RESEARCH CENTRE

Inria Saclay Centre at Université Paris-Saclay

IN PARTNERSHIP WITH:

CNRS, Université Paris-Saclay

2023 ACTIVITY REPORT

Project-Team TAU

TAckling the Underspecified

IN COLLABORATION WITH: Laboratoire Interdisciplinaire des Sciences du Numérique

DOMAIN

Applied Mathematics, Computation and Simulation

THEME

Optimization, machine learning and statistical methods

Contents

Project-Team TAU

Creation of the Project-Team: 2019 July 01

Keywords

Computer sciences and digital sciences

A3.3.3. – Big data analysis

A3.4. – Machine learning and statistics

A3.5.2. – Recommendation systems

A6.2. – Scientific computing, Numerical Analysis & Optimization

A8.2. – Optimization

A8.6. – Information theory

A8.12. – Optimal transport

A9.2. – Machine learning

A9.3. – Signal analysis

Other research topics and application domains

B1.1.4. – Genetics and genomics

B4. – Energy

B9.1.2. – Serious games

B9.5.3. – Physics

B9.5.5. – Mechanics

B9.5.6. – Data science

B9.6.10. – Digital humanities

1 Team members, visitors, external collaborators

Research Scientists

- Marc Schoenauer [Team leader, INRIA, Senior Researcher, HDR]
- Guillaume Charpiat [INRIA, Researcher]
- Alessandro Ferreira Leite [INRIA, Advanced Research Position]
- Cyril Furtlehner [INRIA, Researcher, HDR]
- Flora Jay [CNRS, Researcher]
- Michèle Sebag [CNRS, Senior Researcher, HDR]

Faculty Members

- Philippe Caillou [UNIV PARIS SACLAY, Associate Professor]
- Sylvain Chevallier [UNIV PARIS SACLAY, Professor, HDR]
- Cécile Germain [UNIV PARIS SACLAY, Emeritus, HDR]
- Matthieu Kowalski [UNIV PARIS SACLAY, Associate Professor, from Sep 2023]
- François Landes [UNIV PARIS SACLAY, Associate Professor]
- Beatriz Seoane Bartolomé [UNIV PARIS SACLAY, Associate Professor, from Feb 2023]

Post-Doctoral Fellows

- Birhanu Hailu Belay [Université Paris-Saclay]
- Shuyu Dong [INRIA, Post-Doctoral Fellow]
- Leo Benoit Planche [UNIV PARIS SACLAY, Post-Doctoral Fellow, from Oct 2023]
- Sara Sedlar [UNIV PARIS SACLAY, LISN, from May 2023]
- Lisheng Sun Hosoya [Université Paris-Saclay]

PhD Students

- Anaclara Alvez-Canepa [Paris-Saclay University]
- Bruno Aristimunha Pinto [Université Paris-Saclay, from Mar 2023]
- Nicolas Atienza [Thales, CIFRE]
- Nicolas Bereux [UNIV PARIS SACLAY, from Oct 2023]
- Guillaume Bied [UNIV PARIS SACLAY]
- Eva Boguslawski [RTE, CIFRE]
- Ragansu Chakkappai [CNRS]
- Thibault De Surrel de Saint Julien [Université Paris Dauphine, from Nov 2023]
- Romain Egele [Argone Nathional Laboratory]
- Emmanuel Goutierre [CNRS]

- Isabelle Hoxha [Université Paris-Saclay, until Jul 2023]
- Armand Lacombe [INRIA]
- Wenzhuo Liu [IRT SYSTEM X, until Jun 2023]
- Jean-Baptiste Malagnoux [CentraleSupelec]
- Jazeps Medina Tretmanis [UNIV BROWN, from Sep 2023]
- Apolline Mellot [Inria, co-supervision with MIND team]
- Emmanuel Menier [IRT SYSTEM X]
- Thibault Monsel [UNIV PARIS SACLAY]
- Matthieu Nastorg [INRIA]
- Solal Nathan [UNIV PARIS SACLAY, LISN]
- Adrien Pavao [UNIV PARIS SACLAY]
- Francesco Pezzicoli [UNIV PARIS SACLAY]
- Audrey Poinsot [EKIMETRICS, CIFRE]
- Arnaud Quelin [SORBONNE UNIVERSITE]
- Cyriaque Rousselot [UNIV PARIS SACLAY]
- Vincenzo Schimmenti [CNRS, until Nov 2023]
- Nilo Elias Schwencke [UNIV PARIS SACLAY]
- Haozhe Sun [UNIV PARIS SACLAY]
- Antoine Szatkownik [UNIV PARIS SACLAY]
- Manon Verbockhaven [UNIV PARIS SACLAY]

Technical Staff

- Benjamin Bearce [Université Paris-Saclay, Ingénieur]
- Hande Gozukan [INRIA, with AVIZ]
- Dinh-Tuan Tran [Université Paris-Saclay, Ingénieur]
- Sebastien Treguer [INRIA, Engineer]
- Ihsan Ullah [ChaLearn, Ingénieur]

Administrative Assistant

- Julienne Moukalou [INRIA]

External Collaborators

- Michele Alessandro Bucci [Safran Tech]
- Jean Cury [UNIV PARIS SACLAY]
- Aurelien Decelle [Universidad Complutense de Madri, en détachement]
- Isabelle Guyon [Google, from Sep 2023]
- Anne-Catherine Letournel [CNRS, LISN]
- David Rousseau [CNRS, LAL]
- Maria Sayu Yamamoto [Université Paris-Saclay, from Feb 2023]
- Burak Yelmen [UNIV PARIS SACLAY]

2 Overall objectives

2.1 Presentation

Building upon the expertise in machine learning (ML) and stochastic optimization, and statistical physics of the former TAO project-team, the TAU team aims to tackle **the vagueness of the Big Data purposes**. Based on the claim that (sufficiently) big data can to some extent compensate for the lack of knowledge, Big Data is hoped to fulfill all Artificial Intelligence commitments.

This makes Big Data under-specified in three respects:

- A first source of under-specification is related to **common sense**, and the gap between observation and interpretation. The acquired data do not report on “obvious” issues; still, obvious issues are not necessarily so for the computer. Providing the machine with common sense is a many-faceted, AI hard, challenge. A current challenge is to **interpret the data** and cope with its blind zones (e.g., missing values, contradictory examples, ...).
- A second source of under-specification regards the **steering of a Big Data system**. Such systems commonly require lifelong learning in order to deal with open environments and users with diverse profiles, expertises and expectations. A Big Data system thus is a dynamic process, whose behavior will depend in a cumulative way upon its future environment. The challenge regards **the control of a lifelong learning system**.
- A third source of under-specification regards its social acceptability. There is little doubt that Big Data can pave the way for Big Brother, and ruin the social contract through modeling benefits and costs at the individual level. What are the **fair trade-offs between safety, freedom and efficiency**? We do not know the answers. A first practical and scientific challenge is to first assess, and then enforce, **the trustworthiness of solutions**.

However, several concerns have emerged in the last years regarding Big Data models. First, in industrial context, data is now always big, and many practical problems are relevant to **small data**. On the opposite, when big data is available, the arms race around LLMs has given birth to increasingly big models, involving hundreds of billions of parameters, and environmental concerns are becoming increasingly high, for their training, but even for their use and the inference process.

Our initial overall under-specification considerations, mitigated with the concerns above, have lead the team to align its research agenda along four pillars:

- Frugal Learning, addressing the environmental concerns, in terms of deep network architecture and considering the small data regimes;
- Causal Learning, a grounded way to address the trustworthiness issue by improving explainability of the results;

- Bidirectional links with Statistical Physics, to better understand very large systems and improve their performances, both in terms of accuracy of the models and energy consumption in their use;
- Hybridization of Machine Learning with Numerical Simulations, again aiming to reach better efficiency while decreasing the computing needs.

Last but not least, the organization of challenges and the design of benchmarks, a cornerstone of Machine Learning nowadays, remains an active thread of the team activity, in particular through the Codalab platform and its new version Codabench.

3 Research program

3.1 Frugal Learning

Frugality is a must for machine learning: because of scientific concerns (monster models imply non-reproducible science); because of sustainability concerns (energy consumption to train and use models); because of applicability concerns: in most non-GAFAM/GAMAM settings, we deal with small data, and PhD students not infrequently receive the promised data in the last months of their PhDs.

We target in particular three domains: data frugality, computational complexity at test time (to minimize environmental footprint when using the trained network at large scales), and computational complexity of neural architecture search (i.e. of the automatically finding of neural architectures suitable for a given machine learning task at hand, at training time). The mainstream strategy suggests finding a model in a large (overparameterized) model space, in order to avoid optimization and expressivity issues, and then pruning it [133]. An alternative to the above strategy, named neural network growth, consists in starting from a tiny architecture and grafting additional neurons or layers to extend its representation power on demand, on the fly during training. This raises interesting mathematical questions regarding optimization, generalization, and statistical significance.

An approach we are currently developping is following the preliminary proof of concept in M. Verbockhaven's PhD where we seek to adapt the neural tangent kernel to the directions desired by the functional gradient descent. This kind of approach could be useful not only to automatically (and frugally) design from scratch a neural network architecture suitable for a new task, but could also be of prime interest in classical Neural Architecture Search starting from large networks, providing directly optimal architecture variations instead of searching for them in a computationally-heavy trial-and-error fashion.

A nice byproduct is that by building smaller models, one potentially requires smaller data, and is potentially less prone to overfit. This opens interesting questions regarding regularization in deep learning and advocates for a more reasonable, guided use of combinatorics, that appear through traditional random initializations of numerous neurons (lottery ticket hypothesis [121]).

3.2 Causal Learning

The rise of causal modelling (CM) has an impact on the general agenda of machine learning, more aware of, and more robust w.r.t. the potential and usual differences of data distributions between training and testing times or along lifelong learning. This new agenda focuses on sorting out distribution-independent relations (hopefully causal ones) among the observed features, and other relations, possibly reflecting spurious correlations. The expected benefits of this causality-inspired focus is to deliver learned models that are: i) more robust w.r.t. the non iid setting; ii) more interpretable; iii) possibly humanly verified. The last two properties only hold, naturally, if the features are expressed at a sufficient level of generality.

A key scientific question is whether and how the main lesson of Deep Learning (It's the representation, stupid !) can be ported to causal modelling, particularly so when dealing with raw, redundant and/or high dimensional data. The use of latent variables and structures in e.g. [123, 144, 146] has shown its potential to disentangle root causes (sources of the observed data) and cope with hidden confounders. However, causal modelling comes with the key requirement of identifiability/uniqueness of the learned causal models, that is in general *not* satisfied in mainstream machine learning.

A promising research direction toward model identifiability is to investigate the stability of causal discovery. Formally, one might want that, if data $\mathscr{D}$ yields model $\mathscr{G}$, then data $\mathscr{D}'$ generated after $\mathscr{G}$

yields a model $\mathscr{G}'$ that is in essence same as $\mathscr{G}$. This direction opens to two strategies: i) observing the differences between $\mathscr{G}$ and $\mathscr{G}'$ sheds some light about the diversity in the data with some/no impact on the causal modelling output, i.e. the biases of the causal discovery algorithms; ii) and more deeply, the issue of stability can inspire new learning criteria, enforcing the stability of the causal models under such changes of distribution. Another hot research direction investigates how to improve the interpretability of a model, without degrading too seriously its accuracy. Let us focus on the task of interpreting hidden variables and their interactions. A possible strategy – at the core of the AI2 French-German proposal, 2023-2026; coll. Fraunhofer Bonn – takes inspiration from the Multi-Criteria Decision Aid literature (and the lessons learned in R. Bresson's PhD [96, 97]). The idea is that i) if the last say two layers of a deep net were structured as a hierarchical choquet integral (HCI); ii) and if their input (the nodes in the layer before) were interpretable (giving a feature name to each node), then the black box could be made transparent, expressing sparse hierarchical interactions (HCI) of these features. The first condition can be handled by retraining a trained efficient deep net, and imposing HCI constraints on the last two layers. A pending question is how these constraints would degrade the loss accuracy (depending on the number of would-be features). The second condition will be met by associating a supervised binary learning problem to each node, and involving the expert in the loop (or possibly exploiting textual information about the samples) to solve it.

3.3 Machine Learning with/for Statistical Physics

Concerning the links between statistical physics and machine learning, we are working on both aspect of ML with Statistical Physics and Statistical Physics for ML.

1- The first line of research, based on our expertise on generative models, will be headed toward efficient methods for frugal and interpretable generative models, typically energy based models (EBM)[128]. In particular concerning explainability we will look for physically-inspired interpretable feature extraction processes, exploring the possibilities of using EBMs as data-driven fitness landscapes.

2- This explainability aspect will be actually important for our second axes concerning applications of EBMs in bioinformatics. For instance, given data of protein's families with common ancestors, we expect to be able to learn a model describing the statistics of the family, and then use this model to predict the mutation of the amino-acid. More broadly we will develop methods for direct coupling extraction with RBMs, clustering of data in families and subfamilies, semi-supervised strategies and use EBM for pattern extraction in genomics/proteomics sequence datasets.

3- Our third axes will focus on symmetries both for methods and applications. "It is only slightly overstating the case to say that physics is the study of symmetry" (Philip Anderson 1972), and enforcing symmetries into models or finding symmetries in the data[132] is also key to ML. CNNs can enforce translation equivariance, GNNs enforce permutation equivariance, and more recently, rules for building roto-translation-equivariant networks have been devised[109]. The importance of symmetries has been acknowledged in [107], coining the term "geometric deep learning" to refer to group-invariance aware neural networks. We are working on pushing roto-translation equivariance further, with application to molecular systems or amorphous materials. Furthermore, from statistical physics we know that sytems display scale-invariant distributions at their critical point. Starting from simple avalanche models as benchmarks, we want to design networks that would be genuinely scale-equivariant (or invariant). Applications range from seismic hazard to solar wheather forecast, i.e. any area where large events-related data are scarce. Such networks would de facto perform extrapolation, a rare feature in Machine Learning. This avenue of research is being studied within Anaclara Alvez' PhD (co-supervised by Cyril Furtlehner and François Landes).

4- Our last axes deals with fundamental properties of ML like for instance neural scaling laws[105] and is based on recent theoretical progresses like the formulation of the neural tangent kernel[124] and the lazy regime[108]. Various asymptotic results can be obtained thanks to random matrix theory or replica approaches. Equipped with such tools we would like to explore for instance the learning dynamics beyond the lazy regime, the out-of-equilibrium regimes of EBMs via dynamical mean field theory but also the utility-privacy trade-off with solvable models.

3.4 Machine Learning for Numerical Simulations

Until recently, applying off-the-shelf neural nets to numerical simulations (e.g., approximating the solution of PDEs) could only compete with numerical solvers in a few situations: when the problem is simple and of reasonable size, and when a limited accuracy, that does not need to be guaranteed, is sufficient. For instance, in cases involving chaotic behaviors (e.g. turbulent flows in fluid dynamics), current models fail to fit the target trajectory in the mid to long term. The situation is rapidly evolving (see e.g., GraphCast, by DeepMind [126]), but there remains a need for tighter coupling between ML and simulations.

Building upon TAU expertise in numerical engineering, it is suggested that the diversity of use cases tackled in applications (recent and on-going PhDs of W. Liu, E. Menier, M. Nastorg, E. Goutierre; T. Monsel, and collaboration with the IRT SystemX IA2 program as well as with IFPEN) can lead to formulating general principles and methodology.

One research direction is to consider more structured losses/architectures. This research direction evolves at a rapid pace: from convolutional architectures, to distributional architectures enforcing invariance or equivariance properties [102], to optimal transport based embeddings [141]. It is believed that new losses, aimed at preserving statistical quantities (e.g. high order moments; extreme value exponents), might help to learn and reproduce chaotic data trajectories, better than MSE losses. Nevertheless, until theoretical guidelines are available to the practionner, it is important to be able to experimentally guide and validate users' choices in terms of architecture/loss, for any new use-case. There is today a lack of well-grounded and widely accepted benchmarks, and we contribute to the IRT SystemX LIPS platform (Learning Industrial Physical Simulation benchmark suite) [129], lead by our collaborators from IRT (Mouadh Yagoubi) and RTE (Benjamin Donnot and Antoine Marot).

Another direction of research concerns how the domain know-how can best be conveyed to the learning process: through priors; or warm-starting the solution; or enforcing the required solution properties through specific loss terms; or maybe simply choosing the right training samples.

A theoretically and practically important domain concerns the coupling of an ML model and a numerical simulator, with mutual benefits (compensating for insufficient data; adjusting the simulator hyper-parameters; prioritizing new experiments toward optimal design or model identification; providing a fast sampler; addressing inverse problems). Mimicking the structure of the simulator/the physical phenomenon through the neural architecture helps to guide the optimization, all the more so as it supports the definition of auxiliary losses (e.g. based on internal states of the simulator). Again, the use of auxiliary losses can be very useful, *if* an appropriate learning schedule has been defined (controlling the impact/weight of each auxiliary loss depending on the current state of the model and of the learning trajectory).

Last but not least, unleashing the power of the recently emerged Foundation Models and Transformers resulted in low hanging fruits (e.g., more powerful surrogate models) that have not yet been picked up, and will also open new avenues for hybrid/multidisciplinary research.

3.5 Challenge Organization

In the rapidly evolving field of machine learning (Data-Driven Artificial Intelligence) empirical evaluations of new algorthms to confirm their effectiveness and reliability is even more essential. This trend is intensifying with the increasing complexity of methods, particularly with the emergence of deep neural networks, generative AI, and large language models, which are difficult to explain and interpret. Empirical evaluation is essential, in particular because of the complexity of the algorithms and the unpredictable nature of the data.

The approach taken in ths pillar is that of organizing scientific competitions (also called "challenges"). Scientific competitions systematize large-scale experiments and show the effectiveness of participants in solving complex problems. Annual competitions, organized on the Codalab competition platform, address various scientific or industrial questions, evaluating the automatic algorithms submitted by participants. The newer version of Codalab, called Codabench, extends the capabilities of Codalab to benchmarks.

Both challenges and benchmarks are crucial for comparing models and understanding their behavior. Recent applications include: improving decision-making, particularly useful in fields like finance and medicine; helping to combat climate change by optimizing the use of resources; personalizing the customer experience in e-commerce, banking, and other industries; improving security and preventing fraud; and improving accessibility for people with disabilities, for example, through voice recognition systems, visual aids for the visually impaired, and other assistive technologies.

The importance of impartial evaluations of algorithms is constantly increasing with the acceleration of progress in Artificial Intelligence. According to David Donoho: "The emergence of Frictionless Reproducibility flows from 3 data science principles that matured together after decades of work by many technologists and numerous research communities. The mature principles involve data sharing, code sharing, and competitive challenges, however implemented in the particularly strong form of frictionless open services." He cites the Codalab project as being exemplary in this area [116].

