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

Foundational benchmarks and challenges v.2 (fused with D8.2 Report on all Hackathons and Benchmarks)

Document type (nature)	Report
Deliverable No	2.6
Work package number(s)	2
Date	Due 30 June 2024
Responsible Beneficiary	INRIA, ID 3
Author(s)	Sébastien Treguer and Marc Schoenauer
Publicity level	Public
Short description	Description of all Data Challenges organized during the second half of the project (M19-M36)

History			
Revision	**Date**	**Modification**	**Author**
1	17/5/24	Initial (almost empty) version	Sébastien Treguer and Marc Schoenauer
2	26/6/24	Mostly complete version sent to Guiseppe and Peter	Sébastien Treguer and Marc Schoenauer
3	28/6/24	Final(?) version, without Peter's review	Sébastien Treguer and Marc Schoenauer

Document Review		
Reviewer	**Partner ID / Acronym**	**Date of report approval**
Guiseppe De Giacomo	UNIROMA	26/6/24
Peter Flach	UNIBRIS	1/7/24

Table of Contents

Summary of the Deliverable

This report surveys the second part of the activities of the TAILOR project related to the organization of Data Challenges[1] since M22 and Deliverable 2.3. However, because (almost) none of the Data Challenges were complete when Deliverable 2.3 was written, this report will in fact contain a description of all TAILOR Data Challenges that took place during the project. However, the description of the Data Challenges that were already presented in Deliverable 2.3 will be very brief, focusing on the results obtained since then.

This include four academic Data Challenges

- Two *MetaLearning Data Challenges* (Meta Learning from Learning Curves 2, and Cross-Domain MetaDL, proposed by Inria TAU (TAILOR partner #3), already presented in Deliverable 2.3;
- *WebCrow*, a Data Challenge about crossword puzzles, proposed by CINI, University of Siena (TAILOR partner #37);
- *Mind the Avatar's Mind,* a first step toward reducing energy consumption in smart buildings, organized by TNO (TAILOR partner #39) and DFKI (TAILOR partner #26).

and five industrial Data 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, with the help of Inria TAU, already presented in Deliverable 2.3;
- *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 TAU, though not a TAILOR partner, also presented in Deliverable 2.3.
- *Brain age prediction challenge*, and *Sleep States challenge*, both proposed by Inria TAU in collaboration with NeurotechX Global Hackathon respectively in 2022 and 2023.
- *Machine Learning for Physical Simulations*, proposed by IRT SystemX with the participation of Inria TAU.

These Data Challenges will be described in turn, and their results only surveyed. Indeed, these results will not be analyzed in detail here, but in Deliverable 2.4 "Lessons learned from TAILOR Challenges", that will in particular analyze a posteriori the relevance of the proposed solutions to TAILOR principles, both regarding the methodology used by the winning teams in light of Learning, Optimizing and Reasoning, and the level of Trustworthiness of the proposed solutions.

[1] In this document, following the advice given by the reviewers of Deliverable 2.3, we will use the term "Data Challenge" for the challenges / competitions / benchmarks, even if Data is not the main focus of these challenges, in order to avoid confusion with the challenges designating scientific hurdles to be tackled. We will however omit "Data" in cases without ambiguity (e.g., the word "challenge" is associated with the name of the Data Challenge).

Introduction

[The first part of this Introduction - in italics - is the Introduction of Deliverable 2.3]
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.

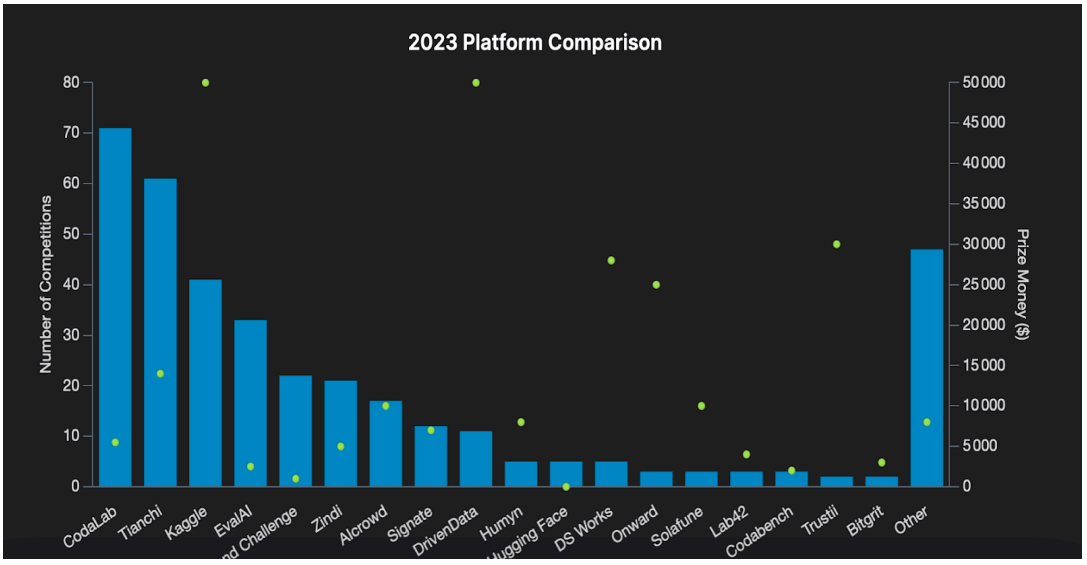
It became obvious that we would not be able to organize the promised number of challenges on our own, limited to inputs from TAILOR partners. Therefore, we identified existing challenge series, linked to TAILOR topics, that we could contribute to. We started with the activities of INRIA's TAU group on 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.

The above introduction was written in July, 2002. It is still valid, but since then, things have progressed fast, as reported in the present Deliverable.

Some important information regarding Codalab, the platform that we have chosen to run most TAILOR-related Data Challenges.

- End 2022 - beginning 2023, Codalab version 2 became live under the name of Codabench. The purpose of the change of name is to indicate that this is a completely different version of the challenge platform - even the word "challenge" is replaced by "benchmarks" to account for the fact that most Data Challenges remain indefinitely open, becoming de facto recognized benchmarks in their domain. The Inria partner was a crucial partner of this move, thanks to Isabelle Guyon and Adrien Pavao (see the explanatory paper).
- The 2023 comparative study published by MLContest **ranked Codalab first**, before Kaggle and Alibaba's Tianchi, with respect to number of competitions run, in particular competitions affiliated with prestigious academic conferences, as is the case for several of TAILOR academic competitions. Interestingly, this happened in spite of the fact that Codalab/Codabench offered in total for 2023 much less prize money than Kaggle or Tianchi: "only" less than 250k$ for Codalab, vs more than 850k$ for Tianchi and 2.3M$ for Kaggle (see plot below giving the statistics for 2023 for the most prominent platforms: the blue bars are the numbers of Data Challenge, the green dots the total prize money).

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



Content of the Deliverable

The first part of this Deliverable is a complement of the descriptions of the five Data Challenges that were already presented in Deliverable 2.3, i.e., that were either completed or on-going as of July 2022. They have all terminated today and their results have been published: The two Meta-Learning Data Challenges (one had already ended in July 2022, the other one ended soon after); The L2RPN challenge, ended in October 2022; The Smarter Mobility Data Challenge, that, after being postponed several times, was finally run from October to December 2022, and for which the jury met and decided on the winners in March 2023. Some details will also be given regarding the Inductive Links Prediction Challenge, that started to be prepared together with TAILOR partner #29 Fraunhofer, was abandoned, at least from a Tailor perspective, due to staff changes at Fraunhofer.

On the other hand, five new Data Challenges were designed and run since then, that are presented next: The Brain Age Prediction Challenge; The Sleep States challenge; The Crossword Puzzles Challenge, aka WebCrow; The Mind the Avatar's Mind Challenge (which turned out to be more an extended Hackathon than a Data Challenge in the end); And the two Data Challenges around Machine Learning for Numerical Simulations organized on Codalab by IRT-SystemX, ML4PhySim, completed in March 2024, and ML4CFD, planned to start on July 1. 2024 and to end by the end of October 2024, i.e., two months after the end of the TAILOR project.

Participants

This section lists the main contributors of the Data Challenges described in this Deliverable.

