just by being able to look at them. Last issue on the legal side: it is not obvious, be it for public institutions or for private businesses (at least in France) to give cash prizes to people outside their organization.

As for future plans, as said regarding industrial challenges, we have identified a few companies that participated in the Theme Development Workshops and seemed to be possible candidates for providing test cases and data for challenges (or, eventually, for hackathons). We will start direct discussions with them after the Summer break, rather than count on the results of open calls to the community at large. A dedicated Task Force has been created to investigate both academic and industrial contexts and come up with contact persons among TAILOR partners and beyond, who could possibly be interested in creating new challenges.