4 Application domains

4.1 Computational Social Sciences

Computational Social Sciences (CSS) studies social and economic phenomena, ranging from technological innovation to politics, from media to social networks, from human resources to education, from inequalities to health. It combines perspectives from different scientific disciplines, building upon the tradition of computer simulation and modeling of complex social systems [122] on the one hand, and data science on the other hand, fueled by the capacity to collect and analyze massive amounts of digital data.

The emerging field of CSS raises formidable challenges along three dimensions. Firstly, the definition of the research questions, the formulation of hypotheses and the validation of the results require a tight pluridisciplinary interaction and dialogue between researchers from different backgrounds. Secondly, the development of CSS is a touchstone for ethical AI. On the one hand, CSS gains ground in major, data-rich private companies; on the other hand, public researchers around the world are engaging in an effort to use it for the benefit of society as a whole [127]. The key technical difficulties relate to data and model biases, and to self-fulfilling prophecies. Thirdly, CSS does not only regard scientists: it is essential that the civil society participate in the science of society [145].

TAO/TAU was involved in CSS for the last five years, and its activities had been strengthened thanks to P. Tubaro's and I. Guyon's expertises respectively in sociology and economics, and in causal modeling. Their departures has negatively impacted the team activities in this domain, but many projects are still on-going and CSS remains a domain of choice (see Section 8.6).

4.2 Energy Management

Energy Management has been an application domain of choice for TAO since the mid 2000s, with main partners SME Artelys (METIS Ilab INRIA; ADEME projects POST and NEXT), RTE (three CIFRE PhDs), and IFPEN (bilateral contract, DATAIA project ML4CFD). The goals concern i) optimal planning over several spatio-temporal scales, from investments on continental Europe/North Africa grid at the decade scale (POST), to daily planning of local or regional power networks (NEXT); ii) monitoring and control of the French grid enforcing the prevention of power breaks (RTE); iii) improvement of house-made numerical methods using data-intense learning in all aspects of IFPEN activities (Section 8.4.2).

The daily maintainance of power grids requires the building of approximate predictive models on the top of any given network topology. Deep Networks are natural candidates for such modelling, considering the size of the French grid (~ 10000 nodes), but the representation of the topology is a challenge when, e.g. the RTE goal is to quickly ensure the "n-1" security constraint (the network should remain safe even if any of the 10000 nodes fails). Existing simulators are too slow to be used in real time, and the size of actual grids makes it intractable to train surrogate models for all possible (n-1) topologies (see Section 8.5 for more details).

Another aspect of Power Grid management regards the real-time control of the grid topology, man-made at the moment. Its automation is yet a difficult challenge, but results on the L2RPN challenge have demonstrated its feasibility with Reinforcement Learning, opening the way to more ambitious goals (e.g., decentralized control via multi-agent Reinforcement Learning, see Section 8.5).

4.3 Data-driven Numerical Modeling

In domains where both first principle-based models and equations, and empirical or simulated data are available, their combined usage can support more accurate modelling and prediction, and when appropriate, optimization, control and design, and help improving the time-to-design chain through fast interactions between the simulation, optimization, control and design stages. The expected advances regard: i) the quality of the models or simulators (through data assimilation, e.g. coupling first principles and data, or repairing/extending closed-form models); ii) the exploitation of data derived from different distributions and/or related phenomenons; and, most interestingly, iii) the task of optimal design and the assessment of the resulting designs.

A first challenge regards the design of the model space, and the architecture used to enforce the known domain properties (symmetries, invariance operators, temporal structures). When appropriate, data from different distributions (e.g. simulated vs real-world data) will be reconciled, for instance taking inspiration from real-valued non-volume preserving transformations [113] in order to preserve the natural interpretation.
Another challenge regards the validation of the models and solutions of the optimal design problems. The more flexible the models, the more intensive the validation must be. Along this way, generative models will be used to support the design of "what if" scenarios, to enhance anomaly detection and monitoring via refined likelihood criteria.

In the application domains described by Partial Differential Equations (PDEs), the goal of incorporating machine learning into classical simulators is to speed up the simulations while maintaining as much as possible the accuracy ad physical relevance of the proposed solutions. Many possible tracks are possible for this; one can build surrogate models, either of the whole system, or of its most computationaly costly parts; one can search to provide better initialization heuristics to solvers, which make sure that physical constraints are satisfied. Or one can inject physical knowledge/constraints at different stages of the numerical solver.

5 Social and environmental responsibility

5.1 Footprint of research activities

The Laboratory (LISN) is currently actively re-thinking its carbon footprint, being part of the Labo1.5 initiative. We participate in working groups about GreenAI (being able to measure, compare and mitigate the negative impact of training and inference for large models). To start changing practices, the simple fact of reporting the cost of training one's model in publications has been spotted as en efficient tool. Ideally, the development cost (all the trainings performed during the research, not just the training of the model presented in the paper) should also be mentionned.

Another axis studied by the lab is the limitation of (aerial) transport, keeping in mind that the younger members should be allowed to build their own research network and foreign experiences.

5.2 Impact of research results

All our work on Energy (see Sections 4.2) is ultimately targeted toward optimizing the distribution of electricity, be it in planning the investments in the power network by more accurate previsions of user consumption, or helping the operators of RTE to maintain the French Grid in optimal conditions.

A collaboration with IDEEV has just started, with the idea of leveraging Deep Learning as a tool to help unlock agro-ecological research. In particular, we aim to help measure the yields in mixed cropping (requiring to be able to classify grains of a given species but of different varieties – something impossible to the naked eye) and detect pollinators on video footage taken outside (including wind, change in light conditions, etc).

6 Highlights of the year

6.1 Awards

- **Best paper award**: Romain Egele, Isabelle Guyon, Venkatram Vishwanath, and Prasanna Balaprakash, for the paper *Asynchronous Decentralized Bayesian Optimization for Large Scale Hyperparameter Optimization* at the 2023 IEEE 19th International Conference on e-Science (e-Science) [43].
- **ACM Ten-Year Technical Impact Award**: Isabelle received the 2023 Ten-Year Technical Impact Award by the committee of the 15th ACM International Conference on Multimodal Interaction. The award recognized the lasting impact of the publication: "Multi-modal gesture recognition challenge 2013: Dataset and Results", which was pioneering work at the time on gesture recognition using multi-modal data (RGB camera + depth camera). The publication was co-authored by Isabelle and appeared in the Proceedings of the 15th ACM International Conference on Multimodal Interaction [120].

7 New software, platforms, open data

7.1 New software

7.1.1 Codalab

Keywords: Benchmarking, Competition

Functional Description: Challenges in machine learning and data science are competitions running over several weeks or months to resolve problems using provided datasets or simulated environments. Challenges can be thought of as crowdsourcing, benchmarking, and communication tools. They have been used for decades to test and compare competing solutions in machine learning in a fair and controlled way, to eliminate "inventor-evaluator" bias, and to stimulate the scientific community while promoting reproducible science. See our news: https://codalab.lisn.upsaclay.fr/highlights.

The new Codalab infrastructure deployed in 2021 includes vast amounts of storage over a distributed Minio (4 physical servers, each with 12 disks of 16 TB) spread over 2 buildings for robustness, and 20 GPU workers in the backend, thanks for the sponsorship of région Ile-de-France, ANR, Université Paris-Saclay, CNRS, INRIA, and ChaLearn, to support 50,000 users, organizing or participating each year to hundreds of competitions.

Some of the areas in which Codalab is used include Computer vision and medical image analysis, natural language processing, time series prediction, causality, and automatic machine learning. Codalab has been selected by the Région Ile de France to organize industry-scale challenges.

TAU continues expanding Codalab to accommodate new needs, including teaching. Check recent student projects: https://saclay.chalearn.org/

URL: `http://competitions.codalab.org`

Contact: Isabelle Guyon

7.1.2 Cartolabe

Name: Cartolabe

Keyword: Information visualization

Functional Description: The goal of Cartolabe is to build a visual map representing the scientific activity of an institution/university/domain from published articles and reports. Using the HAL Database, Cartolabe provides the user with a map of the thematics, authors, and articles. ML techniques are used for dimensionality reduction, cluster, and topic identification, visualization techniques are used for a scalable 2D representation of the results.

Cartolabe has, in particular, been applied to the Grand Debat dataset (3M individual propositions from French Citizen, see https://cartolabe.fr/map/debat). The results were used to test both the scaling capabilities of Cartolabe and its flexibility to non-scientific and non-English corpora. We also added sub-map capabilities to display the result of a year/lab/word filtering as an online generated heatmap with only the filtered points to facilitate the exploration. Cartolabe has also been applied in 2020 to the COVID-19 Kaggle publication dataset (Cartolabe-COVID project) to explore these publications.

URL: http://www.cartolabe.fr/

Publication: hal-02499006

Contact: Philippe Caillou

Participants: Philippe Caillou, Jean Daniel Fekete, Michèle Sebag, Anne-Catherine Letournel, Hande Gozukan

Partners: LRI - Laboratoire de Recherche en Informatique, CNRS

7.1.3 DeepHyper

Keywords: Deep learning, Autotuning, HPC

Functional Description: Machine learning algorithms are continually evolving to serve diverse applications, yet their development often entails a significant trial-and-error process to identify optimal learning pipelines.This is compounded by the multitude of data preprocessing techniques, prediction (or generative) models, and learning procedures available, each offering a range of configurable parameters, also referred to as hyperparameters. DeepHyper addresses this challenge by automating the selection and configuration of algorithms and their corresponding hyperparameters, facilitating a streamlined approach for engineers and scientists to comprehend and optimize the learning pipeline. At its core, DeepHyper employs parallel Bayesian optimization, validated through rigorous testing involving up to 8,000 parallel tasks. This methodology is adaptable for both single and multi-objective tasks, enabling efficient early discarding of costly training steps. Furthermore, DeepHyper seamlessly integrates with various parallel backends, including multi-threading, multi-processing, Clouds (via the Ray library), and MPI-based schedulers on supercomputers, enhancing its scalability and versatility across different computing environments. The development of DeepHyper is supported by the TAU-team through advances in learning theory for improving and explaining its core algorithms.

URL: https://github.com/deephyper/deephyper

Contact: Romain Egele

7.1.4 OmniPrint

Keyword: Open data

Functional Description: Benchmarks and shared datasets have been fostering progress in deep learning. While there is an increasing number of available datasets, there is a need for larger ones. However, collecting and labeling data is time-consuming and expensive, and systematically varying environmental conditions is difficult and necessarily limited. Therefore, resorting to artificially generated data is helpful to drive fundamental research in deep learning. OmniPrint is geared to generating an unlimited amount of printed characters.

Character images provide excellent benchmarks for deep learning problems because of their relative simplicity and visual nature while opening the door to high-impact real-life applications. A conjunction of technical features is required to meet our specifications: pre-rasterization manipulation of anchor points, post-rasterization distortions, natural background and seamless blending, foreground filling, anti-aliasing rendering, and importing new fonts and styles. Modern fonts such as TrueType or OpenType are made of straight line segments and quadratic Bezier curves, connecting anchor points. Thus, it is easy to modify characters by moving anchor points. This allows users to perform vectors-space pre-rasterization geometric transforms (rotation, shear, etc.) as well as distortions (e.g., modifying the length of ascenders of descenders) without incurring aberrations due to aliasing when transformations are done in pixel space (post-rasterization).

The key technical contributions include implementing transformations and styles such as elastic distortions, natural background, foreground filling, and so on, selecting characters from the Unicode standard to form alphabets from more than 20 languages around the world, further grouped into partitions, to facilitate creating meta-learning tasks, identifying fonts, implementing character rendering with a low-level FreeType font rasterization engine, which enables direct manipulation of anchor points, adding anti-aliasing rendering, implementing and optimizing utility code to facilitate dataset formatting. To our knowledge, OmniPrint is the pioneering text image synthesizer geared toward ML research, supporting pre-rasterization transforms, which allows Omniprint to imitate handwritten characters to some degree. More details can be found in the paper (https://openreview.net/forum?id=R07XwJPmgpl, https://arxiv.org/abs/2201.06648).

URL: https://github.com/SunHaozhe/OmniPrint

Contact: Haozhe Sun

7.1.5 codabench

Keywords: Competition, Benchmarking

Functional Description: Obtaining standardized crowdsourced benchmark of computational methods is a major issue in data science communities. Dedicated frameworks enabling fair benchmarking in a unified environment are yet to be developed. Here we introduce Codabench, an open-source, community-driven platform for benchmarking algorithms or software agents versus datasets or tasks. A public instance of Codabench (this https URL) is open to everyone, free of charge, and allows benchmark organizers to compare fairly submissions, under the same setting (software, hardware, data, algorithms), with custom protocols and data formats. Codabench has unique features facilitating the organization of benchmarks flexibly, easily and reproducibly, such as the possibility of re-using templates of benchmarks, and supplying compute resources on-demand. Codabench has been used internally and externally on various applications, receiving more than 3000 users and 25000 submissions.

@article{codabench, title = {Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform}, author = {Zhen Xu and Sergio Escalera and Adrien Pavão and Magali Richard and Wei-Wei Tu and Quanming Yao and Huan Zhao and Isabelle Guyon}, journal = {Patterns}, volume = {3}, number = {7}, pages = {100543}, year = {2022}, issn = {2666-3899}, doi = {https://doi.org/10.1016/j.patter.2022.100543}, url = {https://www.sciencedirect.com/science/article/pii/S2666389922001465} }

URL: https://www.codabench.org/

Contact: Isabelle Guyon

Partner: Région Île-de-France

7.1.6 pyriemann-qiskit

Keywords: Quantum programming, Riemannian geometry, Symmetric positive definite matrices

Functional Description: Literature on quantum computing suggests it may offer an advantage compared with classical computing in terms of computational time and outcomes, such as for pattern recognition or when using limited training sets. Bulding on the Qiskit library on quantum computing, pyriemann-qiskit implements a wrapper around quantum-enhanced support vector classifiers (QSVCs) and variational quantum classifiers (VQCs), to use quantum classification with Riemannian geometry. It also introduces a quantum version of the MDM algorithm, a classifier operating on the manifold of symmetric positive definite matrices.

URL: https://pyriemann-qiskit.readthedocs.io/en/latest/

Publication: https://hal.science/hal-04040814v1

Contact: Sylvain Chevallier

Partner: IBM

7.1.7 pyriemann

Keywords: Riemannian geometry, Hermitian positive definite matrices, Symmetric positive definite matrices

Functional Description: Pyriemann is a Python machine learning package based on scikit-learn API. It provides a high-level interface for processing and classification of real (resp. complex)-valued multivariate data through the Riemannian geometry of symmetric (resp. Hermitian) positive definite (SPD) (resp. HPD) matrices.

pyRiemann aims at being a generic package for multivariate data analysis but has been designed around biosignals (like EEG, MEG or EMG) manipulation applied to brain-computer interface (BCI), estimating covariance matrices from multichannel time series, and classifying them using the Riemannian geometry of SPD matrices. It is widely used in the scientific community with more than one million download.

URL: https://pyriemann.readthedocs.io

Contact: Sylvain Chevallier

7.1.8 braindecode

Keywords: Brain-Computer Interface, Deep learning

Functional Description: BrainDecode is an open-source Python toolbox for decoding raw electrophysiological brain data with deep learning models. It includes dataset fetchers, data preprocessing and visualization tools, as well as implementations of several deep learning architectures and data augmentations for analysis of EEG, ECoG and MEG. It is design for neuroscientists who want to work with deep learning and deep learning researchers who want to work with neurophysiological data.

URL: https://braindecode.org/stable/index.html

Contact: Sylvain Chevallier

Partner: Roche

7.1.9 MOABB

Name: Mother of all BCI Benchmarks

Keywords: Brain-Computer Interface, Open data, Benchmarking

Functional Description: Mother of all BCI Benchmarks (MOABB) allows to build a comprehensive benchmark of popular brain-computer interface algorithms applied on an extensive list of freely available EEG datasets. This is an open science initiative, serving as a reference point for the future algorithmic developments. Build on reference libraries like scikit-learn and MNE-python, machine learning pipelines can be ranked and promoted on a website, providing a clear picture of the different solutions available in the field. This software has 80k downloads and an active international development community.

URL: https://neurotechx.github.io/moabb/

Contact: Sylvain Chevallier

7.1.10 dnadna

Name: Deep Neural Architectures for DNA

Keywords: Deep learning, Population genetics

Functional Description: DNADNA provides utility functions to improve development of neural networks for population genetics and is currently based on PyTorch. In particular, it already implements several neural networks that allow inferring demographic and adaptive history from genetic data. Pre-trained networks can be used directly on real/simulated genetic polymorphism data for prediction. Implemented networks can also be optimized based on user-specified training sets and/or tasks. Finally, any user can implement new architectures and tasks, while benefiting from DNADNA input/output, network optimization, and test environment.

URL: `https://mlgenetics.gitlab.io/dnadna/`

Contact: Flora Jay

7.2 New platforms

Participants: Isabelle Guyon, Anne-Catherine Letournel, Adrien Pavao, Hande Gozukan.

- **CODALAB:** The TAU group is community lead (under the leadership of Isabelle Guyon) of the open-source Codalab project, hosted by Universite' Paris-Saclay, whose goal is to host competitions and benchmarks in machine learning [137]. We have replaced the historical server by a dedicated server hosted in our lab. Since inception in December 2021, over 40000 participants entered 640 public competitions (see statistics). The engineering team, overseen by Anne-Catherine Letournel (CNRS engineer) includes two engineers dedicated full time to administering the platform and developing challenges: Adrien Pavao, financed by a project started in 2020 with the Re'gion Ile-de-France, et Dinh-Tuan Tran, financed by the ANR AI chaire of Isabelle Guyon, Ihsan Ullah, financed by a collaboration with LBNL/CERN and IJCLAB, and Benjamn Bearce financed by the ANR AI chaire of Isabelle Guyon. Several other engineers are engaged as contractors on a needs-be basis. The rapid growth in usage led us to put in place a new infrastructure. We have migrated the storage over a distributed Minio (4 physical servers, each with 12 disks of 16 TB) spread over 2 buildings for robustness, and added 10 more GPUs to the existing 10 previous ones in the backend. A lot of horsepower to suport Industry-strength challenges, thanks for the sponsorship of re'gion Ile-de-France, ANR, Universite' Paris-Saclay, CNRS, INRIA, and ChaLearn.

- **CODABENCH:** Codabench [147] is a new version of Codalab emphasizing the orgnization of benchmarks, which can be thought of as ever-lasting challenges, de-emphasizing competiton, and favoring the comparison between algorithms. Codabench has also all the capabilities of Codalab and will progressively replace it. When Codabench is fully stable, we will retire Codalab.