TAILOR partners

- Inria TAU (partner #3) is leading the tasks related to Data Challenges (2.3 and 2.4), and participated in the organization of all Data Challenges described here except Crossword Puzzle and the Mind the Avatar's Mind challenges.
- TNO (partner #39) proposed and led the Mind the Avatar's Mind challenge. Note that DFKI (partner #26) also participated in the organization of this Data Challenge.
- CINI-U. Siena (partner #37) designed and ran the Crossword Puzzle Challenge
- TUE (partner #12) participated in both the design and the organization of the Cross-Domain MetaDL challenge. Note that Univ. Leiden (partner #7) also contributed to the design of the Cross-Domain MetaDL challenge.
- Fraunhofer (partner #29) initiated the Inductive links prediction Challenge, but unfortunately their staff left the project before the challenge was run.
- EDF, the French national electricity provider (partner #48) is the only TAILOR industrial partner that contributed to the Data Challenges, initiating and being instrumental in the organization of the Smarter Mobility Data Challenge, and bringing the whole French Manifesto for AI with them.

Other Main Contributors

- Chalearn, a non-profit organization, is specialized in organizing Data Challenges. It is lead by Isabelle Guyon, member of Inria TAU, and lead the organization of the MetaLearn challenges
- The French Manifesto for AI, a group of 16 French industries including EDF, which was the main driver of the Smarter Mobility Data Challenge.
- RTE, the French public company operating the French Power Grid co-organized the series of Data Challenges L2RPN (Learning to Run a Power Network).
- IRT-SystemX, a French Public Research and Transfer Institute, organized the Machine Learning for Physical Simulations challenges, together with some of their industrial partners contributing with use cases (RTE, Airbus, Michelin).
- NeuroTechX is organizing its yearly Global Hackathon, that welcomed both the Brain Age Prediction and the Sleep States challenges. Among the main organizers of these challenges beside Inria TAU are TimeFlux, an open-source framework for real-time biosignal processing), and École 42, an unconventional French coding school, that hosted the challenges in its building in Paris.

Smarter Mobility Data Challenge

This industrial challenge was described in detail in Deliverable 2.3. We only summarize here the salient points of this Data Challenge, and describe the progress since July 2022. This Data Challenge has now ended.

Description

The Smarter Mobility Data Challenge aims at testing statistical and machine learning forecasting models to predict the states of a set of charging stations in the Paris area. This is a hierarchical forecasting problem, the objective of the challenge is to provide state forecasts at 3 different aggregation levels: individual stations (91 stations,with 3 plugs each), area level (east, north, west, south, each containing about 20 stations), and global level (all Paris stations grouped together).

Data

Each plug can be charging a car, or plugged into a car already charged, or available, or out of order. Each station reports the states of its 3 plugs. The following real measures have been recorded (plus date, time of the day, day of the week):
Training data: One value every 15mn for each station, from 2020-07-03 00:00 to 2021-02-18 23:45 CET.
Test data: similar measures, following in time the training data, i.e., from 2021-02-19 00:00 to 2021-03-10 23:45 CET (i.e., ground truth is known). 20% of these points are kept secret until the Test Phase.
However, due to technical issues, a significant percentage of the values were missing - and this was one of the difficulties of the task.

Evaluation

The L1 error is used throughout. The loss per zone (area or global, resp.) is simply the sum of the losses of the subzones (station or area resp.). The total loss is the sum of the total loss, the loss per area and the 91 losses per station – to be minimized. The teams could either train one single model for the stations, or one model per area and one global model.

Note that the final ranking was decided by a (human) jury that also took into account an oral presentation of the top teams of the leaderboard, for quality and reproducibility of the proposed approaches. This jury was chaired by Cédric Villani, famous Mathematician (Field Medal) and Member of Parliament (at that time).

Prizes

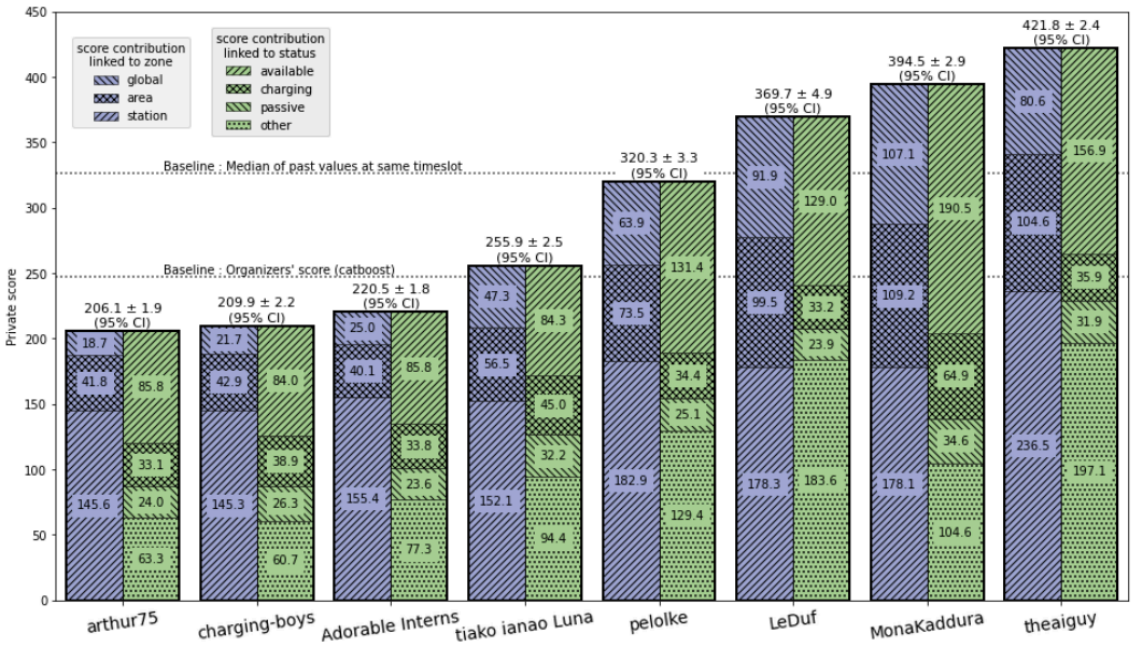
The main goal of the companies in the Manifesto when organizing this Data Challenge was to attract students to the industrial world in general, and in their company in particular. Hence, beyond some small prize money, the winning teams were offered

one day in immersion in one of the participant companies (choice by ranking in the challenge).

Calendar

- **3 Oct. 2022**: Kick-off webinar (recording, password: smarter-kickoff22) and opening of the Development phase
- **13 Oct**.: Data Viz Presentation and Q&A (recording)
- **20 Oct**.: Benchmark Presentation and Q&A (recording)
- **30 Nov.**: Final phase
- **5 Dec.**: Challenge closing
- **10 Feb**. 2023: Jury interviews top of leaderboard and declares the winners
- **10 Mar**.: Prize ceremony

Results



The figure above, taken from the ArXiV paper describing the whole challenge, displays the performances of the top 8 teams on the final leaderboard. The horizontal dotted lines are the results of the baselines provided by the organizers: a very basic one predicting a median value of past measures (only 5 teams have outperformed it, and were interviewed by the jury), and a more advanced one using a tree-based gradient boosting algorithm named CatBoost, that only 3 teams succeeded to beat. These 3 teams were indeed the final winning teams of the challenge.

Links

- Data, code and tutorials on Gitlab.
- A paper describing the challenge and the results
- The codalab page of the challenge

TAILOR Contribution

This industrial Data Challenge was brought by EDF, **TAILOR partner #48**, during TAILOR kickoff (Oct. 2020). However, because it was run under the umbrella of the French Manifesto for AI, a group of 16 large French companies, legal discussions about GDPR and Intellectual Properties took a long time, and the Challenge only started in Oct. 2022. During this period, however, the preparation of the challenge was on-going, and INRIA TAU, **TAILOR partner #3** contributed actively: Marc Schoenauer was part of the initial design discussions and vice-chair of the final jury (also in charge of bringing in Cédric Villani), while Sébastien Treguer helped to technically set up the Challenge on Codalab, design and run the baselines and test the whole pipeline.

Inductive links prediction Challenge

This academic Data Challenge was described in more detail in Deliverable 2.3. However, this description was still preliminary (e.g., the loss function was not completely defined) at that time.

Unfortunately, Mehdi Ali, Jens Lehman and Riccardo Usbeck, from Fraunhofer (TAILOR partner #29), who had proposed this Data Challenge as a TAILOR challenge, left their positions there. In the end, the challenge was actually run, but organized by the other authors of the inspirational paper [*Improving Inductive Link Prediction Using Hyper-Relational Facts*, by Mehdi Ali, Max Berrendorf, Mikhail Galkin, Veronika Thost, Tengfei Ma, Volker Tresp, and Jens Lehmann]. The challenge itself is detailed in this ArXiV paper, but TAILOR was not involved any more, and we have no particular information regarding the results.

TAILOR Contribution

The initial discussions involved Mehdi Ali, Jens Lehman and Riccardo Usbeck, from Fraunhofer (**TAILOR partner #29**) and Marc Schoenauer and Sébastien Tréguer, from Inria TAU (**TAILOR partner #3**).
Furthermore, because this challenge implied using huge amounts of data for its learning phase (the whole wikipedia, for instance, text and knowledge graph), and because at that time Codalab was running short of disk storage, TAILOR made a financial contribution to Codalab by **buying on Inria TAU budget a 170 Tb disk server**.
Nevertheless, even though this challenge was not run on Codalab in the end, several other TAILOR Data Challenges did actually use Codalab (or Codabench, its improved version), and hence TAILOR contribution to this Open Source platform was useful for the progress of science at large.

Learning to Run a Power Network (L2RPN)

Energies of the future and carbon neutrality

This industrial Data Challenge was described in detail in Deliverable 2.3. We only summarize here its salient points, and what happened since July 2022.

Description

The "Learning to Run a Power Network" (L2RPN) Data Challenge is part of a series of competitions designed to address the complex challenges facing modern power networks. This iteration focuses on two critical aspects: achieving carbon neutrality by 2050 from a sustainability perspective, and leveraging Reinforcement Learning (RL) to control power grids from a computer science standpoint.
This Data Challenge is motivated by the increasing complexity of power network operations due to several factors:

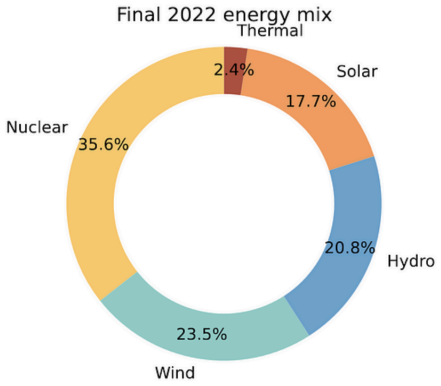
- The integration of renewable energy sources, which introduce unpredictability in power generation.
- The rise of electric mobility, and new changing demand patterns.
- Limitations on new grid infrastructure projects, necessitating more efficient use of existing resources.

These factors are making the task of controlling power grids, such as the French National Grid maintained by RTE, increasingly difficult. Grid operators are now faced with the challenge of doing "more with less."
The L2RPN challenge not only seeks to advance the field of AI in power systems but also to bridge the gap between the AI community and power system experts. It provides a platform for AI researchers to apply cutting-edge techniques to a critical infrastructure problem, while allowing power system professionals to explore innovative solutions beyond traditional methods.
By fostering innovation in this crucial area, the L2RPN challenge contributes to the development of more sustainable and efficient solutions for future power network operations. It encourages the creation of adaptive, robust AI agents that can potentially revolutionize grid management, supporting the transition to a more renewable-based energy system while maintaining grid stability and reliability.

Data

For this Data Challenge, time series describing the electricity injections into the power network, referred to as chronics, have been generated. These chronics take into account the amount of electricity injected into the network by generators, loads, and batteries. In order to train their RL agents, participants were provided with chronics, representing 32 years worth of scenarios.

In order to generate this edition’s chronics, the use of renewable energy generators have been prioritized and penalties have been applied to fossil fuel generators. Chronics have been created with a very limited carbon emission energy mix, as depicted in the following figure:



L2RPN 2022 energy mix over a year.

Evaluation

The submitted agents were evaluated on different chronics of demand/production, and changes in the topology of the grid (because of accidents, maintenance operations, etc). The first goal of the controller is to avoid blackout, i.e., to keep the grid in security, making sure that all current values remain within the safety bounds defined by some Public Authority.

To rank the participants, a score function was needed that would evaluate the performance of an agent and assign it a numerical score. The score function was designed as the average of three cost functions (in Euros), computed over unseen test scenarios:

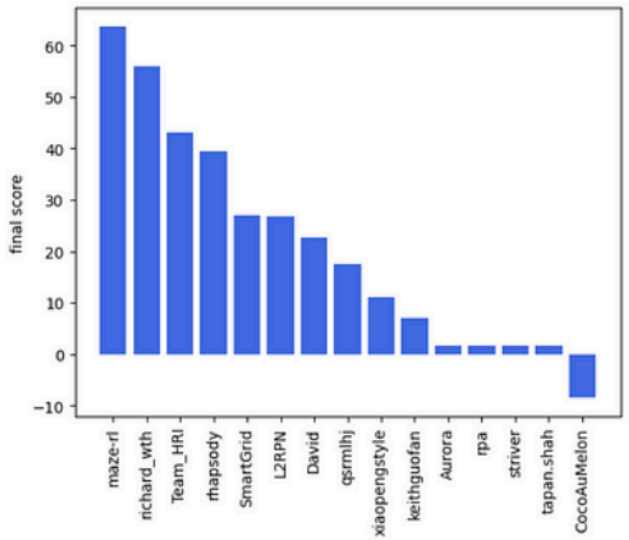
Blackout Cost: When the agent fails to maintain the power network within its security bounds until the end of the scenario, this cost is calculated by multiplying the remaining electricity to be supplied by the current price per MWh.

Operation Cost: This cost is the sum of all expenses incurred by the actions of the agent during operation.

Energy Losses Cost: Due to the Joule effect, some energy is lost during its transportation. The corresponding cost is calculated by multiplying the electricity loss by the current price per MWh.

Results

A total of 16 participating teams made an entry on the final phase of the competition, among which 5 were ranked above the baseline (called “L2RPN” in the following), as can be seen on the figure.
Interestingly, there was no overfitting to the training data in this Data Challenge: the leaderboards of the Development phase and the Final phase were perfectly correlated.
All winners had to provide their code in Open Source (links in the detailed results description).



Calendar

- **June 15., 2022**: Warmup Phase, no submission
- **July 4., 2022**: Development Phase, candidates can make as many submissions as they want, and get feedback on public scenarios
- **Sept. 13., 2022**: Final Phase, each team can make one single submission, that is evaluated on some unknown scenarios, leading to the final leaderboard of the competition and that determines the winners.
- **Sept. 30, 2022**: Legacy Phase: all test scenarios have been made publicly available.

From there on, the Data Challenge became an Open benchmark, for experimentation purposes.

Links

- The description of the setup
- The Codalab web site of the competition, with the leaderboards of all 4 phases.
- Detailed description of the results

TAILOR Contribution

This industrial Data Challenge was organized and run on the one hand by Inria TAU (**TAILOR partner #3**), more precisely by Isabelle Guyon, Adrien Pavao (Isabelle Guyon's PhD student), Eva Boguslawski (Marc Schoenauer's PhD student joint with RTE) and Gaëtan Serré (intern), and on the other hand by RTE (the operator of the French Power Grid), a long-lasting collaborator of Inria TAU, with Antoine Marot and Benjamin Donnot (former joint PhD student of Isabelle Guyon and Marc Schoenauer, now with RTE). It was run on the Codalab platform.

Two Meta-Learning Data Challenges

The two following Data Challenges are the 2022 part of the MetaLearn series of Data Challenges run by Chalean, a non-for-profit organization lead by Isabelle Guyon (Inria TAU, TAILOR partner #3). The following text recalls the context in which these two Data Challenges were run - it is cut-and-pasted from Deliverable 2.3 for the sake of completeness.

History

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.

TAILOR Contribution

The lead of the organizing team of both series of Data Challenges was Isabelle Guyon, from Inria TAU (**TAILOR partner #3**). Several other members of the TAU team participated, as well as many other volunteers from the community (see e.g., the list of authors of the paper about Cross-Domain MetaDL at NeurIPS 2022), in general not related to TAILOR, with the exception of Joaquin Vanschoren, from Eindhoven Technische Universiteit - TUE (**TAILOR partner #12**) for the Cross-Domain challenge. Also, Leiden University (**TAILOR partner #7**) participated in the previous round of Cross-Domain MetaDL, and in particular were among the co-authors of the NeurIPS 2021 Competition Track paper that helped design this second round.

Another contribution of TAILOR in these Data Challenges was the money for **all prizes offered to the winners**, all funded by TAILOR on Inria TAU budget. Indeed, these Data Challenges were pure academic, and no commercial company was willing to fund them.