  The V1 of Codabench was launched in August 2023. The user base is rapidly growing (over 3000 users, 67 public competitions, and 25000 submissions)

7.3 Open data

- We created a new OCR dataset for historical handwritten Ethiopic script (**HHD-Ethiopic**) [61] , characterized by a unique syllabic writing system, low resource availability, and complex orthographic diacritics. The dataset consists of roughly 80,000 annotated text-line images from 1700 pages of 18 th to 20 th century documents, including a training set with text-line images from the 19 th to 20 th century and two test sets. One is distributed similarly to the training set with nearly 6,000 text-line images, and the other contains only images from the 18 th century manuscripts, with around 16,000 images. The former test set allows us to check baseline performance in the classical IID setting

(Independently and Identically Distributed), while the latter addresses a more realistic setting in which the test set is drawn from a different distribution than the training set (Out-Of-Distribution or OOD). Multiple annotators labeled all text-line images for the HHD-Ethiopic dataset, and an expert supervisor double-checked them. We assessed human-level recognition performance and compared it with state-of-the-art (SOTA) OCR models using the Character Error Rate (CER) and Normalized Edit Distance(NED) metrics. Our results show that the model performed comparably to human-level recognition on the 18 th century test set and outperformed humans on the IID test set. However, the unique challenges posed by the Ethiopic script, such as detecting complex diacritics, still present difficulties for the models. Our baseline evaluation and HHD-Ethiopic dataset will encourage further research on Ethiopic script recognition.
URL: https://github.com/bdu-birhanu/HHD-Ethiopic.
Contact: Birhanu Hailu Belay, `birhanu-hailu.belay@upsaclay.fr`

- Thanks to **OmniPrint**, a synthetic data generator of isolated printed characters (see Section 7.2), we created a dataset made of a wide variety of printed characters from various languages, fonts and styles, with customized distortions. It currently includes 935 fonts from 27 scripts and many types of distortions. It has been used in NeurIPS competitions and various scientific publications, it allows benchmarking algorithms across diverse domains such as meta-learning, few-shot learning, domain adaptation, etc.
  URL: https://github.com/SunHaozhe/OmniPrint
  Contact: `omniprint@chalearn.org`

8 New results

8.1 Frugal Learning

Participants: Guillaume Charpiat, Isabelle Guyon, Alessandro Ferreira Leite, Marc Schoenauer, Michèle Sebag, Sylvain Chevallier, Alice Lacan, Romain Egele, Manon Verbockhaven, François Landes, Maria Sayu Yamamoto, Bruno Aristimunha Pinto, Blaise Hanczar(Univ. Evry).

In Manon Verbockhaven's PhD thesis, we study how to optimally grow a neural network architecture, to increase the performance (in terms of loss) while keeping the network as small as possible (in particular, avoiding redundancy). We showed how to formulate the notion of "expressivity bottleneck" in an easily computable manner, and obtain optimal neuron weights as the result of a small SVD. We showed the approach can scale up, with an experiment using ResNet18 on CIFAR-100. With Barbara Hajdarevic's internship, we also started to extend the addition of neurons to existing layers, to the addition of layers to an existing computation graph.

Within the context of Vincenzo Schimmenti's PhD [59], we demonstrate [79] that the information contained in GPS stations, that monitor the Earth surface deformation, can be leveraged to predict aftershocks to large earthquakes, reaching a balanced accuracy of 70%. This is true both for a very robust model (logistic regression, 2 parameters) and for our proof-of-concept CNN, which we manage to avoid to overfit using a robust ensembling technique, despite the very small number of training samples (48 spatial maps only).

Applications to brain imaging provide an effective testbed for frugal learning, embrassing both the data frugality and computational complexity constraints. The brain-computer interfaces (BCI), using the online decoding of human neural activity to interact with a machine, prove to be a flagship application with the rise of large scale neurotech projects like Elon Musk's Neuralink or Meta's neural interface. The Riemannian BCI, based on machine learning algorithms exploiting the geometry of spatial covariance matrices, yields the state-of-the-art models. This success is tentvatively explained from the ability to endow ML models with the appropriate invariances, make most of the scarse and noisy data at hand, while keeping a computational complexity well within the realm of real time algorithms.

Special effort have been made to find richer representations [49, 91] and to harness the individual variability of human subject with robust classifiers [41]. Indeed, this enriched representation leads to increasing the dimensionality of the problem and calls for methods to select relevant information [87] or to remove unwanted samples [29]. These methods lead to sucessful application in human-machine interfaces [32]. as well as in medical domain [20]. These representations can be generalized with minor modifications to temporal data.

The adaptation of deep models to the multivariate timeseries extracted from brain signal has proven to be more difficult than expected. The data frugality, with the very limited availability of labeled data, is the main obstacle to the large scale deployment of deep models. To overcome this issue, we explore the opportunity to learn multiple views of small datasets [85], and achieve data augmentation of the latent space in the Riemannian framework [86]. While correctly labeled data are very scarse, there are large amounts of recorded brain activity while the subjects are not performing any tasks like resting states or sleep recording. We explore the possibility to train diffusion model to generate relevant brain activities [51].

In the domain of bioinformatics, a main issue is to deal with with sparse data. Data augmentation based on generative models is used to compensate for the data shortage [22]. In this context, the frugality consists of achieving data augmentation in a reduced dimensionality space and thereafter expanding the generated data in the original data space.

8.2 Toward Good AI

Participants: Philippe Caillou, Isabelle Guyon, Alessandro Leite, Michèle Sebag, Sylvain Chevallier, Flora Jay, Burak Yelmen, Cyril Furtlehner, Guillaume Charpiat, Aurelien Decelle, Armand Lacombe, Cyriaque Rousselot, Nicolas Atienza, Romain Egele, Haozhe Sun, Shuyu Dong, Antoine Szatkownik, Olivier Allais (INRAE), Julia Mink (Univ. Bonn), Jean Pierre Nadal (CAMS EHESS), Annick Vignes (CAMS EHESS), David Lopez-Paz (Facebook/Meta).

8.2.1 Causal Learning

Spatio-temporal causal modelling is tackled within the Horapest DataIA project (coll. INRAE; Cyriaque Rousselot's PhD). The goal is to assess the causal effects of the diffusion of pesticides in French residential areas, through exploiting the data from the Health Data Hub together with the newly available dataset reporting the concentrations of diverse molecules in 50 stations on a weekly basis (CNEP), and the overall amount of products bought yearly in every postal code (BNVD). The potential effects that will be investigated concern the children' health in the 2019-2022 period, born between 2013 and 2019. The study will contrast the children resident in places with high or low pesticide average concentration on average, and the children with high or low pesticide concentration *in utero*. The access to the HDH data (IRB demand asked in 2018, accepted in 2019, data expected in 2023) should be delivered in 2024. A major difficulty lies in the hidden confounders, suggesting that leveraging additional data is needed. In parallel, a joint proposal with CHU Toulouse and UAG has been submitted to secure the access to external data with support from medical experts.

A second direction is explored in partnership with Fujitsu (Shuyu Dong's postdoc). The goal is to achieve linear Structural Equation Model (SEM) identification from observational data in the large p small n context. The famed characterization of DAG graphs through the exponential trace of the graph proposed by [149] is of cubic complexity in the number of variables. A low rank decomposition of the inverse covariance matrix combined with an approximation of the gradient has been proposed with a significantly better scalability, at the expense of a moderate loss of accuracy in [114]. The challenge is to distinguish the impact of the statistical and the geometrical tasks involved in SEM identification, respectively aimed to estimate the inverse covariance matrix, and to infer the causal graph from the inverse covariance matrix [67]. Another challenge is related to the scalability w.r.t. the number of variables. Along this line, a promising direction based on a Divide-and-Conquer approach, has been proposed. Formally, sub-problems involving the Markov blanket of each variable are defined and solved (thus with moderate complexity); the reconciliation of these partial solutions is formulated as a constraint satisfaction problem, reaching a very good trade-off between time-complexity and accuracy.

A third direction is considered with Nicolas Atienza (PhD Cifre Thales), co-supervised with Johanne Cohen, LISN. The goal is to extend algorithmic recourse [125] to the identification and correction of inappropriate tuning for a critical system. Preliminary investigations have conducted to determining a sufficient and inexpensive characterization of the system state and a patent has been filed on this characterization. Notably, this direction also led to proposing a new approach for model explanations, based on multi-modal embeddings.

A fourth direction is tackled in Audrey Poinsot's PhD (Cifre Ekimetrics), concerned with counterfactual reasoning in the context of strategic and marketing decisions. Because data in that area does not pertain to Big Data, the PhD currently focuses on Data Augmentation in Causal context [77], as known causality links can be leveraged to ease learning.

A joint work involving Paola Tubaro while she was still in the team was published in 2023. It tries to explain the links between gender and the risk of developing severe COVID-19 (men are more vulnerable than women to severe Covid) through the lens of causality [16].

Causality is also at the core of TAU participation in the INRIA Challenge *OceanIA*, that started in 2021 [100]. The main challenge is related to out-of-distribution learning, motivated by the analysis of the TARA images to identify the ecosystems in the diverse sites of the data collection. The high imbalance of the data among the classes, the prevalence of outliers, suggest that the use of multi-modal embeddings as explored in N. Atienza's PhD, might support the design of relevant metrics in the considered space.

Other motivating applications for causal modeling are described in section 4.1.

In Isabelle Hoxha's PhD, the ability to anticipate choice decisions is investigated within theoretical models [21]. This step towards a better understanding of human decision-making processes is tightly related to establishing causal relationships. On one hand, the applicative goals of this PhD are more oriented towards the explainability of human behavior; on the other hand, it offers some contributions to causal machine learning. An original contribution of the proposed model is to support and explain the emergence of causal representations, based on observations at a single-trial level. This work yields novel principles to learn causal representations from sparse input; it also delivers insights into lifelong learning.

Still in the domain of neurosciences, the development of appropriate representations is hoped to support the extraction of robust causal relationships. The neural avalanche model [13] captures the dynamics of activation sequences observed in the brain; it yields a promising and general feature space to interpret these activations. Formally, the challenge is to exploit such avalanche models to characterize and explain the sequence of temporal activations related with multiple sources, supporting a rich categorisation of neuroscience data.

8.2.2 AutoML

Romain Egele's PhD (coll Argonne National Labs, USA) is actively working on Neural Architecture Search (NAS). He developed a package called DeepHyper, allowing users to conduct NAS with genetic algorithms using TensorFlow or PyTorch, the principal Deep Learning framework [119]. His contributions include applying Recurrent Neural Network Architecture Search for Geophysical Emulation and Scalable Reinforcement-Learning-Based Neural Architecture Search for Cancer Deep Learning Research.

Romain Egele has made advances in 2023 in the design of exploration strategies for optimizing workflows on parallel computing platforms of any scale. These strategies encompass single-objective and multi-objective black-box, gray-box, and transfer-learning optimization techniques. He introduced the design of configurable workflows, encompassing HPC data services, neural architecture and optimization hyperparameters of deep neural networks. He developed novel methods for improved quantification of data and model uncertainty in deep neural networks, building upon the afore-mentioned contributions.

In Asynchronous Decentralized Bayesian Optimization for Large Scale Hyperparameter Optimization [43], the centralized architecture with 1 manager and N workers to a decentralized architecture with N agents sharing information. This makes a new agent suggestion overhead independent of the quantity of available resources and therefore allows for large scale Bayesian optimization. This work has been expanded to optimize multiple objectives in [68] and then applied to large language models [65].

A benchmark of early discarding strategies was conducted to compare state of the art algorithm. It was noticed that a very simple strategy, dubbed 1-Epoch, performed significantly better when "computing duration" is the bottleneck. A method based on Bayesian regression (including both aleatoric and epistemic uncertainties) for learning curve extrapolation was also proposed and dubbed Robust Bayesian Early Rejection (RoBER) [42].

Learning workflows often lack uncertainty estimates which are essential to evaluate the confidence in our predictions. The optimization strategies Romain proposed allow to evaluate a large quantity of "learning workflows"; this makes it possible to checkpoint such workflows once they are trained and leverage such history to create an ensemble of models which provides predictive estimates with lower variance (main prediction and aleatoric uncertainty estimates of P(y|x)) and more informative epistemic uncertainty estimates (model uncertainty) thanks to model diversity. This has been applied to quantifying uncertainty for deep learning based forecasting and flow-reconstruction using neural architecture search ensembles [23].

In 2023 Haozhe Sun defended his PhD on the problem of modularity in Deep Learning [60]. This thesis will contribute to reduce the barrier of entry in using DL models for new applications, a step towards "democratizing AI". The current trend in Artificial Intelligence (AI) is to heavily rely on systems capable of learning from examples, such as Deep learning (DL) models, a modern embodiment of artificial neural networks. While numerous applications have made it to market in recent years (including self-driving cars, automated assistants, booking services, and chatbots, improvements in search engines, recommendations, and advertising, and health-care applications, to name a few) DL models are still notoriously hard to deploy in new applications. In particular, they require massive numbers of training examples, hours of GPU training, and highly qualified engineers to hand-tune their architectures.

In 2023, Haozhe Sun published a survey under revision and recently submitted a paper on novel algorithms for low-cost AI exploiting modularity [46]. He also published a new algorithm dubbed RRR-net, considerably reducing the size of ResNet networks in application to image classification [47].

8.2.3 Privacy in deep learning models that generate DNA

In collaboration with the Institute of Genomics of Tartu, we have been leveraging two types of generative neural networks (Generative Adversarial Networks and Restricted Boltzmann Machines) to learn the high dimensional distributions of real genomic datasets and create artificial genomes [95]. These artificial genomes retain important characteristics of the real genomes (genetic allele frequencies and linkage, hidden population structure, ...) without copying them and have the potential to be valuable assets in future genetic studies by providing anonymous substitutes for private databases (such as the ones hold by companies or public institutes like the Institute of Genomics of Tartu.

The main challenges lie in scaling up to the full genome and in making sure that no personal genetic data is leaked. For this, we developed various deep learning generative architectures, from plain GANs and RBMs, to convolutional GANs, with or without attention [30, 31], applied either in the SNP data space (ie., the space of DNA sequences, removing sites that are constant accross the dataset) or in the Principal Component space [48], which is much smaller when facing datasets with fewer individuals (5 000) than SNP sequence length (60 000). Such a low ratio between number of samples and sample dimension makes the task proned to overfitting and calls for careful check of possible privacy leaks. We studied various privacy scores, including in particular the AATS metric (nearest neighbor adversarial accuracy) proposed by [148], and sorted the different models accordingly.

8.3 Machine Learning with/for Statistical Physics

Participants: Cyril Furtlehner, François Landes, Beatriz Seoane, Guillaume Charpiat, Michele Sebag, Anaclara Alvez-Canepa, Nicolas Béreux, Emmanuel Goutierre, Decelle Aurélien (UCM), Catania Giovanni (UCM), Rahul Chako (external post-doc), Andrea Liu (UPenn), David Reichman (Columbia), Giulio Biroli (ENS), Olivier Dauchot (ESPCI), Johanne Cohen (LISN), Christelle Bruni (IJCLAB), Hayg Guler (IJCLAB).

Generative models constitute an important piece of unsupervised ML techniques, which is under rapid development. In this context, insights from statistical physics are important, especially for energy-based models such as restricted Boltzmann machines. The information content of a trained Restricted Boltzmann Machine (RBM) and its learning dynamics can be precisely analyzed using ensemble averaging techniques [111, 112]. We have also described in great detail the effects of inadequate MCMC sampling on the quality and performance of RBMs [98], and energy-based generative models in general [35]. These two papers came to an unexpected conclusion: the most efficient way to train energy-based models for accurate sampling generation applications is to use non-convergent Markov chains and exploit the out-of-equilibrium memory effects imprinted in the model. This strategy has proven to be extremely effective when generating highly structured data such as DNA/RNA/protein sequences in a very short time [62, 52]. The possibility of rapid generation according to the out-of-equilibrium strategy, comes at the loss of the link between the Boltzmann weight of the model and the statistics of the data set, which limits the use of models trained in this way for interpretability applications. For this reason, another line of research in the group has focused on proposing alternative training strategies that ensure correct equilibrium training. In this sense, important insights can be gained by considering data with low intrinsic dimension where exact solutions of the RBM can be obtained [94], thanks to a convex relaxation. In particular, we have found a 1st order transition mechanism that may plague the learning in a more advanced part of the learning. In [12], we further explore this question and show that sampling the equilibrium distribution with the Markov chain Monte Carlo method can be dramatically accelerated when using biased sampling techniques, in particular the Tethered Monte Carlo (TMC) method. This sampling technique can also be used to improve the calculation of the log-likelihood gradient during training, leading to dramatic improvements when training RBMs with artificial clustered datasets. On real low-dimensional datasets, this new training method fits RBM models with significantly faster relaxation dynamics than those obtained with standard persistent contrastive divergence recipes.

Learning dynamics has also been adressed in a different context of feature learning processes in [17] where closed form expressions are obtained for train and test errors via random matrix theory, yielding a characterization of good alignment between the features and the signal, and the derivation of a set autonomous equations driving the process at large scale. In a different perspective it is shown in [14] how a hierarchical clustering on biological sequences data can be efficiently performed from a sequence of learned RBMs by extracting their TAP fixed points which happen to be hierarchically organized, summarizing the learned free energy landscape. In [66] we illustrate another possible interpretation offered by RBMs by inferring couplings of a generalized p-spin model representing for instance direct coupling interactions in biological sequences. We derive explicit formulas to obtain these coupling from a trained RBM – which beyond paiwise order cannot be otherwise obtained in a tractable way from explicit p-spin models due to combinatorial explosion – and show effective detection of 3-spin interactions on various experiments.

As mentioned earlier, the use of ML to address fundamental physics problems is quickly growing. A place where ML can help address fundamental physics questions is the domain of glasses (how the structure of glasses is related to their dynamics), which is one of the major problems in modern theoretical physics [106] and play a key role in Giorgio Parisi's career (2021 Nobel prize laureate). This year, with controlled numerical experiments, we clarified the important role of Dynamical Facilitation in the melting or equilibration of glasses, discarding first-order transition style analogies [64]. There are various ways in which ML can help address fundamental questions about the physics of glasses, that we participated in reviewing in a Roadmap paper, to be published as a Nature Roadmap [70]. Our angle is to learn the hidden structures (features) that control the flowing or non-flowing state of matter, discriminating liquid from solid states, using rotation-equivariant neural networks.

We prove that rotation-equivariant GNNs outperform other approaches in terms of generalization power, displaying especially good generalization to unseen temperatures [139]. Our approach was benchmarked against other recent works in the roadmap [70], confirming that our approach is extremely promising; we currently are actively exploring this avenue of research.