Meta Learning from Learning Curves 2 (MetaLearn 2022)

Description

This academic Data Challenge was described in detail in Deliverable 2.3. We only summarize here its salient points, and what happened since July 2022, i.e., the publication of the results.

The context of the Meta Learning from Learning Curves challenge is that of a portfolio of learning algorithms / hyperparameters: it is then possible to run in parallel several of them and to dynamically decide after every evaluation which one to try next, choosing between exploitation (continue with the current best performing) or exploration (try some yet untested algorithm).
Whereas the first round of this Data Challenge attempted to learn from classical learning curves performance vs learning time (see Figure below, left), the second round used the performance vs dataset size curves (Figure below, right). The question the Data Challenge tries to answer is: learning on a fraction of the whole dataset can be viewed as a proxy for full learning - but how good a proxy is it, and how small can the training dataset be without excessively damaging the performance?

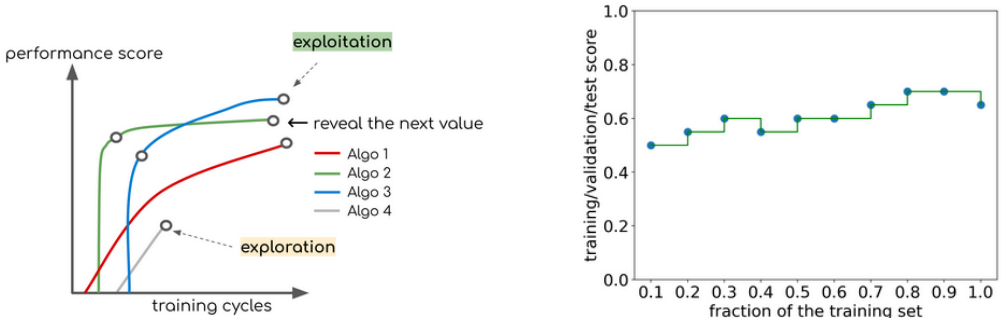


Figure: Example of learning curves for round 1 and round 2 of *MetaLearning from Learning Curves* challenges.

Data

For meta-training, available data consists of a meta-dataset of pre-computed learning curves of 40 algorithms (K-Nearest Neighbors, Multilayer Perceptron, Adaboost, Stochastic Gradient Descent) with different hyperparameters on 30 datasets used in the AutoML challenge, augmented with meta-features of datasets and hyperparameters of algorithms.
The standard data split of the AutoML challenge was used to produce three sets of learning curves for each task, from the training, validation, and test sets.
Also, a new synthetic meta-dataset was generated that contains 12000 learning curves as in FIgure above - right.

For meta-testing, the program must suggest which algorithm from the portfolio to use, and the amount of training data to evaluate the algorithm on a new task (dataset)

efficiently. The agent observes information on both the training learning curve and validation learning curve to plan for the next step.

Evaluation

The scoring program automatically chooses the best algorithm at each time step (i.e. the algorithm with the highest validation score found so far) to compute the test learning curve (as a function of time spent). The metric used for ranking on the leaderboard is the Area under the agent's Learning Curve (ALC).
Note that the final ranking was obtained by running the algorithms with 3 difference random seeds (1, 2, and 3) and taking into account only the **worst** result, in order to favor the robustness of the algorithms.

Results

Interestingly, only one team was able to outperform, on average ALC over the 15 new datasets, the provided baseline that was using DDQN (Double Deep Q Network, now a routine reinforcement learning algorithm). However, they only got the best ALC on 6 datasets, while DDQN also performed best on 6 other datasets.
Nevertheless, the three best teams (including two who failed to outperform the baseline) received prizes of 500€, 300€, and 200€ respectively, funded by TAILOR.
Detailed results and a short description of the winning approaches are available in the final report of the Data Challenge.

Calendar

- **May 16, 2022**: **Public phase**, on public meta-dataset.
- **May 23, 2022**: **Development phase**, the submissions were meta-trained and meta-tested on 15 hidden datasets.
- **July 4, 2022**: **Final/test phase**, the last submission of each participant was meta-learned and meta-tested on 15 fresh hidden datasets, never seen before, giving the final ranking of competitors.
- **July 11, 2022**: **End of competition**, start or the **Legacy Phase**: all test datasets became public, and the Data Challenge became an Open benchmark.
- **July 15, 2022**: winners have been announced at AutoML conference.

Links

- The Codalab web page of the Data Challenge.
- An ArXiV paper describing the setup of this Data Challenge (and the results of the first round).
- Results - a public report

Cross-Domain MetaDL (MetaLearn 2022)

Description

This academic Data Challenge was described in detail in Deliverable 2.3. We only summarize here its salient points, and what happened since July 2022, i.e., the end of the competition and the results.

The context is that of Computer Vision, and the first round MetaDL challenge focused on transferring knowledge between tasks of the same domain so only small data is needed to learn new tasks (aka within-domain few-shot learning). The aim was to efficiently learn N-way (number of classes in a task) k-shot (number of examples per class) tasks, for given N and k. This second competition challenges the participants to solve “any-way” and “any-shot” problems drawn from various domains chosen. Last but not least, these domains were chosen for their humanitarian and societal impact (healthcare, ecology, biology, manufacturing, …).

Data

A meta-dataset (set of datasets) called the Meta-Album has been gathered for few-shot learning and meta-learning beyond this Data Challenge. It is made available through the OpenML platform, maintained by TU Eindhoven (TAILOR partner #12).
At the time of the challenge (but the Meta-ALbum is intended to be continuously updated and augmented, both on the data side and on the algorithmic side), it contained 40 datasets from 10 domains, uniformly formatted as 128x128 RGB images, carefully resized with anti-aliasing, cropped manually, and annotated with various meta-data. Also available on OpenML are the codes and the results of some baseline algorithms.
This Data Challenge used only 30 datasets containing 40 images per class, grouped into three sets (Set-0, Set-1, and Set-2) of ten datasets each, one from each domain. The final test datasets were new to the meta-learning community and had not been previously included in any past meta-learning benchmarks. See the Meta-Album web site for more information (or to retrieve code and images).

Evaluation

Here, the number of classes N in the meta-test tasks ranges from 2 to 20, the training set contains k from 1 to 20 labeled examples per class, and the test set contains 20 unlabeled examples per class. Furthermore, since this Data Challenge focuses on cross-domain meta-learning, all the tasks are carved out from a meta-dataset that contains multiple datasets from ten domains.

During the Feedback Phase, each submission was meta-trained on Set-0 and evaluated on 1 000 any-way any-shot tasks from Set-1 (100 tasks per dataset). The participants received feedback per dataset. During the Final Phase, the last submission of all participants was meta-trained on Sets 0–1 and evaluated on 6 000 any-way

any-shot tasks carved out from Set-2 (600 tasks per dataset). The **worst** of three runs with different random seeds was retained to rank the candidates.
Because of the any-way any-shot aspect of the challenge, the balanced accuracy (bac), also known as macro-averaging recall, normalized with respect to the number of ways N, was used as the evaluation metric (again, all details in Deliverable 2.3, or in the NeurIPS paper describing the results).

Results

The Data Challenge proposed 5 different Leagues to apply, with 3 of them meant to encourage poorly represented communities (New in ML, Women, poorly represented countries), that we will not discuss here. The two main leagues were the Meta-Learning League (no pre-trained backbone was allowed), and the Free-Style League, in which, as the name says, everything was allowed w.r.t. pre-trained backbones. The rules (and hence the results) for the other 3 Leagues were that of the Free Style League (see below) Also, six different baselines were proposed, and trained in both of the above Leagues.

Surprisingly, though 47 teams made more than 1000 submissions during the Feedback Phase, only 4 succeeded in outperforming the baselines provided by the challenge in the Free-Style League, and none in the Meta-Learning League (some teams did not even submit there)!

Whereas the baseline Prototypical Networks remained the best performing in the Meta-Learning League (i.e., learning from scratch), some prizes were nevertheless given to the best performing candidates of that League.
More precisely, in each League, three cash prizes were planned to be awarded to the best 3 best performing teams, 400$, 250$ and 150$ respectively (all prizes funded by TAILOR on Inria TAU budget). In the Free-Style League, the 4 teams that outperformed the baseline were officially ranked, (with cash prize for the top 3). In the Meta-Learning League, 2 prizes were awarded, going to the teams ranked respectively 4th and 3rd in the Free-Style. The prizes in the other 'incentive' Leagues were shared among the same 4 winning teams, weighted by the proportion of the team members that belong to the target category of the League, resulting in respectively 2, 1 and 3 prizes awarded in the New in ML, Women and the Poorly Represented Countries.