Another line of research in the group is the application of the physics of spin glasses (originally disordered magnetic alloys) to better understand the role of parameters in ML models by calculating the phase diagrams of the model. Following this approach, we have described the storage capacity of bidirectional associative memories as a function of parameters of the model such as the asymmetry between the layers or the number of patterns considered [11]. Along the same line, we analytically described the phase diagram of the paradigmatic binary perceptron model in the teacher-student scenario and showed that faster training (which also requires significantly less data) can be achieved when multiple systems are coupled together [63], a strategy used in ML to maximize local entropy. In addition, an attempt is being made to better characterize the peculiar and rich out-of-equliibrium phenoma of spin glass systems. In this sense, it is worth mentioning that we were recently able to reproduce and describe (for the first time) the surprising memory and rejuvenation experiments by simulations [9, 75], to show that relaxation in spin glasses is indeed multifractal [10], and to describe the nonlinear responses[26]. A book chapter has also been published on the physics of low dimensional spin glasses [56].

Another line of research is concerned with learning a fast surrogate model for particle accelerator (coll. IJCLab, Univ. PSaclay, and GALAC, LISN; Emmanuel Goutierre's PhD). The challenge is both to model the particle beam (e.g., through aggregate indicators or extending point cloud approaches) and to model the loss (as the simulator involves the propagation of the beam through a sequence of chambers with different physics mechanisms [44, 88].

8.4 Machine Learning for Numerical Simulations

8.4.1 ML and Reduced Order Models for Dynamical Systems

Participants: Michele Alessandro Bucci, Marc Schoenauer, Emmanuel Menier, Thibault Monsel, Mouadh Yagoubi (IRT-SystemX), Lionel Mathelin (DATAFLOT team, LISN), Petros Koumoutsakos (Harvard SEAS).

Numerical simulations of fluid dynamics in industrial applications require the spatial discretization of complex 3D geometries with consequent demanding computational operations for the PDE integration. The computational cost is mitigated by the formulation of Reduced Order Models (ROMs) aiming at describing the flow dynamics in a low dimensional feature space. The Galerkin projection of the driving equations onto a meaningful orthonormal basis speeds up the numerical simulations but introduces numerical errors linked to the underrepresentation of dissipative mechanisms.

Deep Neural Networks can be trained to compensate missing information in the projection basis. By exploiting the projection operation, the ROM correction consists in a forcing term in the reduced dynamical system which has to: i) recover the information living in the subspace orthonormal to the projection one; ii) ensure that its dynamics is dissipative. A constrained optimization is then employed to minimize the ROM errors but also to ensure the reconstruction and the dissipative nature of the forcing [24, 53], improving the prediction while preserving the guarantees of the ROM. The approach was extended on Michelin use case of rubber calendering process [135].

During his PhD thesis, Emmanuel Menier spent 3 months in Spring 2023 in the group of Prof. Petros Koumoutsakos, at SEAS - Harvard, John A. Paulson School of Engineering and Applied Sciences. It was a perfect match between his previous work and the group's expertise in high dimensional dynamical complex system (e.g., CFD), and resulted in the Interpretable Learning Effective Dynamics (iLED) framework, a novel framework based on nonlinear dimension reduction thanks to deep neural networks, that offers comparable accuracy to state-of-the-art recurrent neural network-based approaches while providing the added benefit of interpretability [89, 71] (submitted). The basic idea of iLED is grounded on the Mori-Zwanzig formalism, an approach that has been later generalized to other dynamical systems [54], hybridizing with previous work of the team on Neural Delay ODEs [72].

Thibault Monsel's PhD has been indeed focusing on the learning of dynamical sytems involving delays, i.e. Delayed Differential Equations (DDE). While Neural ODE was a conceptual breakthrough, it cannot learn partially-observable dynamical systems, which lead to Augmented Neural ODE, where supplementary state variables help keeping a memory of the past, which can be sufficient to infer the missing information. Inspired by the Mori-Zwanzig formalism and Takens theorem, we develop another way to extract this information, using delays, i.e. using past observable states of the system [54]. These delays may depend on the current state. Thibault contributed to implement efficiently Delayed Differential Equations in deep learning frameworks (Ajax, PyTorch) and showed the advantage of DDEs over (Augmented) Neural DDEs and recurrent networks (LSTM) under certain circonstances [72]. Furthermore, constant delays can now be learned during training.

8.4.2 Graph Neural Networks for Numerical Simulations

Participants: Guillaume Charpiat, Michele Alessandro Bucci (Safran Tech), Marc Schoenauer, Wenzhuo Liu, Matthieu Nastorg, Emmanuel Menier, Mouadh Yagoubi (IRT-SystemX), Lionel Mathelin (DATAFLOT team, LISN), Thilbault Faney (IFPEN), Jean-Marc Gratien (IFPEN).

The use of GNNs to approximate the numerical solutions of PDEs on unstructured meshes, rather than using CNNs on grid meshes was systematicall studied in Wenzhuo Liu's PhD: After porting ideas from multi-grid approaches to Finite Elements, and comparing the CNN and GNN approchaes [131], she tackled the poor Out-of-Distribution generalization issue using Meta-Learning [130], improving the OoD learning on CFD simulations of air flow around an airfoil by considering each airfoil shape as a separate task. The last part of her PhD, defended in March 2023 [57], applies Transfer Learning to decrease the amount of data needed to learn accurate simulation on fine meshes using numerous costless simulations on coarse meshes. Note that in parallel, GNNs had been tested and adopted in close collaborators of the DATAFLOT team at LISN [78], resulting in the hiring of Alessandro Bucci thanks to a bilateral contrat with IFPEN.

During the 2.5 years that he spent at TAU, Alessandro worked on several use cases of IFPEN, with the goal of accelerating some softwares that IFPEN uses daily. This IFPEN/TAU collaboration led to a successful DataIA proposal, ML4CFD, that funds Matthieu Nastorg's PhD (to be defended in March 2024). Matthieu's work aims to significantly accelerate [136] the numerical resolution of the Poisson equation (ubiquitous in CFD, e.g., to compute the pressure in Navier Stokes simulations), based on B. Donon's Deep Statistical Solvers (DSS) [118]. A significant improvement was brought by replacing the arbitrary number of iterations of the message passing mechanism by the solution of a fixed point equation [73]. The scale-up issue has been addressed in the final part of Matthieu's PhD by considering the DSS approach as a preconditionner for a domain decomposition method [74].

8.4.3 Advances in sparse recovery for inverse problem and application in M/EEG

Participants: Matthieu Kowalski, Pierre Barbault, Benoit Malézieux, Charles Soussen (L2S), Thomas Moreau (Inria Mind), Diego Delle Donne (Essec), Leo Liberti (LiX), Clément Bonnet (Univ. Bretagne Sud) and his supervisors.

In[40], we present an innovative method for brain-age prediction employing a novel Sliced-Wasserstein distance metric, enhancing the analysis of covariance matrices that epitomize EEG and MEG signal distributions. This method excels by integrating the strong theoretical underpinnings of Riemannian geometry with the computational efficiency of kernel methods, leading to significant advancements in domain adaptation for Brain Computer Interface (BCI) applications. The proposed distance showcases effective surrogate qualities to the Wasserstein distance and outperforms contemporary algorithms in both computational efficiency and prediction accuracy.

In [37] (see also [36]), by employing a Bernoulli-Gaussian prior and an innovative Expectation-Maximization framework to optimize the so-called Marginal MAP estimation of the support instead of the usual joint-MAP of both the support and the activation value, we have significantly improved the identification of the support of sparse signals. This new methodology demonstrates superior performance over traditional L_0 norm-based techniques in terms of error rates and signal-to-noise ratio. Furthermore, in [15] we provide groundbreaking perspective on the L_2+L_0 sparse approximation problems, introducing a comprehensive, big-M independent integer linear programming formulation. This development effectively circumvents the shortcomings of existing methods, ensuring the accurate recovery of global minimizers without requiring predefined bounds.

8.4.4 Population genetics: demography inference by inverting simulators

Participants: Guillaume Charpiat, Flora Jay, Aurélien Decelle, Cyril Furtlehner, Jérémy Guez, Arnaud Quelin, Jazeps Medina Tretmanis, Burak Yelmen.

Collaboration: Bioinfo Team (LISN), Estonian Biocentre (Institute of Genomics, Tartu, Estonia), UNAM (Mexico), U Brown (USA), U Cornell (USA), TIMC-IMAG (Grenoble), MNHN (Paris), Pasteur Institute (Paris), ETH Zurich.

Thanks to the constant improvement of DNA sequencing technology, large quantities of genetic data should greatly enhance our knowledge about evolution and in particular the past history of a population. This history can be reconstructed over the past thousands of years, by inference from present-day individuals: by comparing their DNA, identifying shared genetic mutations or motifs, their frequency, and their correlations at different genomic scales. Still, the best way to extract information from large genomic data remains an open problem; currently, it mostly relies on drastic dimensionality reduction, considering a few well-studied population genetics features.

For the past decades, simulation-based likelihood-free inference methods have enabled researchers to address numerous population genetics problems. As the richness and amount of simulated and real genetic data keep increasing, the field has a strong opportunity to tackle tasks that current methods hardly solve. However, high data dimensionality forces most methods to summarize large genomic datasets into a relatively small number of handcrafted features (summary statistics). With Theophile Sanchez' PhD [142], we had proposed an alternative to summary statistics, based on the automatic extraction of relevant information using deep learning techniques. This lead to the release of DNADNA, a flexible open-source python-based software for deep learning inference in population genetics[1] (cf section 7). It is task-agnostic and aims at facilitating the development, reproducibility, dissemination, and reusability of neural networks designed for genetic polymorphism data. dnadna defines multiple user-friendly workflows[27].

We continued that line of research, based on training from simulations, and extended it to ancient DNA simulations [80], and to the inference of other traits from DNA statistics, such as the cultural transmission of reproductive success [90, 19]. We also theoricized genetic phenomena obervable through statistics [18, 25]. Finally, we applied this framework to real data such as pre-Hispanic Central Mexico [28] and the MUC19 gene in Denisovans, Neanderthals and Modern Humans [82].

We also worked on deep generative models of DNA sequences (cf section 8.2.3).

8.5 Energy Management

Participants: Isabelle Guyon, Alessandro Leite, Marc Schoenauer, Eva Boguslawski, Benjamin Donnot (RTE), Matthieu Dussartre (RTE).

Our collaboration with RTE has a long history, starting with Benjamin Donnot's (2016-2019) [115] and Balthazar Donon's [117] CIFRE PhDs, and is centered on the maintainance of the national French Power Grid. Eva Boguslawski's CIFRE PhD, co-supervised by Alessandro Leite and Marc Schoenauer, started in Sept. 2022. It addresses the control of the grid through decentralized decision process using multi-agent Reinforcement Learning, in the line of the LR2PN challenge that Eva contributed to organize during her Master internship [143]. During the first year of her PhD she presented her first results in two Summer Schools, one on Reinforcement Learning [84], and on on Energy Management [83].

We also have a long history of collaboration with the SME *Artelys*, the most recent project being the ADEME project NEXT, regarding the optimization of regional grid extensions, that was completed only this year, after a 2-years extension due to Covid-19. However, beyond Victor Berger's PhD in 2021 [101] and Herilalaina Rakotoarion's PhD in 2022 [140], we didn't publish any new result in 2023.

[1] https://gitlab.com/mlgenetics/dnadna

8.6 Computational Social Sciences

Participants: Philippe Caillou, Michèle Sebag, Cyriaque Rousselot, Guillaume Bied, Armand Lacombe, Soal Nathan, Hande Gozukan.

Computational Social Sciences (CSS) is making significant progress in the study of social and economic phenomena thank to the combination of social science theories and new insight from data science. While the simultaneous advent of massive data and unprecedented computational power has opened exciting new avenues, it has also raised new questions and challenges.

Several studies are being conducted in TAU, about labor (labor markets), about health (in relation with pollution issues), about diffusion of information around Cartolabe, a platform for scientific information system and visual querying.

8.6.1 Labor Studies

Participants: Philippe Caillou, Michèle Sebag, Guillaume Bied, Armand Lacombe, Solal Nathan, Assia Wirth, Hande Gozukan, Jean-Pierre Nadal (EHESS), Bruno Crépon (ENSAE).

A first area of activity of TAU in Computational Social Sciences is the study of labor and chiefly the recommendation of job offers/job seeker profiles.

Job markets The DataIA project Vadore [34] (partners ENSAE and Pôle Emploi/France Travail) benefits from the sustained cooperation and from the wealth of data gathered by France Travail. The data management is regulated along a 3-partite convention (GENES-ENSAE, Univ Paris-Saclay, Pôle Emploi). Extensive efforts have been required to achieve the data pipelines required to enable learning recommendation models and exploiting them in a confidentiality preserving way (G. Bied's PhD). In 2022, panels and beta-test campaigns have been conducted to assess the suitability of the recommendations. A second beta-test campaign, involving 200,000 job seekers in the region Rhone-Alpes scale has taken place in 2023.

The learned models [39] have been successfully assessed along several criteria. A first criterion regards the performance vs state of the art algorithms and the current Pole Emploi recommender system. A second criterion regards the time-to-recommendation; circa 0.02 seconds are currently required to deliver a bunch of recommended job offers to each job seeker.

A third criterion regards the fairness of the recommendation model [33, 38]. A comprehensive study examining gender-related gap in several utilities (wages, types of contract, distance-to-job) has been conducted, comparing the gaps observed in actual hirings, in applications, and in recommendations. Interestingly, the gap in recommendations closely mimics that in actual hirings and in applications (if any, the recommendation algorithm tends to decrease the gap). Algorithmic fairness in domains as sensitive as employment is under scrutiny of French and European regulations. The difficulty is to decouple the biases observed in applications (thus reflecting job seekers' preferences, that should be respected) from those due to recruiters (that should not be perpetuated in the learned models).

A last criterion concerns the congestion of the job market (share of job offers paid attention to by job seekers). Recommender systems tend to increase the congestion due to the so-called popularity bias. Early attempts to prevent the congestion have been investigated in [50], using optimal transport.

Both fairness and congestion issues are at the core of S. Nathan's PhD (coll. Univ. Ghent, Belgium). A first research direction is concerned with integrating the congestion in the recommendation loss; this requires a global view of the market dynamics, and the difficulty is to design a loss term that is both computationally affordable and differentiable. A second research direction along this line is to integrate an estimation of job offer popularity within the recommendation system, enabling job seekers to anticipate and react to, competition. Interestingly, such compound architecture (integrating the job offer popularity estimate and its effects on the decision of applying) could enable to model competition-avoidance strategies, in particular in relation with gender effects.

A key difficulty for research on ML-based job recommendation is the lack of open and representative datasets, owing to the very sensitive nature of the data and the protection of vulnerable persons. We have co-organized a workshop (Feb. 2023) gathering researchers and industrials on this topic, in collaboration with Actiris and VDAB (public employment services in Belgium), to identify how this lack of open datasets, hindering the benchmarking of existing systems, can be addressed.

8.6.2 Health and practices

Participants: Philippe Caillou, Michèle Sebag, Armand Lacombe, Cyriaque Rousselot, Olivier Allais (INRA), Julia Mink (Univ. Bonn, DE).

Continuing our former partnership with INRAE (in the context of the *Initiative de Recherche Stratégique* Nutriperso; [99]), we proposed the HORAPEST DATAIA project to uncover the potential causal relationships between pesticide dissemination and children's health (Cyriaque Rousselot's PhD). The demand of access has been approved by the CNIL and the Health Data Hub; the data are expected in Feb. 2024, and contacts have been taken with the CHU Toulouse for cooperation on complementary data.

8.6.3 Scientific Information System and Visual Querying

Participants: Philippe Caillou, Michèle Sebag, Anne-Catherine Letournel, Hande Gozukan, Jean-Daniel Fekete (AVIZ, Inria Saclay).

A third area of activity concerns the 2D visualisation and querying of a corpus of documents. Its initial motivation was related to scientific organisms, institutes or Universities, using their scientific production (set of articles, authors, title, abstract) as corpus. The Cartolabe project (see also Section 7) started as an Inria ADT (coll. TAO and AVIZ, 2015-2017). It received a grant from CNRS (coll. TAU, AVIZ and HCC-LRI, 2018-2019).

The originality of the approach is to rely on the content of the documents (as opposed to, e.g. the graph of co-authoring and citations). This specificity allowed to extend Cartolabe to various corpora, such as Wikipedia, Bibliotheque Nationale de France, or the Software Heritage. Cartolabe was also applied in 2019 to the *Grand Debat* dataset: to support the interactive exploration of the 3 million propositions; and to check the consistency of the official results of the *Grand Debat* with the data. Cartolabe has also been applied in 2020 to the COVID-19 kaggle publication dataset (Cartolabe-COVID project) to explore these publications.

Among its intended functionalities are: the visual assessment of a domain and its structuration (who is expert in a scientific domain, how related are the domains); the coverage of an institute expertise relatively to the general expertise; the evolution of domains along time (identification of rising topics). A round of interviews with beta-user scientists has been performed in 2019-2020. Cartolabe usage raises questions at the crossroad of human-centered computing, data visualization and machine learning: i) how to deal with stressed items (the 2D projection of the item similarities poorly reflects their similarities in the high dimensional document space; ii) how to customize the similarity and exploit the users' feedback about relevant neighborhoods. A statement of the current state of the project was published in 2021 [93].

8.7 Organization of Challenges

Participants: Isabelle Guyon, Marc Schoenauer , Anne-Catherine Letournel, Sébastien Tréguer, Adrien Pavao, Eva Boguslawski, Haozhe Sun, Romain Egele, Lisheng Sun, Mouadh Yagoubi (IRT SystemX), Antoine Marot (RTE), Benjamin Donnot (RTE), Sergio Escalera (U. Barelona).

The TAU group uses challenges (scientific competitions) as a means of stimulating research in machine learning and engage a diverse community of engineers, researchers, and students to learn and contribute advancing the state-of-the-art. The TAU group is community lead of the open-source **Codalab** platform (see Section 7), hosted by Université Paris-Saclay. The project had grown since 2019 and includes now an engineer dedicated full time to administering the platform and developing challenges (Adrien Pavao), financed in 2021 by a 500k€ project with the Région Ile-de-France. This project also received the support of the Chaire Nationale d'Intelligence Artificielle of Isabelle Guyon (2020-2024), Lawrence Berkeley Labs (2022-2025, Fair Universe project), and TAILOR ICT48 Network of Excellence.