Calendar

- **June 15, 2022: Public phase,** using 10 public datasets
- **July 1, 2022: Feedback phase,** using 10 hidden datasets.
- **Sept. 1, 2022: Legal phase**, the last submissions of each participant are ranked on 10 new hidden datasets.
- **Oct. 1, 2022: End of competition**, the winners are announced, and invited to publish their approach at NeurIPS Competitions workshop.
- **Oct. 1, 2022**: **Legacy Phase**, all datasets have been made publicly available, and the Data Challenge became an Open benchmark.

Links

- The Codalab page of the Data Challenge
- A didactic tutorial that runs on Google Colab
- An ArXiV paper describing competition design and baseline results
- The results presented at NeurIPS 2022

Brain Age Prediction Data Challenge

Description

Brain age prediction has emerged as a valuable tool in neuropsychiatry, with discrepancies between predicted and chronological age associated with various psychiatric and neurological conditions, even in their early stages. While structural and functional magnetic resonance imaging (MRI) data have yielded accurate predictions in adults, the application of this approach to young subjects using electroencephalography (EEG) remains largely unexplored. EEG offers significant advantages in terms of affordability, accessibility, and tolerability for younger populations, which could facilitate widespread adoption in developmental studies.

This industrial Data Challenge aims to investigate the feasibility and accuracy of brain age prediction in young subjects (under 25 years) using EEG data. We hypothesize that EEG-based models can provide valuable insights into neurodevelopmental trajectories, potentially identifying deviations from typical brain maturation patterns. The goal is to apply machine learning techniques to extract age-relevant features from EEG recordings and develop robust prediction models tailored to this younger age group. The results of this Data Challenge could pave the way for more accessible and youth-friendly brain age estimation tools, potentially improving our understanding of normal and atypical brain development. Furthermore, this approach may contribute to early detection of neurodevelopmental disorders and inform targeted interventions during critical periods of brain plasticity.

Data

The dataset for the Brain Age Prediction Data Challenge consists of resting-state EEG recordings from a diverse group of participants. Each subject is represented by time series of EEG data, collected during both eyes-open and eyes-closed conditions. In total, the dataset includes EEG recordings from more than 2000 participants across four international cohorts, representing diverse countries and cultural contexts. Specifically, the training dataset comprises 2400 raw EEG files (in MNE format) from 1200 subjects. Each recording contains data from 129 electrodes, sampled at 500 Hz. The duration of recordings is 20 seconds for the eyes-open condition and 40 seconds for the eyes-closed condition.

For evaluation purposes, the dataset is divided into three parts: the training set (1200 subjects), a public test set (400 subjects) used for the public leaderboard, and a private test set (400 subjects) used for the final ranking. This structure allows for robust model development and fair evaluation of the participants' algorithms.

Evaluation

The Data Challenge is framed as a regression problem, where the goal is to predict the age of participants based on their EEG recordings. Submissions are evaluated using the Mean Absolute Error (MAE) metric[1]. MAE measures the average magnitude of

errors between the predicted ages and the actual ages, providing a straightforward and interpretable assessment of model performance. Lower MAE values indicate more accurate predictions, and thus, better-performing models.

$$MAE(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

In addition to submitting their result file evaluated with this metric, participants had to share their code with a description of their approach, through a gitlab or github repository. These requirements aimed to promote reproducibility, transparency, and the sharing of innovative methodologies, fostering a collaborative and scientifically rigorous environment.

Results

36 teams competed for the rewards announced on this industrial Data Challenge: 1st place: 1000$, 2nd place: 500$, 3rd place: 250$, jury prize - 250$.

As is often the case in Data Challenges, some teams overfitted the public test set, therefore the teams leading the public leaderboard were not the ones winning on the final private test set. For instance the first and second on the public leaderboard don't appear in the final top 3 ranking, while the winner "tsneurotech" was only 10th on the public leaderboard.

The final ranking:
1st place: tsneurotech (MAE score: 1.156811)
2nd place: MethodA, team State++ (MAE score: 1.600948)
3rd place: thatsvenyouknow (MAE score: 1.603094)
jury prize: robintibor (MAE score*: 1.4453888)

Calendar

- **November 4, 2022: Development phase starts.** feed-back was provided on the public leaderboard, displaying results on the validation set only.
- **November 18, 2022: Final phase:** Only one submission on final dataset.
- **November 24, 2022: Competition Ends.** Final results have been released.

Links

- The Codalab web site of the competition, with the leaderboard of both development and final phases.
- The website of NeuroTechX Global Hackathon 2022

TAILOR Contribution

The Brain Age Prediction Data Challenge was a key component of the NeuroTechX Global Hackathon 2022, an international event organized by NeuroTechX, a non-profit organization dedicated to advancing neurotechnology and fostering a global, interdisciplinary community with the goal of making neurotechnology more inclusive and beneficial for all of society. The hackathon adopted a hybrid format, taking place online and simultaneously across five locations: Paris, Barcelona, Belgrade, San Francisco, and Munich.

Inria TAU (TAILOR Partner #3) and TAILOR joined forces to propose and design this Data challenge, which formed one of the three official tracks of the global hackathon. Inria involvement underscores the challenge's scientific rigor and its alignment with cutting-edge AI and neurotechnology research.
The collaboration extended to partnerships with TimeFlux (an open-source framework for real-time biosignal processing), CogLab (a cognitive science research lab), and École 42 (an innovative French coding school where the Hackathon took place), further enriching the challenge resources and expanding its reach within the tech and research communities.
This initiative exemplifies how academic institutions, research networks, and industry partners can come together to advance the field of computational neuroscience and promote the development of AI applications in healthcare and neurotechnology.

Sleep States

Description

Electroencephalography (EEG) is a powerful, non-invasive technique for recording brain electrical activity, offering high temporal resolution and real-time data crucial for both clinical and research applications. It plays a vital role in clinical settings for diagnosing and monitoring neurological disorders such as epilepsy, sleep disorders, and brain injuries. It's also essential for assessing brain function in patients under anesthesia or in comas. The real-time nature of EEG data enables clinicians to make rapid, informed decisions about diagnosis and treatment.
In cognitive neuroscience research, EEG is instrumental in exploring various cognitive processes, including attention, perception, memory, and emotional responses to stimuli. This provides valuable insights into the neural mechanisms underlying human cognition and emotional processing.

In this Data Challenge, participants are tasked with developing machine learning models to accurately predict sleep states using EEG data collected from IDUN Guardian Earbuds. This Data Challenge addresses the growing need for accessible, consumer-grade BCI devices capable of providing reliable sleep monitoring and analysis.

Successful models from this competition could contribute to:

- Improved sleep disorder diagnosis and treatment monitoring
- Development of more accurate and user-friendly consumer sleep tracking devices
- Advancement of BCI technology for various applications beyond sleep analysis
- Enhanced understanding of neural correlates of different sleep stages

By bridging the gap between clinical-grade EEG analysis and consumer-friendly devices, this Data Challenge aims to democratize access to sophisticated brain monitoring tools and expand our understanding of sleep neurophysiology.

Data

The Data Challenge utilizes data collected from IDUN Guardian Earbuds, an innovative in-ear EEG device featuring two dry electrodes in each ear. With a sampling rate of 250Hz, these earbuds capture high-quality EEG data in a non-invasive, user-friendly manner. Participants had access to both raw and filtered EEG recordings, accompanied by labels derived from the Guardian system and a validated reference system for cross-validation
.
The dataset encompasses a comprehensive set of sleep-related markers, enabling detailed analysis of various sleep phenomena:

1. Sleep Spindles (SS): Binary indicator (0 or 1) of these brief bursts of oscillatory brain activity.
2. K-Complexes (K): Binary marker (0 or 1) for these sudden, sharp waveforms.

3. Rapid Eye Movements (REM): Binary signal (0 or 1) indicating periods of rapid eye movement.
4. Sleep Onset (Son): Binary markers (0 or 1) denoting the beginning of sleep periods.
5. Sleep Offset (Soff): Binary markers (0 or 1) denoting the end of sleep periods.
6. Arousals (A): Binary indicator (0 or 1) of brief awakenings or shifts to lighter sleep.
7. Microsleep (MS): Binary marker (0 or 1) for very brief episodes of sleep.