Adrien Pavao has defended his thesis [58] on the theoretical rationalization of judging competitions, building on the theory of social choice. This is applicable, in particular to judging multi-task or multi-objective challenges: each task or objective can be thought of as a "judge" voting towards determining a winner. He devised novel empirical criteria to assess the quality of ranking functions, including the generalization to new tasks and the stability under judge or candidate perturbation and conducted empirical comparisons on 5 competitions and benchmarks. While prior theoretical analyses indicate that no single ranking function satisfies all desired theoretical properties, our empirical study reveals that the classical "average rank" method (often used in practice to judge competitions) fares well. However, some pairwise comparison methods can get better empirical results. Beyond his PhD, Adrien also wrote a survey of the different types of competitions in Machine Learning, detailing the recommended protocol for each of them [81]. He also published a genera tutorial to guide future challenge organisers [76].

The challenge series in Reinforcement Learning for Power Grid control, started in 2021 with the company RTE France on the theme "Learning to run a power network" [134] (**L2RPN**, http://l2rpn.chalearn.org), has led to a highly impactful challenge "AI for Industry", for startuo companies. The challenge was sponsored by Paris Région Ile-de-Frances wth a prze of 500 KEuros. The goal is to test the potential of Reinforcement Learning to solve a real world problem of great practical importance: controlling electricity transportation in smart grids while keeping people and equipment safe. The 2023 edition had objective to devise control policies for the French electricity grid under scenarios of energies of the future, towards attaining carbon neutrality. The participants will be tackling prospective productions and consumption scenarios of the future, emphasing renewable energies. This poses particular difficulties because of solal and wind energies have irregular productions [143].

We launched in 2023 a new challenge series called AutoSurvey [45], sponsored by Google and ChaLearn https://auto-survey.chalearn.org/.

The goal of these challenges is to advance the generation of systematic review reports, overview papers, white papers, and essays that synthesize on-line information. The coverage spans multiple domains including Literary or philosophical essays (LETTERS), Scientific literature (SCIENCES), and topics surrounding the United Nations Sustainable Development Goals (SOCIAL SCIENCES). The participants submit code (AI-agents) capable of composing survey papers, using internet resources. Such AI-agents will thus operate as AI-Authors.

A second track of the competition aims are advancing the automated reviewing of papers. The partcipants must submit AI-reviewers, capable of revewing papers such as those written by AI-authors.

We continue using challenges in teaching. The masters students of the AI master designed several small challenges, which are then given to other students in labs, and both types of students seem to love it. In 2023, they organized biomedicine, particle physics and computer vision challenges, on the theme of bias in data and fairness.

In terms of dissemination, a collaborative book "AI competitions and benchmarks: The science behind the contests " written by expert challenge organizers is under way and will appear in the Journal of Data-centric Machine Learning Research (DMLR), see http://www.chalearn.org/books.html. Challenge organizaton is now better grounded in theory, with such effort. The thesis of Adrien Pavao includes several advances in devising sound challenge protocols, including two-stage challenges, as described in his recent paper "Filtering participants improves generalization in competitions and benchmarks" [138]. Chapter 2 [55] was presented as a tutorial at NeurIPS 2023 (Data-Centric AI for reliable and responsible AI: from theory to practice).

Last but not least, our long-lasting collaboration with the particle-physics community (in particular with David Rousseau from LAL in Orsay), that started with the world-wide successful HiggsML challenge [104, 103], temporarily ended with the TrackML challenge [110], whose second "Throughput" phase was detailed in [7].

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

TAU continues its policy about technology transfer, accepting any informal meeting following industrial requests for discussion (and we are happy to be often solicited), and deciding about the follow-up based upon the originality, feasibility and possible impacts of the foreseen research directions, provided they fit our general canvas. This lead to the following 4 on-going CIFRE PhDs, with the corresponding side-contracts with the industrial supervisor, and the continuation until September 2023 of the bilateral contract with Fujitsu (within the national "accord-cadre" Inria/Fujitsu).

- **Fujitsu**, 2021-2022 renewed 2022-2023 (200k€ per year), *Causal discovery in high dimensions*. Project ended September 2023.
  Coordinator: Marc Schoenauer
  Participants: Shuyu Dong and Michèle Sebag
- **CIFRE RTE** 2021-2024 (72 kEuros), with RTE, related to Eva Boguslawski's CIFRE PhD *Decentralized Partially Observable Markov Decision Process for Power Grid Management*
  Coordinator: Marc Schoenauer and Matthieu Dussartre (RTE)
  Participants: Eva Boguslawski, Alessandro Leite
- **CIFRE Ekimetrics** 2022-2025 (45 kEuros), with Ekimetrics, related to Audrey Poinsot's CIFRE PhD *Causal incertainty quantification under partial knowledge and low data regimes*
  Coordinator: Marc Schoenauer and Nicolas Chesneau (Ekimetrics)
  Participants: Guillaume Charpiat, Alessandro Leite, Audrey Poinsot and Michèle Sebag
- **CIFRE MAIR** 2022-2025 (75 kEuros), with Meta (Facebook) AI Research, related to Mathurin Videau's CIFRE PhD *Reinforcement Learning: Sparse Noisy Reward*
  Coordinator: Marc Schoenauer and Olivier Teytaud (Meta)
  Participants: Alessandro Leite and Mathurin Videau
- **CIFRE MAIR** 2022-2025 (75 kEuros), with Meta (Facebook) AI Research, related to Badr Youbi's CIFRE PhD *Learning invariant representations from temporal data*
  Coordinator: Michèle Sebag and David Lopez-Paz (Meta)
  Participants: Badr Youbi
- **Google initiater grant** (20 kEuros): AutoSurvey
  Coordinator: Marc Schoenauer and Isabelle Guyon
  Participants: Benedictus Kent Rachmat, Khuong Thanh Gia Hieu, Ihsan Ullah, and Lisheng Sun
- **AI Verse**, related to Abir Affane's post-doc
  Coordinator: Pierer Alliez (INRIA Titane)
  Participant: Guillaume Charpiat

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program

- **MDG-TAO** Associate Team involved in the International Lab IIL CWI-Inria
  Title: Data-driven simulations for Space Weather predictions
  International Partner : CWI (Holland) - Multiscale Dynamics Group - Enrico Camporeale
  Start year: 2017 – See also: web site
  Participants: Cyril Furtlehner, Michèle Sebag
  Abstract: We propose an innovative approach to Space Weather modeling: the synergetic use of SoA simulations with Machine Learning and Data Assimilation techniques, in order to adjust for errors due to non-modeled physical processes, and parameter uncertainties.

2019-2024

10.1.2 Participation in other International Programs

HFSP RGY0075/2019

Title: Evolutionary changes in human hosts and their pathogens during first contact in the New World

Coordinator: Emilia Huerta-Sanchez (U Brown, USA)

Date/Duration: 2019-2024

Participants: Flora Jay

Collaboration: M Avila-Arcos (UNAM, Mexico)

10.2 European initiatives

10.2.1 Horizon Europe

Adra-e Adra-e project on cordis.europa.eu

Title: AI, Data and Robotics ecosystem

Duration: From July 1, 2022 to June 30, 2025

Partners:

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- LINKOPINGS UNIVERSITET (LIU), Sweden
- NATIONAL UNIVERSITY OF IRELAND GALWAY (NUI GALWAY), Ireland
- DUBLIN CITY UNIVERSITY (DCU), Ireland
- AI DATA AND ROBOTICS ASSOCIATION, Belgium
- TRUST-IT SERVICES SRL, Italy
- COMMISSARIAT A L ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES (CEA), France
- UNIVERSITEIT TWENTE (UNIVERSITEIT TWENTE), Netherlands
- DEUTSCHES FORSCHUNGSZENTRUM FUR KUNSTLICHE INTELLIGENZ GMBH (DFKI), Germany
- ATOS SPAIN SA, Spain

- HRVATSKA UDRUGA ZA UMJETNU INTELIGENCIJU (CROATIAN ARTIFICIAL INTELLIGENCE ASSOCIATION), Croatia
- COMMPLA SRL (Commpla Srl), Italy
- ATOS IT SOLUTIONS AND SERVICES IBERIA SL (ATOS IT), Spain
- SIEMENS AKTIENGESELLSCHAFT, Germany
- UNIVERSITEIT VAN AMSTERDAM (UvA), Netherlands

Inria contact: Marc Schoenauer

Coordinator: Marc Schoenauer

Summary: AI, Data and Robotics (ADR) is omnipresent in our daily lives and key to addressing some of the most pressing challenges facing our society. Europe has excellent research centres, innovative start-ups, a world-leading position in robotics and competitive manufacturing and services sectors, from automotive to healthcare, energy, financial services and agriculture. While the essentials are present, European ADR is waiting for exploitation to achieve its full potential. The ADR ecosystem is inherently complex because many stakeholders at many different levels require a holistic strategy towards collaboration to be effective and efficient. The Adra Association, representing the private side of the ADR Partnership, leverages this diversity through its founding organisations (BDVA, euRobotics, CLAIRE, ELLIS, EurAI) and channels it to the benefit of the European ecosystem. The Adra-e CSA proposal is set up in close liaison with Adra Association and includes it as a partner, committed to sustain its outcomes. Adra-e should be seen as the operational arm of the partnership to foster collaboration, convergence and interoperability between communities and disciplines to advance European ADR while safeguarding the interest of European citizens. This is achieved by supporting the ADR Partnership in the update and implementation of the SRIDA, creating the conditions for an inclusive, sustainable, effective, multi-layered, and coherent European ADR ecosystem, leading to increased trust and adoption of ADR, a more competitive supply and demand sides in the EU and raising private investments at the same time.The consortium is composed of leading industry and research organisations with significant expertise in all three disciplines. All are involved in Adra and the associations and partnerships shaping European research. Many of them are supporting the Digitising European Industry initiative from the EC participating in the constitution of Digital Innovation Hubs Network and Digital platforms.

10.2.2 H2020 projects

TAILOR

Participants: Marc Schoenauer, Sébastien Treguer.

TAILOR project on cordis.europa.eu

Title: Foundations of Trustworthy AI - Integrating Learning, Optimization, and Reasoning.

Duration: From September 1, 2020 to August 31, 2024

Partners:

- LINKOPINGS UNIVERSITET, Sweden (coordinator)
- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- and 56 other partners

Inria contact: Marc Schoenauer

Coordinator: Fredrik Heinz

Summary: Maximising opportunities and minimising risks associated with artificial intelligence (AI) requires a focus on human-centred trustworthy AI. This can be achieved by collaborations between research excellence centres with a technical focus on combining expertise in theareas of learning, optimisation and reasoning. Currently, this work is carried out by an isolated scientific community where research groups are working individually or in smaller networks. The EU-funded TAILOR project aims to bring these groups together in a single scientific network on the Foundations of Trustworthy AI, thereby reducing the fragmentation and increasing the joint AI research capacity of Europe, helping it to take the lead and advance the state-of-the-art in trustworthy AI. The four main instruments are a strategic roadmap, a basic research programme to address grand challenges, a connectivity fund for active dissemination, and network collaboration activities.

VISION VISION project on cordis.europa.eu

Title: Value and Impact through Synergy, Interaction and coOperation of Networks of AI Excellence Centres

Duration: From September 1, 2020 to August 31, 2024

Partners:

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- UNIVERSITEIT LEIDEN (ULEI), Netherlands
- NEDERLANDSE ORGANISATIE VOOR TOEGEPAST NATUURWETENSCHAPPELIJK ONDERZOEK TNO (NETHERLANDS ORGANISATION FORAPPLIED SCIENTIFIC RESEARCH), Netherlands
- THALES SIX GTS FRANCE SAS (THALES SIX GTS France), France
- DEUTSCHES FORSCHUNGSZENTRUM FUR KUNSTLICHE INTELLIGENZ GMBH (DFKI), Germany
- CESKE VYSOKE UCENI TECHNICKE V PRAZE (CVUT), Czechia
- FONDAZIONE BRUNO KESSLER (FBK), Italy
- INTELLERA CONSULTING SPA (INTELLERA CONSULTING), Italy
- UNIVERSITY COLLEGE CORK - NATIONAL UNIVERSITY OF IRELAND, CORK (UCC), Ireland

Inria contact: Jozef Geurts

Coordinator: Holger Hoos

Summary: A broad and ambitious vision is needed for artificial intelligence (AI) research and innovation in Europe to thrive and remain internationally competitive. Building upon its strength across allareas of AI and commitment to its core values, Europe is excellently positioned to take a human-centred, ethical and trustworthy approach to AI. However, to establish itself as a powerhouse in AI, Europe needs to overcome the present fragmentation of the AI community, which creates inefficiencies and limits progress on next-generation, trustworthy AI tools and systems that draw from methods across a broader spectrum of AI techniques. Europe also needs to improve interaction between AI researchers, innovators and users. Following on from the European Commission's Communication on AI for Europe and the Coordinated Action Plan between the European Commission and the Member States, European efforts in AI should be strongly coordinated to be internationally competitive. Europe must scale up existing research capacities and reach critical mass through tighter networks of European AI excellence centres. Towards this end, effective coordination between the four networks of AI excellence centres to be established under ICT-48-2020 (Research and Innovation Actions) is of crucial importance. VISION will reinforce and build on Europe's assets in AI, including its world-class community of researchers, and thus enable Europe to stay at the forefront of AI developments, which is widely recognised as critical in maintainingEurope's strategic autonomy in AI. VISION will achieve this in the most efficient and effective manner possible, by strongly building on the success and organisation of CLAIRE (the Confederation of Laboratories for AI Research inEurope, claire-ai.org) as well as on AI4EU, and by leveraging the expertise and connections of several of Europe's leading institutions in AI research and innovation.

10.2.3 Other european programs/initiatives

DeMythif.AI

Participants: Sylvain Chevallier, David Rousseau (LAL).

DeMythif.AI project on cordis.europa.eu

Title: PhD excellence program in Paris-Saclay to unravel AI uncertainty

Duration: From September 2023 date 31 December 2028

Partners:

- Universite Paris-Saclay (coordinator)
- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- and 10 other partners

Inria contact: Sylvain Chevallier

Coordinator: Sylvain Chevallier

Summary:

10.3 National initiatives

10.3.1 ANR

- Chaire IA **HUMANIA** 2020-2024 (600kEuros), *Democratizing Artificial Intelligence.*
  Coordinator: Isabelle Guyon (TAU)
  Participants: Marc Schoenauer, Michèle Sebag, Anne-Catherine Letournel, François Landes.
- PEPR IA **SAIF** (400k€) *Safe AI through Formal methods*
  Coordinator: Caterina Urban (INRIA Antique)
  Participant: Guillaume Charpiat.
- PEPR IA **CAUSALI-T-AI** (400k€) *CAUSALIty Teams up with Artificial Intelligence*
  Coordinator: Marianne Clausel (Université de Loraine)
  Participant: Michèle Sebag, Alessandro Leite.
- **RoDAPoG** 2021-2025 (302k€) *Robust Deep learning for Artificial genomics and Population Genetics*
  Coordinator: Flora Jay,
  Participants: Cyril Furtlehner, Guillaume Charpiat.
- **SPEED** 2021-2024 (49k€) *Simulating Physical PDEs Efficiently with Deep Learning*
  Coordinator: Lionel Mathelin (LISN (ex-LIMSI))
  Participants: Michele Alessandro Bucci, Guillaume Charpiat, Marc Schoenauer.

10.3.2 Others

- **ADEME NEXT** 2017-2021, extended Sept. 2023 due to Covid-19 (675 kEuros). Simulation, calibration, and optimization of regional or urban power grids (Section 4.2).
  ADEME (Agence de l'Environnement et de la Maîtrise de l'Energie)
  Coordinator: SME ARTELYS
  Participants Isabelle Guyon, Marc Schoenauer, Michèle Sebag, Victor Berger (PhD), Herilalaina Rakotoarison (PhD), Berna Bakir Batu (Post-doc)

- **Inria Challenge** (formerly IPL) **HYAIAI**, 2019-2023, *HYbrid Approaches for Interpretable Artificial Intelligence*
Coordinator: Elisa Fromont (Lacodam, Inria Rennes)
Participants: Marc Schoenauer and Michèle Sebag

- **CODALAB** 2019-2023 *soutien à la plate-forme de challenge Codalab*, Financement re'gion Île de France.
Coordinator: Isabelle Guyon
Participants: Adrien Pavao, Marc Schoenauer, Michèle Sebag

- **Inria Challenge OceanAI** 2021-2024, *AI, Data, Models for a Blue Economy*
Coordinator: Nayat Sanchez Pi (Inria Chile)
Participants: Marc Schoenauer, Michèle Sebag and Shiyang Yan

- **DATAIA YARN** 2022-2025, *Automatic Processing of Messy Brain Data with Robust Methods and Transfer Learning*
Coordinator: Sylvain Chevallier
Participants: Florent Bouchard (L2S), Fredéric Pascal (L2S), Alexandre Gramfort (Meta), Sara Sedlar

- **Fair Universe** 2022-2025, We received with Lawrence Berkeley Labs a grant of 6.4 million USD to develop benchmarks in High Energy Physics and implement them on Codabench. Colaboration with David Rousseau of IJCLAB.
Coordinator: Isabelle Guyon
Participants: David Rousseau, Ragansu Chakkappai, Ihsan Ullah

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- Flora Jay - Co-head of LEGO working group of GDR BiM – "Machine learning for genomics" (23-now). Workshops: Paris May 23, Lille Nov 23. https://gt-lego.cnrs.fr/

- Marc Schoenauer, General Chair of the First AI, Data, Robotics Forum, Versaille, Nov. 8-9, 2023.

- Marc Schoenauer, General Chair of the IRT SystemX 3rd International Workshop on Artificial Intelligence and Augmented Engineering, Saclay, Dec. 7, 2023.

Member of organizing committees

- Guillaume Charpiat - co-organizer of the "Geometric Green Learning" session of the Geometric Science of Information conference, 2023; program committee of "Journées de Recherche en Apprentissage Frugal" JRAF 2023

- Flora Jay - Scientific committee of ETEE conference "Empirism & Theory in Ecology and Evolution", Saclay, Nov22

- Marc Schoenauer - Steering Committee, Parallel Problem Solving from Nature (PPSN); Steering Committee, Learning and Intelligent OptimizatioN (LION).

- Michele Sebag - co-Organizer of Workshop: AI4HR, AI for Human Resources, coll. European Conferenc on Machine Learning. 2023.

11.1.2 Scientific events: selection

Reviewer All TAU members are reviewers of the main conferences in their respective fields of expertise.

11.1.3 Journal

Member of the editorial boards

- Isabelle Guyon - Action editor, Journal of Machine Learning Research (JMLR)
- Isabelle Guyon - Co-editor, Data-centric Machine Learning Researchschimmenti:tel-04344778(DMLR)
- Marc Schoenauer - Action editor, Journal of Machine Learning Research (JMLR); Advisory Board, Evolutionary Computation Journal, MIT Press, and Genetic Programming and Evolutionary Machines, Springer Verlag.
- Michèle Sebag - Editorial Board, ACM Transactions on Evolutionary Learning and Optimization.

Reviewer - reviewing activities All members of the team reviewed numerous articles for the most prestigious journals in their respective fields of expertise.

11.1.4 Invited talks

- Guillaume Charpiat, invited talks at *Learning on Graphs and Geometry* online working group from MIT/Montréal, at *Workshop ANR IA Camelot*, Montpellier, and at *Journées de Recherche en Apprentissage Frugal (JRAF)* 2023, Grenoble.
- Sylvain Chevallier, *Les sciences ouvertes en BCI des modèles riemanniens aux évaluations large échelle*, BCI & Neurotech Masterclass France.
- Flora Jay, invited keynote at *RECOMB-Genetics* (Turkey, April23) and at MCEB (Fr, June23).
- Flora Jay, invited speaker at *ISMB/ECCB 2023 (SFBI symposium)*, France (international conference).
- Marc Schoenauer, *Artificial Intelligence 2023: Trends and Risks*, Renault internal Seminar, 4 Jul. 2023.
- Beatriz Seoane Bartolomé, *Statistical Physics and Machine Learning*, Joachim Herz Foundation, Leipzig, Germany, 2023.
- Beatriz Seoane Bartolomé, *Institut Pascal meeting: Probabilistic sampling for physics: finding needles in a field of high-dimensional haystacks*, Institut Pascal, Paris-Saclay, 2023.
- Beatriz Seoane Bartolomé, *At the interface of physics, mathematics and artificial intelligence*, Pollica Physics Center, Italy, 2023.
- Beatriz Seoane Bartolomé, *Journées Interdisciplinaires de Physique Statistique*, LPTMS Paris-Saclay, France, 2023.
- Michele Sebag, plenary speaker, *ANITI Days*, Toulouse, Nov. 2023.
- Michèle Sebag, *Causal Colloquium*, Grenoble, May 2023.

11.1.5 Leadership within the scientific community

- Sylvain Chevallier: President of the academic society CORTICO, promoting the research in brain-computer interface; Executive Committee, Institut de Convergence DataIA;
- Flora Jay: member of GDR BiM science board (23-now)
- Marc Schoenauer: Advisory Board, ACM-SIGEVO, Special Interest Group on Evolutionary Computation; Founding President (since 2015), SPECIES, Society for the Promotion of Evolutionary Computation In Europe and Surroundings, that organizes the yearly series of conferences EvoStar.
- Michèle Sebag: Member of IRSN Scientific Council; Member of scientific council of the AMIES Labex; Area Chair NeurIPS, ECML-PKDD; Senior Meta-Reviewer ECAI

11.1.6 Scientific expertise

- Guillaume Charpiat: member of the Commission Scientifique (CS) at INRIA Saclay (PhD/post-docs grant allocations) until Nov. 2023.; and of graduate school GS ISN evaluation committee (PhD grants) 2023
- Guillaume Charpiat: Jean Zay (GENCI/IDRIS) committee member for resource allocation (GPU) demand expertise
- Guillaume Charpiat : MdC hiring committes: CentraleSupelec (“Tenure track in AI", 26/06/2023) and Evry University (10/05/2023)
- Guillaume Charpiat: jury of "Prix Doctorants STIC du Plateau de Saclay"
- Flora Jay, CR hiring committee, INRAE Toulouse
- Flora Jay, MdC hiring committee, LIX
- Flora Jay, Committee member for i-bio grant funding 2023
- Marc Schoenauer, Scientific Advisory Board, BCAM, Bilbao, Spain
- Marc Schoenauer, "Conseil Scientifique", IFPEN
- Marc Schoenauer, "Conseil Scientifique", Mines Paritech
- Marc Schoenauer, scientific coordinator of the IRT SystemX IA2 program (Artificial Intelligence for Augmented Engineering)
- Michele Sebag, "Conseil scientifique", IRSN
- Michele Sebag, UDOPIA jury (PhDs)
- Michele Sebag, FNRS (PhDs and Post-docs)
- Michele Sebag, professorship hiring committee, Univ. Paris Sorbonne
- Michele Sebag, European Commission’s Scientific Advice Mechanism (SAM), Oct. 2023.

11.1.7 Research administration

- Guillaume Charpiat: head of the Data Science department at LISN, Université Paris-Saclay
- Michele Sebag, elected member of Lab. Council, LISN, Université Paris-Saclay; Member of Council, Institut Pascal

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- Licence : Philippe Caillou, Computer Science for students in Accounting and Management, 192h, L1, IUT Sceaux, Univ. Paris Sud.
- Licence : François Landes, Mathematics for Computer Scientists, 27h, L2, Univ. Paris-Sud.
- Licence : François Landes, Introduction to Statistical Learning, 83h, L2, Univ. Paris-Sud.
- Licence : Isabelle Guyon: Introduction to Data Science, L1, Univ. Paris-Sud.
- Licence : Isabelle Guyon, Project: Resolution of mini-challenges (created by M2 students), L2, Univ. Paris-Saclay.
- Licence : Matthieu Kowalski, Signal Processing, L3, 25h, Univ. Paris-Saclay

- Master : François Landes, Foundational Principles of Machine Learning, 25h, M1 Recherche (AI track), U. Paris-Sud.
- Master : François Landes, Machine Learning, 42h, M2 Recherche, Univ. Paris-sud, physics department (PCS international Master)
- Master : Guillaume Charpiat, Deep Learning in Practice, 24h, M2 Recherche, MVA / Centrale-Supelec / DSBA
- Master : Guillaume Charpiat, Information Theory, 14h, M1 IA Paris-Sud.
- Master : Guillaume Charpiat, Introduction to Deep Learning, 3h, Eugloh.
- Master : Isabelle Guyon, Project: Creation of mini-challenges, M2, Univ. Paris-Sud.
- Master : Michèle Sebag, Deep Learning, 4h; Reinforcement Learning, 12h; M2 Recherche, U. Paris-Sud.
- Master : Flora Jay, Population genetics inference, 11h, M2, U PSaclay.
- Master : Flora Jay, Machine Learning in Genomics, 6h, M2, PSL. Some Principled Methods for Deep Reinforcement Learning and verification of deep learning: theory and practice, July 23rd.
- Master: Sylvain Chevallier, Machine learning algorithms, 21h, M1, Univ. Paris-Saclay.
- Master: Matthieu Kowalski, Signal Processing, 25h, M2, Univ. Paris-Saclay
- Master: Matthieu Kowalski, Sparse Coding, 36h, M2, Univ. Paris-Saclay
- Master : Beatriz Seoane, Applied Statistics, 25h, M1 Recherche (AI track), U. Paris-Saclay.
- Fall school : Flora Jay, Inference using full genome data, 7h, TUM, Germany.
- Workshop : Flora Jay, Teaching Workshop "Machine learning applied to population genomics." EMBO Workshop Population genomics: Background and tools, March 2023, Procida, Italy, $\simeq$ 1d teaching out of 1w workshop
- Continuing education (ie teaching in companies): Guillaume Charpiat, Machine Learning and Deep Learning, 8.5 days.

11.2.2 Supervision

- PhD Wenzhuo LIU, *Graph Neural Networks for Numerical Simulation of PDEs* [57], Mouadh Yagoubi (IRT SystemX) and Marc Schoenauer, Univ. Paris-Saclay, defended 19/04/2023.
- PhD Adrien PAVAO, *Theory and practice of challenge organization* [58], Isabelle Guyon, Univ. Paris-Saclay, defended on 5/12/23.
- PhD - Vincenzo SCHIMMENTI, *Temporal and spatial correlations in earthquake dynamics: physical modeling and data analysis* [59], François Landes and Alberto Rosso (LPTMS), defended on 23/11/2023.
- PhD - Haozhe SUN, *Modularity in Deep Learning* [60], Isabelle Guyon, defended on 19/12/2023.
- PhD in progress - Nicolas ATIENZA, *Causal Learning for Diagnostic Assistance*, from 1/12/2020, Johanne Cohen (LISN/Galac) and Michèle Sebag
- PhD in progress - Guillaume BIED, *Valorisation des Données pour la Recherche d'Emploi*, 1/10/2019, Bruno Crepon (CREST-ENSAE) and Philippe Caillou
- PhD in progress - Eva BOGUSLAWSKI *Congestion handling on Power Grid governed by complex automata*, from 1/05/22, Alessandro Leite, Mathieu Dussartre (RTE) and Marc Schoenauer

- PhD in progress - Romain EGELE, *Data-centric automated deep learning*, from 1/01/22, Isabelle Guyon/Michèle Sebag
- PhD in progress - Emmanuel GOUTIERRE, *Utilisation des méthodes d'apprentissage pour le réglage d'un accélérateur de particules*, from 1/12/2020, Johanne Cohen (LISN/Galac) and Michèle Sebag
- PhD in progress - Jérémy GUEZ, *Statistical inference of cultural transmission of reproductive success*, 1/10/2019, Evelyne Heyer (MNHN) and Flora Jay
- PhD in progress - Isabelle HOXHA, *Neurocognitive mechanisms of perceptual anticipation in decision-making*, from 1/10/2020, Michel-Ange Amorim (Faculté des Sciences du Sport), Sylvain Chevallier and Arnaud Delorme (CerCo)
- PhD in progress - Badr Youbi IDRISSI *Learning an invariant representation through continuously evolving data*, from 01/10/22, David Lopez-Paz (Meta) and Michèle Sebag
- PhD in progress - Armand LACOMBE, *Causal Modeling for Vocational training Recommendation*, 1/10/2019, Michele Sebag and Philippe Caillou
- PhD in progress - Romain LLORIA. *Geometrical Robust Blind Source Separation: Application to EEG classification*, from 1/11/2022, Frédéric Pascal (L2S), Florent Bouchard (L2S), and Sylvain Chevallier
- PhD in progress - Emmanuel MENIER, *Deep Learning for Reduced Order Modeling*, from 1/9/2020, Michele Alessandro Bucci and Marc Schoenauer, to be defended in January 2024
- PhD in progress - Thibault MONSEL, *Active Deep Learning for Complex Physical Systems*, 1/12/21, Alexandre Allauzen (LAMSADE), Guillaume Charpiat, Lionel Mathelin (LISN), Onofrio Semeraro (LISN)
- PhD in progress - Mathieu NASTORG, *Scalable GNN Strategies to Solve Poisson Pressure Problems in CFD Simulations*, 4/1/2021, Guillaume Charpiat and Michele Alessandro Bucci, to be defended in March 2024
- PhD in progress - Solal NATHAN, *Job recommendation, AI Ethics and Optimal Transport.*, 1/1/2023, Michèle Sebag and Philippe Caillou.
- PhD in progress - Francisco PEZZICOLI, *A new generation of Graph Neural Networks to tackle amorphous materials* from 1/11/2021, François Landes and Guillaume Charpiat.
- PhD in progress - Audrey POINSOT, *Causal Uncertainty Quantification under Partial Knowledge and Low Data Regimes*, from 1/03/22, Nicolas Chesneau (Ekimetrics), Guillaume Charpiat, Alessandro Leite, and Marc Schoenauer
- PhD in progress - Arnaud QUELIN, *Infering Human population history with approximated Bayesian computation and machine learning, from ancient and recent genoms' polymorphism data*, from 1/10/22, Frédéric Austerlitz (MNHN), Flora Jay
- PhD in progress - Cyriaque ROUSSELOT, *Spatio-temporal causal discovery – Application to modeling pesticides impact*, from 1/10/22, Michèle Sebag and Philippe Caillou
- PhD in progress - Antoine SZATKOWNIK, *Deep learning for population genetics*, from 1/10/22, Flora Jay, Burak Yelmen, Cyril Furtlehner and Guillaume Charpiat
- PhD in progress - Jazeps MEDINA-TRETMANIS, *Ancient Haplotype Inference for Detecting Population Structure*, Flora Jay and Emilia Huerta-Sanchez
- PhD in progress - Manon VERBOCKHAVEN, *Spotting and fixing expressivity bottlenecks*, from 11/2021, Sylvain Chevallier and Guillaume Charpiat
- PhD in progress - Maria Sayu YAMAMOTO, *Tackling the large variability of EEG data using Riemannian geometry toward reliable Brain-Computer Interfaces*, from 01-04-2021, Sylvain Chevallier and Fabien Lotte (INRIA Bordeaux Potioc)

- PhD in progress - Anaclara ALVEZ, *Scale-Equivariant Neural Networks* from 1/11/2023, François Landes and Cyril Furtlehner.
- PhD in progress - Jean-Baptiste MALAGNOUX, *Convolutional Dictionary Learning and time-frequency Nonmegative Matrix Factorization*, from 1/10/2022, Matthieu Kowalski
- PhD in progress - Florent NICHEL, *Deep Learning for Dictionary Learning*, from 1/10/2022, Matthieu Kowalski and Thomas Moreau (Inria Mind)
- PhD in progress - Nicolas BÉREUX *interpretability and pattern extraction in Restricted Boltzmann Machines* from 1/11/2023, Beatriz Seoane Bartolome, Cyril Furtlehner.
- PhD in progress: Nilo SCHWENCKE *Modélisation des batteries Lithium-Ion par Physics-Informed Neural Networks* from 1/09/2023, Cyril Furtlehner

11.2.3 Juries

- Guillaume Charpiat, PhD jury member, Camille Garcin, Université de Montpellier / IMAG, 29/09/2023
- Flora Jay, PhD jury member, Jeanne Trinquier, "Data-driven generative modeling of protein sequence landscapes and beyond", 13/09/23
- Matthieu Kowalski, PhD reviewer: Félix Mathieu, Telecom ParisTech, 28/11/2023; PhD reviewer: Hoang Trieu Vy LE, ENS Lyon, 13/12/2023
- François Landes: head of the M1 and M2 AI track selection comittee (M1 and M2 combined: 1000+ applicants per year). Also head of the scholarship short-listing comittee.
- Marc Schoenauer, HDR jury member, Xavier Marsault, INSA Lyon, ENSAL, 27/11/23; PhD jury member: Federica Granese, IPP, 11/4/2023; Leo Pouy, Univ. Paris-Saclay, 6/12/23; Kaitlin Maile, 16/12/2022, IRIT, Toulouse, 4.10.23.
- Michèle Sebag: Jury member, Institut Pascal; PhD Emilien Baroux, LMS-X; Yutin Feng, LISN, U. PSaclay; Antoine Caillon, IRCAM; Eduardo Brando, U. St Etienne.

11.3 Popularization

- Flora Jay : “Une population fantôme découverte dans le génome des civilisations préhispaniques", entretien radiophonique, Le Journal des sciences, France Culture, 18/05/23

11.3.1 Internal or external Inria responsibilities

- Marc Schoenauer, Deputy Research Director in charge of AI
- Marc Schoenauer, sherpa for Inria as pilot institution of the PEPR-IA (together with François Terrier for CEA and Jamal Atif for CNRS)
- Marc Schoenauer, scientific coordinator of ICT49 CSA Adra-e (coordinated by Inria)

12 Scientific production

12.1 Major publications

[1] E. Agoritsas, G. Catania, A. Decelle and B. Seoane. ‘Explaining the effects of non-convergent sampling in the training of Energy-Based Models’. In: *Proceedings of the 40th International Conference on Machine Learning, PMLR 202:322-336, 2023*. ICML 2023 - 40th International Conference on Machine Learning. Honolulu (Hawaii), United States, 23rd July 2023. URL: https://inria.hal.science/hal-04344101.

[2] G. Charpiat, N. Girard, L. Felardos and Y. Tarabalka. 'Input Similarity from the Neural Network Perspective'. In: NeurIPS 2019 - 33th Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 8th Dec. 2019. URL: https://hal.science/hal-02394647.

[3] A. Decelle, C. Furtlehner and B. Seoane. 'Equilibrium and non-Equilibrium regimes in the learning of Restricted Boltzmann Machines'. In: NeurIPS 2021. Proceedings NeurIPS 2021. Vancouver, United States, 6th Dec. 2021. URL: https://hal.science/hal-03518796.

[4] B. Donon, W. Liu, A. Marot, Z. Liu, I. Guyon and M. Schoenauer. 'Deep Statistical Solvers'. In: NeurIPS 2020 - 34th Conference on Neural Information Processing Systems. Vancouver / Virtuel, Canada, 6th Dec. 2020. URL: https://hal.inria.fr/hal-02974541.

[5] D. Kalainathan, O. Goudet, I. Guyon, D. Lopez-Paz and M. Sebag. 'Structural Agnostic Modeling: Adversarial Learning of Causal Graphs'. In: *Journal of Machine Learning Research* (2022). URL: https://hal.science/hal-03831338.

[6] H. Rakotoarison, L. Milijaona, A. Rasoanaivo, M. Sebag and M. Schoenauer. 'Learning Meta-features for AutoML'. In: ICLR 2022 - International Conference on Learning Representations (spotlight). Virtual, United States, 26th Apr. 2022. URL: https://hal.inria.fr/hal-03583789.

12.2 Publications of the year

International journals

[7] S. Amrouche, L. Basara, P. Calafiura, D. Emeliyanov, V. Estrade, S. Farrell, C. Germain, V. Vava Gligorov, T. Golling, S. Gorbunov, H. Gray, I. Guyon, M. Hushchyn, V. Innocente, M. Kiehn, M. Kunze, E. Moyse, D. Rousseau, A. Salzburger, A. Ustyuzhanin and J.-R. Vlimant. 'The Tracking Machine Learning challenge : Throughput phase'. In: *Computing and Software for Big Science* 7.1 (2023), p. 1. DOI: 10.1007/s41781-023-00094-w. URL: https://inria.hal.science/hal-03159824.

[8] A. Andreev, G. H. Cattan, S. Chevallier and Q. Barthélemy. 'pyRiemann-qiskit: A Sandbox for Quantum Classification Experiments with Riemannian Geometry'. In: *Research Ideas and Outcomes* 9 (20th Mar. 2023). DOI: 10.3897/rio.9.e101006. URL: https://hal.science/hal-04040814.

[9] M. Baity-Jesi, E. Calore, A. Cruz, L. Fernandez, J. Gil-Narvion, I. Gonzalez-Adalid Pemartin, A. Gordillo-Guerrero, D. Iñiguez, A. Maiorano, E. Marinari, V. Martin-Mayor, J. Moreno-Gordo, A. Muñoz Sudupe, D. Navarro, I. Paga, G. Parisi, S. Perez-Gaviro, F. Ricci-Tersenghi, J. Ruiz-Lorenzo, S. Schifano, B. Seoane, A. Tarancon and D. Yllanes. 'Memory and rejuvenation effects in spin glasses are governed by more than one length scale'. In: *Nature Physics* 19.7 (27th Apr. 2023), pp. 978–985. DOI: 10.1038/s41567-023-02014-6. URL: https://inria.hal.science/hal-04341897.