This rich dataset allows participants to develop and test machine learning models for accurate sleep state prediction, potentially advancing both clinical sleep analysis and consumer sleep tracking technologies.

Evaluation

The Data Challenge is structured as a multi-label binary classification problem. Participants are required to develop models that can simultaneously classify seven distinct sleep-related markers for each segment of the EEG recording.
Submissions are evaluated using a weighted multi-label F1 score. This metric is chosen to:

- Account for potential class imbalances in the dataset
- Provide a comprehensive assessment of model performance across all seven markers
- Balance precision and recall in the predictions

The F1 score is calculated for each marker individually, then a weighted average is computed to produce the final score. The weighting reflects the relative importance or prevalence of each marker in the dataset.
This evaluation approach ensures that models are assessed on their ability to accurately identify all sleep-related phenomena, providing a robust measure of performance in this complex, multi-faceted classification task.

As in 2022, in addition to submitting their result file evaluated with this metric, participants had to share their code with a description of their approach, through a gitlab or github repository. These requirements aimed to promote reproducibility, transparency, and the sharing of innovative methodologies, fostering a collaborative and scientifically rigorous environment.

Results

The winning team, among nine active participating teams, has reached a F1 score of 0.55, combining approaches from signal processing with machine learning.

Calendar

- **December 2, 2023: Development phase starts.** feed-back are provided on the public leaderboard, displaying results on the validation set only.

- **January 12, 2024: Final phase:** Only one submission on final dataset.
- **January 23, 2024: Competition Ends**. Final results are released.

Links

- The Codabench web site of the competition, with the leaderboard of the development phase.
- github repository of the winning team

TAILOR Contribution

Like for the Brain Age Prediction Data Challenge in 2022, the Sleep States Data Challenge was included as an official track of the NeuroTechX Global Hackathon 2023, an international event organized by NeuroTechX, again the hackathon adopted a hybrid format, but taking place online and simultaneously across eight locations this year: Paris, Vienna, Zurich, Delhi, Kharagpur, Brasilia, Buenos Aires, London.
Once again, Inria TAU (TAILOR Partner #3) and TAILOR joined forces to propose and design this Data Challenge, which formed one of the four official tracks of the global hackathon. One track dedicated to ethics was added in 2023. Inria involvement underscores the Data Challenge's scientific rigor and its alignment with cutting-edge AI and neurotechnology research.
The collaboration extended to partnerships with TimeFlux (an open-source framework for real-time biosignal processing), CogLab (a cognitive science research lab), and École 42 (an innovative French coding school, where the Hackathon took place), further enriching the challenge resources and expanding its reach within the tech and research communities.
This initiative exemplifies how academic institutions, research networks, and industry partners can come together to advance the field of computational neuroscience and promote the development of AI applications in healthcare and neurotechnology.

Automated Crossword Solving

Description

Crossword puzzles are an intriguing challenge for humans because of the complexity and nuances of natural language. Solving a crossword puzzle requires different abilities, that include language understanding, access to general knowledge, and reasoning to solve cryptic clues and/or define an optimal grid-filling strategy to exploit the grid constraints to select the correct word among a set of candidates. Language analysis is needed for understanding clues, finding the correct meaning requires inferences and reasoning that make use of knowledge from various sources. An automated crossword solving agent combines advanced Artificial Intelligence techniques, such as Natural Language Understanding, Question Answering and Constraint Satisfaction to identify the candidate words for each clue, evaluate them and insert the answers into the grid.

The CINI (TAILOR partner #37) developed an open software platform (Webcrow 2.0) that allows the integration of experts for crossword solving. The process involves input management (clues, grid), multiple candidate answer generators, candidate list merging, and grid filling. The modules communicate through a redis pub/sub messaging backbone. The Webcrow project also includes a web platform, webcrow arena, for organizing human-vs-AI competitions.

The Data Challenge consists of developing new modules designed to improve the solving performances of the Webcrow 2.0 agent. The set of modules used to evaluate the base system in different human-vs-AI competitions includes experts for various languages (Italian, English, French), implementing clue database search, rule-based solvers, knowledge graph query, web search, candidate list merging, and optimal grid filling.

Data

The material available for the Data Challenge consists of the Webcrow software library available on GitHub and the clue-answer datasets for Italian (127k unique clue-answer pairs) and English (3.1M unique clue-answer pairs).

Evaluation

The base Webcrow agent has been evaluated against human solvers in public competitions organized in 2022-2023 (WCCI23, WAICF23, TAILOR workshop, 3rd TAILOR conference). The evaluation score is based on the number of correct answers/letters inserted into the grid and on the time to complete the puzzle. The puzzles and the results are available on the competition platform webcrow arena.

Calendar

- **January 12, 2023: Intelligence artificielle VS Humain: pouvez-vous rivaliser avec WebCrow?** The first Data Challenge was organized in INRIA Sophia (details)
- **February 9, 2023: Competition in WAICF** Three Data Challenges were organized during the conference, Italian, French and American Crossword were proposed.
- **February 24, 2023: Intelligence artificielle VS Humain: pouvez-vous rivaliser avec WebCrow? Défi de mots croisés en Français - Tailor Workshop.** Italian, and American Crosswords were also proposed.
- **June 5, 2023**: 3rd Tailor conference. Data Challenge presentation

Links

- The main WebCrow page
- The competition platform webcrow arena
- The WCCI'2022 competition page
- The github link for the Webcrow 2.0 agent platform
- Clue-answer datasets (English, Italian)
- An ArXiV paper detailing the French version

TAILOR contribution

The Data Challenge was organized and run by CINI, **TAILOR partner #37**, Prof. Marco Gori's team at Siena University. It was not run on the Codalab/Codabench platform. It was presented at the third TAILOR conference in Siena (June 2023).

Mind the Avatar's Mind

Description

This Data Challenge asks data science problems in the context of urban energy sustainability. More precisely, the focus of the challenge is on sensor data from a smart-building located on a university campus. The building is a so-called multi-tenant building, which means that it is used by different types of organizations.
The building consists of multiple floors. Each floor contains different types of rooms. There are rooms used for lectures, rooms for project meetings and rooms that can be used for demonstrations of projects. Rooms are grouped in zones. Some of the zones, such as the main entrance hall, the stairs, the canteen and the toilets, are public zones. These zones can be accessed by everyone. A few of them, such as, e.g., the demonstration area on the ground floor, are used as a walk-through to other zones. There are also private zones, which can be accessed only by a single organization. Each zone contains one or more sensors, grouped in boxes and mounted on ceilings or walls. They provide data on temperature, movement, light, and/or CO2.

Sometimes topological information about sensors in a building happens to be incomplete or outdated. It can be that sensors are re-mounted, rooms could have been splitted, or the administration fails to be precise. In such situations it can happen that one does have one does have data coming from the sensors captured somewhere in the measurement database.

The Data Challenge addressed the following questions:

1. Can we complete missing information in the sensor-overview by means of deriving patterns in sensor data and identify the zone in which the unknown sensors are mounted?
2. Can we predict the occupancy in an efficient and reliable way for the lecture zones for different times of the day?

Although the data from building sensors indicates a notion of occupancy, there is neither a ground truth nor a unit of which we can express the sensor data directly into an amount of occupancy. One has to find a way to translate sensor data into a useful interpretation of occupancy(-density).
How can we interpret and translate the building sensor data in terms of occupancy?

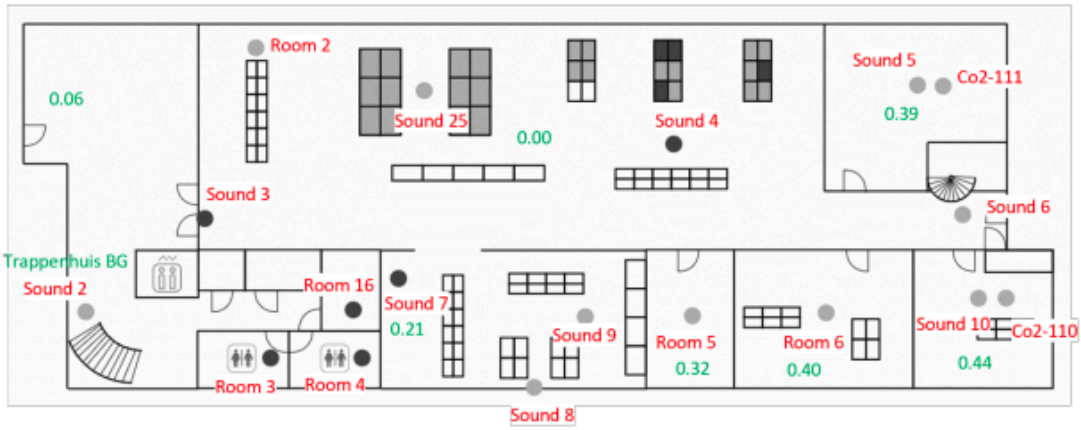
The Data Challenge/hackathon described here is a first step toward answering the long-term goal of this research: predict energy consumption based on such sensor data, and make recommendations that can lead to reduction of the energy consumption, contributing to the efficient and sustainable use of smart buildings.