[10] M. Baity-Jesi, E. Calore, A. Cruz, L. A. Fernández, J. M. Gil-Narvión, I. González-Adalid Pemartín, A. Gordillo-Guerrero, D. Íñiguez, A. Maiorano, E. Marinari, V. Martín-Mayor, J. Moreno-Gordo, A. Muñoz Sudupe, D. Navarro, I. Paga, G. Parisi, S. Pérez-Gaviro, F. Ricci-Tersenghi, J. J. Ruiz-Lorenzo, S. F. Schifano, B. Seoane, A. Tarancón and D. Yllanes. 'Multifractality in spin glasses'. In: *Proceedings of the National Academy of Sciences of the United States of America* 121.2 (7th June 2023). DOI: 10.1073/pnas.2312880120. URL: https://inria.hal.science/hal-04344056.

[11] A. Barra, G. Catania, A. Decelle and B. Seoane. 'Thermodynamics of bidirectional associative memories'. In: *Journal of Physics A: Mathematical and Theoretical* 56.20 (2nd May 2023), p. 205005. DOI: 10.1088/1751-8121/accc60. URL: https://inria.hal.science/hal-04344065.

[12] N. Béreux, A. Decelle, C. Furtlehner and B. Seoane. 'Learning a restricted Boltzmann machine using biased Monte Carlo sampling'. In: *SciPost Physics* 14.3 (14th Mar. 2023), p. 032. DOI: 10.21468/SciPostPhys.14.3.032. URL: https://inria.hal.science/hal-04341856.

[13] M.-C. Corsi, P. Sorrentino, D. P. Schwartz, N. George, L. L. Gollo, S. Chevallier, L. Hugueville, A. E. Kahn, S. Dupont, D. S. Bassett, V. Jirsa and F. de Vico Fallani. 'Measuring Neuronal Avalanches to inform Brain-Computer Interfaces'. In: *iScience* 27.1 (13th Dec. 2023), p. 108734. DOI: 10.1016/j.isci.2023.108734. URL: https://inria.hal.science/hal-04345847.

[14] A. Decelle, B. Seoane and L. Rosset. 'Unsupervised hierarchical clustering using the learning dynamics of restricted Boltzmann machines'. In: *Physical Review E* 108.1 (7th July 2023), p. 014110. DOI: 10.1103/PhysRevE.108.014110. URL: https://inria.hal.science/hal-04341820.

[15] D. Delle Donne, M. Kowalski and L. Liberti. 'A Novel Integer Linear Programming Approach for Global L0 Minimization'. In: *Journal of Machine Learning Research* (23rd Dec. 2023). URL: https://hal.science/hal-04372309.

[16] N. Díaz-Rodríguez, R. Binkytė, W. Bakkali, S. Bookseller, P. Tubaro, A. Bacevičius, S. Zhioua and R. Chatila. 'Gender and sex bias in COVID-19 epidemiological data through the lens of causality'. In: *Information Processing and Management* 60.3 (May 2023), p. 103276. DOI: 10.1016/j.ipm.2023.103276. URL: https://hal.science/hal-03961804.

[17] C. Furtlehner. 'Free Dynamics of Feature Learning Processes'. In: *Journal of Statistical Physics*. Journal of Statistical Physics 190.51 (10th Jan. 2023), p. 42. DOI: 10.1007/s10955-022-03064-5. URL: https://inria.hal.science/hal-03878500.

[18] C. Gain, B. Rhoné, P. Cubry, I. Salazar, F. Forbes, Y. Vigouroux, F. Jay and O. François. 'A Quantitative Theory for Genomic Offset Statistics'. In: *Molecular Biology and Evolution* 40.6 (2023), msad140. DOI: 10.1093/molbev/msad140. URL: https://hal.science/hal-04243951.

[19] J. Guez, G. Achaz, F. Bienvenu, J. Cury, B. Toupance, É. Heyer, F. Jay and F. Austerlitz. 'Cultural transmission of reproductive success impacts genomic diversity, coalescent tree topologies and demographic inferences'. In: *Genetics* 223.4 (14th Feb. 2023), iyad007. DOI: 10.1093/genetics/iyad007. URL: https://hal.science/hal-03875721.

[20] S. Hartley, G. Bao, A. Russo, M. Zagdoun, S. Chevallier, F. Lofaso, A. Leotard and E. Azabou. 'Self-administered non-invasive vagus nerve stimulation therapy for severe pharmacoresistant restless legs syndrome: outcomes at 6 months'. In: *Journal of Sleep Research* (1st Jan. 2023). DOI: 10.1111/jsr.14066. URL: https://hal.science/hal-04258081.

[21] I. Hoxha, S. Chevallier, M. Ciarchi, S. Glasauer, A. Delorme and M. A. Amorim. 'Accounting for endogenous effects in decision-making with a non-linear diffusion decision model'. In: *Scientific Reports* 13.1 (18th Apr. 2023), p. 6323. DOI: 10.1038/s41598-023-32841-9. URL: https://hal.science/hal-04290041.

[22] A. Lacan, M. Sebag and B. Hanczar. 'GAN-based data augmentation for transcriptomics: survey and comparative assessment'. In: *Bioinformatics* 39.Supplement_1 (1st June 2023), pp. i111–i120. DOI: 10.1093/bioinformatics/btad239. URL: https://univ-evry.hal.science/hal-04263316.

[23] R. Maulik, R. Egele, K. Raghavan and P. Balaprakash. 'Quantifying uncertainty for deep learning based forecasting and flow-reconstruction using neural architecture search ensembles'. In: *Physica D: Nonlinear Phenomena* 454 (Nov. 2023), p. 133852. DOI: 10.1016/j.physd.2023.133852. URL: https://hal.science/hal-04219331.

[24] E. Menier, M. A. Bucci, M. Yagoubi, L. Mathelin and M. Schoenauer. 'CD-ROM: Complementary Deep-Reduced Order Model'. In: *Computer Methods in Applied Mechanics and Engineering* 410 (1st May 2023). URL: https://hal.science/hal-03846122.

[25] I. Overcast, G. Achaz, R. Aguilée, C. Andújar, P. Arribas, T. J. Creedy, E. P. Economo, R. S. Etienne, R. Gillespie, C. Jacquet, F. Jay, S. Kennedy, H. Krehenwinkel, A. Lambert, E. Meramveliotakis, V. Noguerales, B. Perez-lamarque, G. Roderick, H. Rogers, M. Ruffley, I. Sanmartin, A. P. Vogler, A. Papadopoulou, B. C. Emerson and H. Morlon. 'Towards a genetic theory of island biogeography: Inferring processes from multidimensional community-scale data'. In: *Global Ecology and Biogeography* 32 (Jan. 2023), pp. 4–23. DOI: 10.1111/geb.13604. URL: https://inserm.hal.science/inserm-04270149.

[26] I. Paga, Q. Zhai, M. Baity-Jesi, E. Calore, A. Cruz, C. Cummings, L. A. Fernandez, J. M. Gil-Narvion, I. G.-A. Pemartin, A. Gordillo-Guerrero, D. Iñiguez, G. Kenning, A. Maiorano, E. Marinari, V. Martin-Mayor, J. Moreno-Gordo, A. Muñoz-Sudupe, D. Navarro, R. Orbach, G. Parisi, S. Perez-Gaviro, F. Ricci-Tersenghi, J. Ruiz-Lorenzo, S. F. Schifano, D. Schlagel, B. Seoane, A. Tarancon and D. Yllanes. 'Superposition principle and nonlinear response in spin glasses'. In: *Physical Review B* 107 (23rd June 2023), p. 214436. DOI: 10.1103/PhysRevB.107.214436. URL: https://inria.hal.science/hal-04344051.

[27] T. Sanchez, E. Madison Bray, P. Jobic, J. Guez, A.-C. Letournel, G. Charpiat, J. Cury and F. Jay. 'dnadna: a deep learning framework for population genetics inference'. In: *Bioinformatics* 39.1 (Jan. 2023). DOI: 10.1093/bioinformatics/btac765. URL: https://hal.science/hal-03352910.

[28] V. Villa-Islas, A. Izarraras-Gomez, M. Larena, E. M. P. Campos, M. Sandoval-Velasco, J. E. Rodríguez-Rodríguez, M. Bravo-Lopez, B. Moguel, R. Fregel, E. Garfias-Morales, J. Medina Tretmanis, D. A. Velázquez-Ramírez, A. Herrera-Muñóz, K. Sandoval, M. Nieves-Colón, G. Zepeda García Moreno, F. Villanea, E. F. V. Medina, R. Aguayo-Haro, C. Valdiosera, A. Ioannidis, A. Moreno-Estrada, F. Jay, E. Huerta-Sanchez, J. V. Moreno-Mayar, F. Sánchez-Quinto and M. Ávila-Arcos. 'Demographic history and genetic structure in pre-Hispanic Central Mexico'. In: *Science* 380.6645 (12th May 2023). DOI: 10.1126/science.add6142. URL: https://hal.science/hal-04244008.

[29] M. S. Yamamoto, K. Sadatnejad, T. Tanaka, M. R. Islam, F. Dehais, Y. Tanaka and F. Lotte. 'Modeling complex EEG data distribution on the Riemannian manifold toward outlier detection and multimodal classification'. In: *IEEE Transactions on Biomedical Engineering* (2023). DOI: 10.1109/TBME.2023.3295769. URL: https://hal.science/hal-04181396.

[30] B. Yelmen, A. Decelle, L. L. Boulos, A. Szatkownik, C. Furtlehner, G. Charpiat and F. Jay. 'Deep convolutional and conditional neural networks for large-scale genomic data generation'. In: *PLoS Computational Biology* 19.10 (30th Oct. 2023), e1011584. DOI: 10.1101/2023.03.07.530442. URL: https://hal.science/hal-04244818.

[31] B. Yelmen and F. Jay. 'An Overview of Deep Generative Models in Functional and Evolutionary Genomics'. In: *Annual Review of Biomedical Data Science* 6.1 (10th Aug. 2023), pp. 173–189. DOI: 10.1146/annurev-biodatasci-020722-115651. URL: https://hal.science/hal-04243980.

[32] A. Zammouri, A. A. Moussa and S. Chevallier. 'Use of cognitive load measurements to design a new architecture of intelligent learning systems'. In: *Expert Systems with Applications* 237 (24th Aug. 2023), p. 121253. DOI: 10.1016/j.eswa.2023.121253. URL: https://universite-paris-saclay.hal.science/hal-04350891.

Invited conferences

[33] G. Bied, C. Gaillac, M. Hoffmann, S. Nathan, P. Caillou, B. Crépon and M. Sebag. 'Gender fairness in job recommendation: a case study'. In: AI for HR and Public Employment Services. Ghent (BE), Belgium, 9th Feb. 2023. URL: https://inria.hal.science/hal-04025006.

[34] G. Bied, S. Nathan, E. Perennes, C. Gaillac, P. Caillou, B. Crépon and M. Sebag. 'Using Data from job seekers, job offers and past hirings to learn a Job Recommender System: the VADORE Project'. In: AI for HR and Public Employment Services. Ghent (BE), Belgium, 9th Feb. 2023. URL: https://inria.hal.science/hal-04025000.

International peer-reviewed conferences

[35] E. Agoritsas, G. Catania, A. Decelle and B. Seoane. 'Explaining the effects of non-convergent sampling in the training of Energy-Based Models'. In: *Proceedings of the 40th International Conference on Machine Learning, PMLR 202:322-336, 2023.* ICML 2023 - 40th International Conference on Machine Learning. Honolulu (Hawaii), United States, 23rd July 2023. URL: https://inria.hal.science/hal-04344101.

[36] P. Barbault, M. Kowalski and C. Soussen. 'Estimation du MAP Marginal d'un signal Bernoulli-Gaussien: une approche par relaxation continue'. In: GRETSI 2023 - XXIXème Colloque Francophone de Traitement du Signal et des Images. Grenoble, France, Aug. 2023, pp. 1–4. URL: https://centralesupelec.hal.science/hal-04383981.

[37] P. Barbault, M. Kowalski and C. Soussen. 'Marginal MAP estimation of a Bernoulli-Gaussian signal: continuous relaxation approach'. In: EUSIPCO 2023 - the 31st European Signal Processing Conference. Helsinki, Finland, 4th Sept. 2023, pp. 1833–1837. URL: https://centralesupelec.hal.science/hal-04383978.

[38] G. Bied, C. Gaillac, M. Hoffmann, P. Caillou, B. Crépon, S. Nathan and M. Sebag. 'Fairness in job recommendations: estimating, explaining, and reducing gender gaps'. In: *Proceedings of the 1st Workshop on Fairness and Bias in AI co-located with 26th European Conference on Artificial Intelligence (ECAI 2023)*. AEQUITAS 2023 - First AEQUITAS Workshop on Fairness and Bias in AI | co-located with ECAI 2023. Vol. 3523. Krakow, Poland: CEUR-WS.org, 2023. URL: https://inria.hal.science/hal-04438512.

[39] G. Bied, S. Nathan, E. Perennes, M. Hoffmann, P. Caillou, B. Crépon, C. Gaillac and M. Sebag. 'Toward Job Recommendation for All'. In: IJCAI 2023 - The 32nd International Joint Conference on Artificial Intelligence. Macau, China: International Joint Conferences on Artificial Intelligence Organization, 19th Aug. 2023, pp. 5906–5914. DOI: 10.24963/ijcai.2023/655. URL: https://hal.science/hal-04245528.

[40] C. Bonet, B. Malézieux, A. Rakotomamonjy, L. Drumetz, T. Moreau, M. Kowalski and N. Courty. 'Sliced-Wasserstein on Symmetric Positive Definite Matrices for M/EEG Signals'. In: ICML 2023 - Fortieth International Conference on Machine Learning. Honololu, United States, 2023, pp. 1–29. URL: https://cnrs.hal.science/hal-04250921.

[41] M.-C. Corsi, S. Chevallier, F. de Vico Fallani and F. Yger. 'Ensemble of Riemannian Classifiers for Multimodal Data: FUCONE Approach for M/EEG Data'. In: ISBI 2023 - IEEE International Symposium on Biomedical Imaging. Cartagena de Indias, Colombia, 18th Apr. 2023. URL: https://inria.hal.science/hal-04140126.

[42] R. Egele, I. Guyon, Y. Sun and P. Balaprakash. 'Is One Epoch All You Need For Multi-Fidelity Hyperparameter Optimization?' In: ESANN 2023 - 31th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges / Hybrid, Belgium, 4th Oct. 2023. URL: https://hal.science/hal-04219301.

[43] R. Egele, I. Guyon, V. Vishwanath and P. Balaprakash. 'Asynchronous Decentralized Bayesian Optimization for Large Scale Hyperparameter Optimization'. In: eS 2023 - 19th IEEE International Conference on e-Science. Limassol (Chypre), Cyprus, 9th Oct. 2023. URL: https://hal.science/hal-04219312.

[44] E. Goutierre, C. Bruni, J. Cohen, H. Guler and M. Sebag. 'Physics-aware modelling of an accelerated particle cloud'. In: MLPS 2023 - Machine Learning and the Physical Sciences Workshop 23023 - At the 37th conference on Neural Information Processing Systems (NeurIPS). New Orleans, United States, 2023. URL: https://hal.science/hal-04396175.

[45] T. G. H. Khuong and B. K. Rachmat. 'Auto-survey Challenge: Advancing the Frontiers of Automated Literature Review'. In: JDSE 2023 - 8th Junior Conference on Data Science and Engineering. Orsay, France, 27th Sept. 2023. URL: https://inria.hal.science/hal-04206578.

[46] H. Sun and I. Guyon. 'Modularity in Deep Learning: A Survey'. In: *Lecture Notes in Networks and Systems*. SAI 2023 - Intelligent Computing Conference. Vol. LNNS-739. Intelligent Computing - Proceedings of the 2023 Computing Conference, Volume 2. London, United Kingdom: Springer Nature Switzerland, 20th Aug. 2023, pp. 561–595. DOI: 10.1007/978-3-031-37963-5_40. URL: https://hal.science/hal-04224278.

[47] H. Sun, I. Guyon, F. Mohr and H. Tabia. 'RRR-Net: Reusing, Reducing, and Recycling a Deep Backbone Network'. In: IJCNN 2023 - International Joint Conference on Neural Networks. Gold Coast, Australia: IEEE, 18th June 2023, pp. 1–9. DOI: 10.1109/IJCNN54540.2023.10191770. URL: https://hal.science/hal-04224281.

[48] A. Szatkownik, C. Furtlehner, G. Charpiat, B. Yelmen and F. Jay. 'Towards creating longer genetic sequences with GANs: Generation in principal component space'. In: *PMLR*. MLCB 2023 - 18th Conference on Machine Learning in Computational Biology. Seattle, United States, 2024. URL: https://hal.science/hal-04419057.

[49] M. S. Yamamoto, A. Mellot, S. Chevallier and F. Lotte. 'Novel SPD matrix representations considering cross-frequency coupling for EEG classification using Riemannian geometry'. In: EUSIPCO 2023 - The 31st European Signal Processing Conference. Helsinki, Finland, 4th Sept. 2023. URL: https://hal.science/hal-04131609.

National peer-reviewed Conferences

[50] G. Bied, E. Perennes, S. Nathan, V. Naya, P. Caillou, B. Crepon, C. Gaillac and M. Sebag. 'RECTO : REcommandation diminuant la Congestion par Transport Optimal'. In: *Actes de la 9ème Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (APIA) PFIA 2023*. 9ème Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle APIA@PFIA2023. Vol. APIA. 2023. Strasbourg, France: AFIA-Association Française pour l'Intelligence Artificielle, 7th July 2023, pp. 89–98. URL: https://hal.science/hal-04158566.

Conferences without proceedings

[51] B. Aristimunha, R. Y. de Camargo, S. Chevallier, A. G. Thomas, O. Lucena, J. Cardoso, W. H. L. Pinaya and J. Dafflon. 'Synthetic Sleep EEG Signal Generation using Latent Diffusion Models'. In: DGM4H 2023 - 1st Workshop on Deep Generative Models for Health at NeurIPS 2023. New orleans, United States, 15th Dec. 2023. URL: https://universite-paris-saclay.hal.science/hal-04350867.

[52] A. Carbone, A. Decelle, L. Rosset and B. Seoane. 'Fast and Functional structured data generator'. In: *https://openreview.net/forum?id=uXkfPvjYeM*. ICML 2023 - Workshop on Structured Probabilistic Inference & Generative Modeling. Honolulu, United States, 28th July 2023. URL: https://inria.hal.science/hal-04342214.