Data

Participants were provided with a floorplan, sensor-overview, and sensor data. There were 4 types of sensors: basic, CO2, sound and door. The basic, sound and CO2 sensors send data on temperature, humidity, light and motion. CO2 (resp. sound)

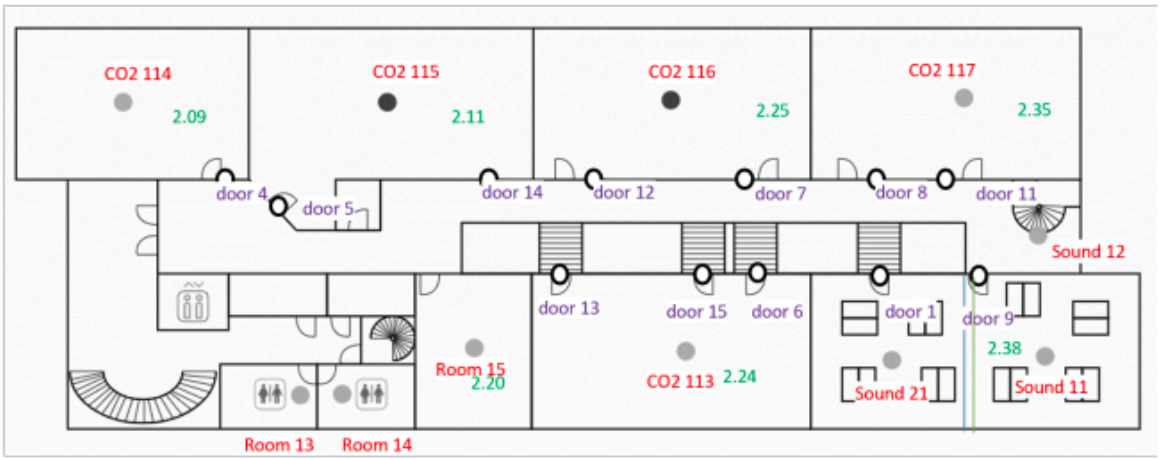
sensors have extra data such as CO2 (resp. sound levels). The door sensors indicate whether the door is closed or open.
The data was anonymized and prepared for the setting of three hackatons moments, where teams can explore, learn and compete in finding solutions to the research questions. The figure below is an example of a floorplan of the 3 levels of the building with different sensors (with their Id in red).

Ground floor



First floor

Second floor



Sensor overview linked the sensors Id and their location, but some locations were “unknown”.
Sensor data included one year recordings of all sensor measurements every 10mn. Unfortunately, there were quite a few impossible values (e.g., temp>1000), and most sensors experienced some periods of “black-out” during which no data was recorded (from 1-3 days to up to 20 days.

Evaluation

The teams were asked to provide prediction models for the occupancy using data from building sensors (movement, temperature, light, sound).

While exploring and learning, the competitive element (the Data Challenge) for 25 participants, which divided themselves into a few groups, was to find the missing sensors, and have a motivation about how data from the sensors can be possibly used to estimate occupancy.

Results

The Data Challenge involved mixed mode competition where discussions and presentations were plenary with all the teams, whereas there was a competitive element in the form of a prize for the best individual team.
The people worked in groups, and challenged themselves to develop algorithms that could pinpoint and repair missing data in incomplete sensor data and/or floor plans of buildings.

On the 3rd evening a winning group was appointed by a small jury. Their results included a good result on the question about completing the sensor map, locating the unknown positions of the sensor, and on the quality of the presentation they gave to the group.

Various datascience approaches were used to cluster data and learning predictive models.

To summarize:

- First, an intense manual data analysis was performed to identify the patterns in building occupancy (yearly, weekly, daily).
- The identification of the zone where sensors of unknown location were situated was done via a huge correlation analysis, based on the idea that sensors with highly correlated data are likely to be in the same room, or at least in the same zone
- Regarding the prediction of occupancy: sensor data could be predicted using a logistic regression model with sliding window. Occupancy was not a hard labeled type of data. The available data was not labeled for such a task, and

though they started labeling it by hand ‘occupancy’ could only be given in a motivational manner, not as a quantitative result from the predictive model.

Links

- Detailed report of the winning solution with steps and graphs is provided here: Mansur Nurmukhambetov (nomomon.github.io)
- Similar data is available on the Kaggle platform.

Calendar

The Data Challenge was spread over 3 evening-workshops (Feb/April 2023), during which the organizers were sitting with the teams in the building itself (in that sense, this Data Challenge is some kind of hybrid challenge/hackathon more than a pure offline Data Challenge):

- **March 16, 2023**: 1st evening 17:00 - 21:00
- **March 30, 2023**: 2nd evening 17:00 - 21:00
- **April13, 2023**: 3rd evening 17:00 - 21:00

TAILOR contribution

The academic Data Challenge was organized by TNO (TAILOR partner #39) and DFKI (TAILOR partner #26), in collaboration with the Hanze university of applied sciences in the Netherlands and the company AIMZ. It was not run on the Codalab/Codabench platform. It was presented at the third TAILOR conference in Siena (June 2023).

Machine Learning for Physical Simulations
ML4PhySim and ML4CFD Data Challenges

Brief history

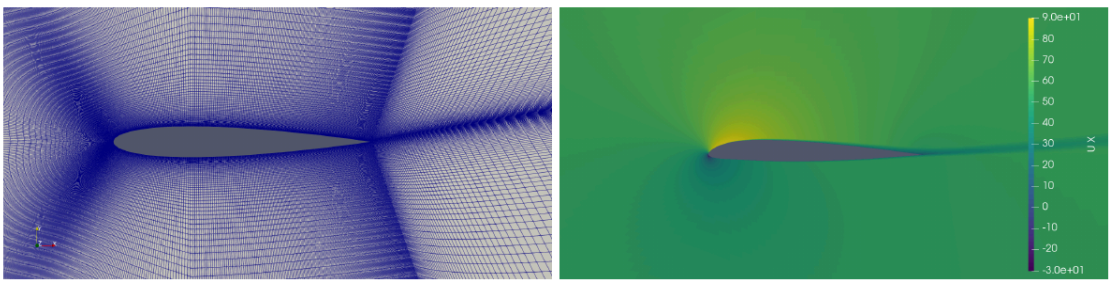
IRT-SystemX is a public institute for industrial maturation and transfer, with a long collaboration history with TAILOR partner #3 Inria. IRT-SystemX, together with several academic (including Inria TAU) and industrial (including NVIDIA, RTE and Criteo) partners, organize Data Challenges to promote the use of Machine Learning-based surrogate models to numerically solve physical problems, through several industrial use cases from their members.

The first attempt was submitted to NeurIPS 2023 Competition track. It included three use cases, the airflow around an airfoil, the current circulation in a Power Grid (context similar to that of the L2RPN challenge above), and the rolling of a tire on a flat surface. Unfortunately, this Data Challenge was not accepted as a NeuIPS challenge. We decided to nevertheless run this Data Challenge, but taking into account the reviews received at NeurIPS 2023, whose main criticism was that three use cases were too many, making it hard to contribute for the average user.

This gave birth to the ML4PhySim challenge, which ran between November 2023 and March 2024. And it was so successful (126 registered teams, and 1165 submissions) that a follow-up was proposed, named ML4CFD, that was accepted to NeurIPS 2024 Competition track. The task, the data, and the evaluation functions are the same, than in ML4PhySim except that the baseline for ML4CFD is the winner of the 2023 ML4PhySim, putting the bar rather high. The following of this Section will present the scientific context of both Data Challenges, but only ML4PhySim results will be briefly surveyed (here, and in Deliverable 2.4) because the new ML4CD challenge has not even started yet (kick-off on July 1st).

Description

Numerous numerical methods exist in the field of Numerical Simulations, based on a discretization of the domain, and different approximations of the Navier Stokes equations that govern the fluid flows. All compute the different physical quantities on a mesh - as illustrated in the Figure below: the mesh is given in input, and the velocity field is the output of the simulation.



(a) Input (b) Output

But such simulations are computationally very costly, and Machine Learning techniques can be used to train a model mimicking the numerical simulations in a fraction of its computational cost. However, such pure data-based models do not take the physics of the problem into account, and the surrogate solutions, though seemingly accurate in terms of the computed fields, completely lack elementary physical properties like mass or energy conservation. Furthermore, one known weakness of data-based models is their poor performance when it comes to Out-of-Distribution test cases, i.e., when the test examples do not come from the same distribution than the training examples. The goal of this Data Challenge is to propose data-based models that also accurately approximate meaningful physical quantities, and that behave reasonably well when facing OoD examples.

The chosen domain is that of Computational Fluid Dynamics (CFD), more precisely the flow of a fluid around an airfoil. The task consists in predicting the incompressible steady-state two-dimensional fields and the force acting over airfoils in a subsonic regime. The ultimate goal of this CFD problem is to find the airfoil that maximizes the lift-over-drag ratio and predict the velocity and pressure fields around it accurately.

Data

The public training dataset is the AirFrans dataset described in the NeurIPS (dataset and benchmarks track) paper, made of 1000 CFD simulations of steady-state aerodynamics over two dimensions airfoils in a subsonic flight regime (5 real values at every point of the point cloud defined by the mesh on the simulation domain), and the participants have access for their simulations to the LIPS (Learning Industrial Physical Simulation) platform described in the NeurIPS (dataset and benchmarks track) paper. The task is to build surrogate models of 5 standard fields of CFD for new airfoils, including Out-ot-Distribution cases, and the evaluation is a mix of accuracy (MSE), computational cost, and, last but not least, respect of the physical constraints (Navier-Stokes equations).

Evaluation

The evaluation is a multi-criteria problem, and the following criteria were aggregated:

- ML accuracy: the accuracy of the five predicted fields on test examples (from the same distribution than the training examples);
- ML speedup: the speedup when compared to a state-of-the-art numerical solver (OpenFoam here);
- Physical relevance: the physical quantities of interest are the drag and the lift, that are not directly computed by the model, but can be inferred from the five fields around the airfoil. The Spearman correlations and the mean relative errors for drag and lift are computed for all test examples
- The same metrics are also computed for OoD examples

Furthermore, rather than adding apples and oranges (the quantities defined above are in different incompatible scales), it was chosen to simply transform their values to three possible status, and these status into points, more precisely: Unacceptable (0 point), Acceptable (1 point), Great (2 points), and to visualize the results with colored large

dots (respectively red, orange and green). The points are then translated into a final score using weights (the arbitrary part), but this leads to some easily explainable comparative results: the figure below gives the results of ML4PhySim in a single table, and the strengths and weaknesses of each method appear clearly.

Results

		Criteria category						
		ML-related (40%)		Physics (30%)	OOD generalization (30%)			Global Score (%)
	Method	Accuracy(75%)	Speed-up(25%)	Physical Criteria	OOD Accuracy(42%)	OOD Physics(33%)	Speed-up(25%)	
		$\overline{u}_x$ $\overline{u}_y$ $\overline{p}$ $\overline{\nu}_t$ $\overline{p}_s$		$\overline{C}_D$ $\overline{C}_L$ ρ_D ρ_L	$\overline{u}_x$ $\overline{u}_y$ $\overline{p}$ $\overline{\nu}_t$ $\overline{p}_s$	$\overline{C}_D$ $\overline{C}_L$ ρ_D ρ_L		
	OpenFOAM		1				1	**82.5**
	Baseline(FC)		750				750	**32.85**
Rank	Preliminary Edition : Top 5 solutions							
1	MMGP [13]		27.40				28.08	**81.29**
2	GNN-FC		570.77				572.3	**66.81**
3	MINR		518.58				519.21	**58.37**
4	Bi-Trans		552.97				556.46	**51.24**
5	NeurEco		44.93				44.78	**50.72**

In the table above, OpenFOAM is the reference numerical solver, used to generate all training and test data: it has only green dots because it is the reference for all quantities used in the evaluation! On the other hand, the Baseline provided with the Data Challenge does not perform very well, with only two green dots.

The five lines below, with a rank, are the five winners of the ML4PhySim challenge. Note that the first one, MMGP, and an excellent accuracy, and performs reasonably well on OoD examples. It will be used as a baseline in the ML4CFD challenge, making it difficult to outperform it.
These top five candidates received the money prizes (respectively 3000€, 2000€, 1000€, 500€ and 500€), offered by IRT-SystemX. All candidates were also offered CPU and GPU usage by Exaion and NVIDIA, two of the sponsors. NVidia also offered usage of its Modulus(c) framework to help the participants in designing augmented physical simulators.

Calendar

ML4PhySim

- **Nov. 16, 2023**: Competition kick-off, start of Warmup phase
- **Jan. 11, 2024:** Start of the Development phase
- **March 14, 2024:** Start of the Final phase
- **March 31, 2024** End of competition

ML4CFD (a NeurIPS 2024 competition)

- **July 1, 2024**: Competition kick-off, start of Warmup phase
- **August 4, 2024**: Start of the Development phase
- **October 15, 2024**: Start of the Final phase
- **October 31, 2024:** End of competition

Links

- The ML4PhySim Codabench page
- The ML4CFD Codabench page
- The NeurIPS 2024 Competition Track paper

TAILOR contribution

These industrial Data Challenges are organized by IRT-SystemX, that has a long-lasting history of collaboration with Inria TAU (TAILOR partner #3). Mouadh Yagoubi, IRT lead here, has co-supervised 3 PhDs with Marc Schoenauer in the past. These Data Challenges are run on Codabench, and Sébastien Treguer (from Inria TAU) was instrumental there too. The organizing committee also included other academic staff from Sorbonne Universté, and the industrial sponsors, NVIDIA, Exaion, ANSYS and Criteo Labs.

Conclusions

The present Deliverable aims to present and describe the Data Challenges and hackathons designed and run by, or with significant contribution from, the TAILOR project. A detailed analysis of the results from the perspectives of Trustworthiness (TAI) and the combination of Learning, Optimization, and Reasoning (LOR) will be addressed in the forthcoming Deliverable 2.4: Lessons learned from TAILOR challenges and Hackathons.
In total, nine Data Challenges were successfully conducted under the TAILOR banner, with one still ongoing. These Data Challenges involved varying levels of participation from TAILOR partners, primarily led by Inria, the partner responsible for challenge-related tasks and WP2. The Data Challenges can be categorized as follows:

Four Academic Data Challenges:

- Two Meta-Learning Data Challenges
- WebCrow crossword puzzle Data Challenge
- Mind the Avatar's Mind Data Challenge

plus the failed (from TAILOR point of view) Inductive links prediction Challenge;

Five Industrial Data Challenges:

- Smarter Mobility Data challenge
- Learning to Run a Power Network (L2RPN) Data Challenge
- Brain Age Prediction Data Challenge
- Sleep States Data Challenge
- Machine Learning for Computational Fluid Dynamics (ML4CFD) Data Challenge (ongoing)

This output surpasses the initial Description of Work, which planned for two Data Challenges per year (one academic and one industrial). The TAILOR project did effectively extend this schema into its fourth year, end beyond. It is important to note that TAILOR involvement varied across these Data Challenges, ranging from end-to-end organization within larger teams to active scientific advisory roles.
A key insight from these Data Challenges, as highlighted in Deliverable 2.3, is the difficulty in motivating stakeholders to provide use-cases and corresponding data for meaningful Data Challenge design. This issue is even more pronounced for hackathons, which require more focused target tasks. Despite the recommendations outlined in Deliverable 2.2, there remains a lack of awareness regarding the substantial effort required beyond the initial ideation phase. The most critical hurdle involves making data publicly available, which often necessitates significant human effort and navigating political decisions, especially for industrial or commercial data.
Nevertheless, we take pride in the success of the eight Data Challenges described herein. We remain convinced that challenges and hackathons are vital tools for advancing AI research and applications. These events not only foster innovation but also bridge the gap between academic research and real-world problem-solving, contributing significantly to the field progress.
Moving forward, addressing the challenges of data availability and stakeholder engagement will be crucial for the continued success and impact of such initiatives in the AI community.