[53] E. Menier, M. A. Bucci, M. Yagoubi, L. Mathelin and M. Schoenauer. 'CD-ROM: Complemented Deep-Reduced Order Model'. In: MORTech 2023 - 6th International Workshop on Model Order Reduction. Saclay, France, 22nd Nov. 2023. URL: https://hal.science/hal-04406567.

[54] T. Monsel, L. Mathelin, O. Semeraro and G. Charpiat. 'End-to-end learning of dynamical systems with the Mori-Zwanzig formalism'. In: CSE 2023 - SIAM Computational Science and Engineering. Amsterdam, Netherlands, 6th Mar. 2023. URL: https://hal.science/hal-04406551.

Scientific book chapters

[55] H. J. Escalante Balderas, I. Guyon, A. Howard, W. Reade and S. Treguer. 'Challenge design roadmap'. In: *AI Competitions and Benchmarks: The Science Behind the Contests*. 2023. URL: https://hal.science/hal-04333280.

[56] V. Martín-Mayor, J. Ruiz-Lorenzo, B. Seoane and A. P. Young. 'Numerical Simulations and Replica Symmetry Breaking'. In: *Spin Glass Theory and Far Beyond*. WORLD SCIENTIFIC, 3rd Aug. 2023, pp. 69–93. DOI: 10.1142/9789811273926_0005. URL: https://inria.hal.science/hal-04341849.

Doctoral dissertations and habilitation theses

[57] W. Liu. 'Deep Graph Neural Networks for Numerical Simulation of PDEs'. Université Paris-Saclay, 19th Apr. 2023. URL: https://hal.science/tel-04156859.

[58] A. Pavão. 'Methodology for Design and Analysis of Machine Learning Competitions'. Université Paris-Saclay, 5th Dec. 2023. URL: https://inria.hal.science/tel-04401932.

[59] V. Schimmenti. 'Temporal and spatial correlations in earthquake dynamics : physical modeling and data analysis'. Université Paris-Saclay, 23rd Nov. 2023. URL: https://theses.hal.science/tel-04344778.

[60] H. Sun. 'Modularity in deep learning'. Université Paris-Saclay, 19th Dec. 2023. URL: https://theses.hal.science/tel-04418605.

Reports & preprints

[61] B. H. Belay, I. Guyon, T. Mengiste, B. Tilahun, M. Liwicki, T. Tegegne and R. Egele. *HHD-Ethiopic A Historical Handwritten Dataset for Ethiopic OCR with Baseline Models and Human-level Performance*. 29th Sept. 2023. URL: https://inria.hal.science/hal-04223188.

[62] A. Carbone, A. Decelle, L. Rosset and B. Seoane. *Fast and Functional Structured Data Generators Rooted in Out-of-Equilibrium Physics*. 13th July 2023. URL: https://inria.hal.science/hal-04344041.

[63] G. Catania, A. Decelle and B. Seoane. *The Copycat Perceptron: Smashing Barriers Through Collective Learning*. 14th Dec. 2023. URL: https://inria.hal.science/hal-04344030.

[64] R. Chacko, F. P. Landes, G. Biroli, O. Dauchot, A. Liu and D. Reichman. *Dynamical Facilitation Governs the Equilibration Dynamics of Glasses*. 2023. DOI: 10.48550/arXiv.2312.15069. URL: https://inria.hal.science/hal-04379217.

[65] S. Dash, I. Lyngaas, J. Yin, X. Wang, R. Egele, G. Cong, F. Wang and P. Balaprakash. *Optimizing Distributed Training on Frontier for Large Language Models*. 26th Sept. 2023. URL: https://inria.hal.science/hal-04393799.

[66] A. Decelle, C. Furtlehner, A. d. J. N. Gómez and B. Seoane. *Inferring effective couplings with Restricted Boltzmann Machines*. 13th Dec. 2023. URL: https://inria.hal.science/hal-04342280.

[67] S. Dong, K. Uemura, A. Fujii, S. Chang, Y. Koyanagi, K. Maruhashi and M. Sebag. *Learning Large Causal Structures from Inverse Covariance Matrix via Matrix Decomposition*. 23rd Oct. 2023. URL: https://inria.hal.science/hal-03885791.

[68] R. Egele, T. Chang, Y. Sun, V. Vishwanath and P. Balaprakash. *Parallel Multi-Objective Hyperparameter Optimization with Uniform Normalization and Bounded Objectives*. 26th Sept. 2023. URL: https://hal.science/hal-04219318.

[69] L. Felardos, J. Hénin and G. Charpiat. *Designing losses for data-free training of normalizing flows on Boltzmann distributions*. 12th Jan. 2023. URL: https://hal.science/hal-03936982.

[70] G. Jung, R. M. Alkemade, V. Bapst, D. Coslovich, L. Filion, F. P. Landes, A. Liu, F. S. Pezzicoli, H. Shiba, G. Volpe, F. Zamponi, L. Berthier and G. Biroli. *Roadmap on machine learning glassy liquids*. 14th Dec. 2023. URL: https://hal.science/hal-04344224.

[71] E. Menier, S. Kaltenbach, M. Yagoubi, M. Schoenauer and P. Koumoutsakos. *Interpretable learning of effective dynamics for multiscale systems*. 11th Sept. 2023. URL: https://inria.hal.science/hal-04392797.

[72] T. Monsel, O. Semeraro, L. Mathelin and G. Charpiat. *Neural State-Dependent Delay Differential Equations*. 22nd June 2023. URL: https://hal.science/hal-04125875.

[73] M. Nastorg, M. A. Bucci, T. Faney, J.-M. Gratien, G. Charpiat and M. Schoenauer. *An Implicit GNN Solver for Poisson-like problems*. 14th Feb. 2023. URL: https://hal.science/hal-03970501.

[74] M. Nastorg, J.-M. Gratien, T. Faney, M. A. Bucci, G. Charpiat and M. Schoenauer. *Multi-Level GNN Preconditioner for Solving Large Scale Problems*. 6th Feb. 2024. URL: https://inria.hal.science/hal-04447099.

[75] I. Paga, J. He, M. Baity-Jesi, E. Calore, A. Cruz, L. A. Fernandez, J. M. Gil-Narvion, I. G.-A. Pemartin, A. Gordillo-Guerrero, D. Iñiguez, A. Maiorano, E. Marinari, V. Martin-Mayor, J. Moreno-Gordo, A. M. Sudupe, D. Navarro, R. Orbach, G. Parisi, S. Perez-Gaviro, F. Ricci-Tersenghi, J. Ruiz-Lorenzo, S. F. Schifano, D. Schlagel, B. Seoane, A. Tarancon and D. Yllanes. *Quantifying memory in spin glasses*. 5th July 2023. URL: https://inria.hal.science/hal-04344048.

[76] A. Pavão. *Hands-on tutorial on how to create your own challenge or benchmark.* 25th Jan. 2024. URL: https://inria.hal.science/hal-04415747.

[77] A. Poinsot and A. Leite. *A Guide for Practical Use of ADMG Causal Data Augmentation.* 2023. URL: https://hal.science/hal-04250612.

[78] M. Quattromini, M. A. Bucci, S. Cherubini and O. Semeraro. *Operator learning of RANS equations: a Graph Neural Network closure model.* 2023. DOI: 10.48550/arXiv.2303.03806. URL: https://hal.science/hal-04290982.

[79] V. M. Schimmenti, G. Petrillo, A. Rosso and F. P. Landes. *Assessing the predicting power of GPS data for aftershocks forecasting.* 17th May 2023. URL: https://hal.science/hal-04352596.

[80] J. M. Tretmanis, F. Jay, M. Ávila-Arcos and E. Huerta-Sanchez. *Simulation-based Benchmarking of Ancient Haplotype Inference for Detecting Population Structure.* 29th Sept. 2023. DOI: 10.1101/2023.09.28.560049. URL: https://hal.science/hal-04244011.

[81] W.-W. Tu and A. Pavão. *Special designs and competition protocols.* 25th Jan. 2024. URL: https://inria.hal.science/hal-04416565.

[82] F. Villanea, D. Peede, E. Kaufman, V. Añorve-Garibay, K. Witt, V. Villa-Islas, R. Zeloni, D. Marnetto, P. Moorjani, F. Jay, P. Valdmanis, M. Ávila-Arcos and E. Huerta-Sánchez. *The MUC19 gene in Denisovans, Neanderthals, and Modern Humans: An Evolutionary History of Recurrent Introgression and Natural Selection.* 25th Sept. 2023. DOI: 10.1101/2023.09.25.559202. URL: https://hal.science/hal-04244022.

Other scientific publications

[83] E. Boguslawski, A. Leite, M. Schoenauer, M. Dussartre and B. Donnot. 'A Decision Support Assistant to Operate a Power Grids with Zonal Automatons'. In: DTU PES Summer School 2023. Future Energy Systems: Advances in OR and AI. Lingby, Denmark, 19th June 2023. URL: https://hal.science/hal-04396136.

[84] E. Boguslawski, A. Leite, M. Schoenauer, M. Dussartre and B. Donnot. 'Decentralized Partially Observable Markov Decision Process for Power Grid Management'. In: Reinforcement Learning Summer School 2023. Barcelone, Spain, 26th June 2023. URL: https://hal.science/hal-04396121.

[85] I. Carrara, B. Aristimunha, S. Chevallier, M.-C. Corsi, T. Papadopoulo and M.-C. Corsi. 'Holographic EEG: multi-view deep learning for BCI'. In: Journée CORTICO 2023 - COllectif pour la Recherche Transdisciplinaire sur les Interfaces Cerveau-Ordinateur. Paris, France, 9th May 2023. URL: https://hal.science/hal-04095900.

[86] I. Carrara, B. Aristimunha, M.-C. Corsi, R. Y. D. Camargo, S. Chevallier and T. Papadopoulo. 'Augmented SPDNet: Second-Order Neural Network for Motor Imagery-Based BCI'. In: Soph.IA summit 2023. Sophia Antipolis, France, 22nd Nov. 2023. URL: https://inria.hal.science/hal-04308549.

[87] R. Frateur, S. Chevallier, F. Yger and M.-C. Corsi. 'Dimensionality Reduction and Frequency Bin Optimization To Improve a Riemannian-based Classification Pipeline'. In: CORTICO 2023 - Journées COllectif pour la Recherche Transdisciplinaire sur les Interfaces Cerveau-Ordinateur. Paris, France, 9th May 2023. URL: https://inria.hal.science/hal-04140107.

[88] E. Goutierre, C. Bruni, J. Cohen, H. Guler and M. Sebag. 'Surrogate Model for Linear Accelerator: A fast Neural Network approximation of ThomX's simulator'. In: IPAC 2023 - 14th International Particle Accelerator Conference. Vol. IPAC2023. Venice, Italy: JACoW Publishing, 2023, pp. 4514–4517. DOI: 10.18429/JACoW-IPAC2023-THPL039. URL: https://hal.science/hal-04396183.

[89] E. Menier, K. Sebastian, M. Yagoubi, M. Schoenauer and K. Petros. 'Interpretable reduced order modeling using neural autoencoders'. In: MORTech 2023 – 6th International Workshop on Model Reduction Techniques. Gif-sur Yvette, France, 22nd Nov. 2023. URL: https://inria.hal.science/hal-04296372.

[90] A. Quelin, J. Guez, F. Petit, F. Jay and F. Austerlitz. 'Inference of the Cultural Transmission of Reproductive Success from human genomic data: ABC and machine learning methods'. In: Alphy/AIEM 2023 - Rencontres Alphy & AIEM. Grenoble, France, 23rd Jan. 2023. URL: https://hal.science/hal-03960408.

[91] M. S. Yamamoto, S. Chevallier and F. Lotte. 'Effectiveness of cross-frequency phase-amplitude covariance as additional features for Riemannian BCIs'. In: BCI 2023 - 10th International BCI Meeting Balancing Innovation and Translation. Brussels, Belgium, 6th June 2023. DOI: 10.3217/978-3-85125-962-9-179. URL: https://hal.science/hal-04181391.

12.3 Other

Softwares

[92] [SW] A.-C. Letournel, P. Caillou, J.-D. Fekete, M. Sebag, H. Gözükan and J. Renault, *Cartolabe-visu* version v3.0, 2023. LIC: BSD 3-Clause. HAL: ⟨hal-04417569⟩, URL: https://hal.science/hal-04417569, VCS: https://gitlab.inria.fr/cartolabe/cartolabe-visu, SWHID: ⟨swh:1:dir:b6e26b967f0056d9933baad4cb94136678123e50;origin=https://gitlab.inria.fr/cartolabe/cartolabe-visu.git;visit=swh:1:snp:22a1dc08954a7534bdb1cb6efe90aceb742cb409;anchor=swh:1:rev:1578dbe49772534ef5681a72fbd2a1973eabaef4⟩.

12.4 Cited publications

[93] P. Caillou, J. Renault, J.-D. Fekete, A.-C. Letournel and M. Sebag. 'Cartolabe: A Web-Based Scalable Visualization of Large Document Collections'. In: *IEEE Computer Graphics and Applications* 41.2 (Apr. 2021), pp. 76–88. DOI: 10.1109/MCG.2020.3033401. URL: https://hal.inria.fr/hal-02499006.

[94] A. Decelle and C. Furtlehner. 'Exact Training of Restricted Boltzmann Machines on Intrinsically Low Dimensional Data'. In: *Physical Review Letters* (6th Sept. 2021). URL: https://hal.inria.fr/hal-03432350.

[95] B. Yelmen, A. Decelle, L. Ongaro, D. Marnetto, C. Tallec, F. Montinaro, C. Furtlehner, L. Pagani and F. Jay. 'Creating artificial human genomes using generative neural networks'. In: *PLoS Genetics* (4th Feb. 2021). DOI: 10.1371/journal.pgen.1009303. URL: https://hal.archives-ouvertes.fr/hal-03149930.

[96] R. Bresson, J. Cohen, E. Hüllermeier, C. Labreuche and M. Sebag. 'Neural Representation and Learning of Hierarchical 2-additive Choquet Integrals'. In: IJCAI-PRICAI-20 - Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence. Yokohama, France, 11th July 2020, pp. 1984–1991. DOI: 10.24963/ijcai.2020/275. URL: https://hal.archives-ouvertes.fr/hal-03017078.

[97] R. Bresson, J. Cohen, E. Hüllermeier, C. Labreuche and M. Sebag. 'On the Identifiability of Hierarchical Decision Models'. In: 18th International Conference on Principles of Knowledge Representation and Reasoning (KR-2021). Online, France: International Joint Conferences on Artificial Intelligence Organization, 12th Nov. 2021, pp. 151–162. DOI: 10.24963/kr.2021/15. URL: https://hal.archives-ouvertes.fr/hal-03453996.

[98] A. Decelle, C. Furtlehner and B. Seoane. 'Equilibrium and non-Equilibrium regimes in the learning of Restricted Boltzmann Machines'. In: NeurIPS 2021. Proceedings NeurIPS 2021. Vancouver, United States, 6th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03518796.

[99] K. Gasnikova, P. Caillou, O. Allais and M. Sebag. 'Towards causal modeling of nutritional outcomes'. In: Causal Analysis Workshop Series (CAWS) 2021. Vol. 5. 19. online, United States, 2021. URL: https://hal.inria.fr/hal-03620867.

[100] N. Sanchez-Pi, L. Marti, J. Salomon, J. Sainte-Marie, O. Bernard, M. Sebag, M. Schoenauer, A. Maass, D. Eveillard, A. Abreu, C. de Vargas and P. A. Marquet. *OcéanIA: AI, Data, and Models for Understanding the Ocean and Climate Change.* 1st July 2021, pp. 1–64. URL: https://hal.archives-ouvertes.fr/hal-03274323.

[101] V. Berger. 'Deep Latent Variable Models : from properties to structures'. Université Paris-Saclay, 13th Oct. 2021. URL: https://tel.archives-ouvertes.fr/tel-03528577.

[102] G. De Bie, H. Rakotoarison, G. Peyré and M. Sebag. *Distribution-Based Invariant Deep Networks for Learning Meta-Features.* 26th Feb. 2021. URL: https://hal.inria.fr/hal-03153200.

[103] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kégl and D. Rousseau. 'How Machine Learning won the Higgs Boson Challenge'. In: *Proc. European Symposium on ANN, CI and ML.* 2016.

[104] C. Adam-Bourdarios, G. Cowan, C. Germain-Renaud, I. Guyon, B. Kégl and D. Rousseau. 'The Higgs Machine Learning Challenge'. In: *Journal of Physics: Conference Series* 664.7 (2015). DOI: 10.1088/1742-6596/664/7/072015. URL: https://hal.inria.fr/hal-01745998.

[105] J. K. et al. *Scaling Laws for Neural Language Models.* arXiv:2001.08361. 2020.

[106] F. Arceri, F. P. Landes, L. Berthier and G. Biroli. 'Glasses and aging: A Statistical Mechanics Perspective'. In: *Encyclopedia of Complexity and Systems Science (Living Reference).* 50 pages, 24 figs. This is an updated version of a chapter initially written in 2009 for the Encyclopedia of Complexity and Systems Science (Springer). Mar. 2022. DOI: 10.1007/978-3-642-27737-5_248-2. URL: https://inria.hal.science/hal-02942375.

[107] M. M. Bronstein, J. Bruna, T. Cohen and P. Veličković. 'Geometric deep learning: Grids, groups, graphs, geodesics, and gauges'. In: *arXiv preprint arXiv:2104.13478* (2021).

[108] L. Chizat, E. Oyallon and F. Bach. 'On Lazy Training in Differentiable Programming'. In: *NeurIPS.* Ed. by H. Wallach et al. Vol. 32. 2019.

[109] T. Cohen and M. Welling. 'Group Equivariant Convolutional Networks'. In: *Proc. ICML.* Ed. by M. F. Balcan and K. Q. Weinberger. Vol. 48. PMLR. 2016, pp. 2990–2999.

[110] David Rousseau et al. 'The TrackML challenge'. In: *NIPS 2018 - 32nd Annual Conference on Neural Information Processing Systems.* Montreal, Canada, Dec. 2018, pp. 1–23. URL: https://hal.inria.fr/hal-01745714.

[111] A. Decelle, G. Fissore and C. Furtlehner. 'Spectral dynamics of learning in restricted Boltzmann machines'. In: *EPL (Europhysics Letters)* 119.6 (2017), p. 60001.

[112] A. Decelle, G. Fissore and C. Furtlehner. 'Thermodynamics of Restricted Boltzmann Machines and Related Learning Dynamics'. In: *J. Stat. Phys.* 172 (2018), pp. 1576–1608.

[113] L. Dinh, J. Sohl-Dickstein and S. Bengio. 'Density estimation using Real NVP'. In: *Int. Conf. on Learning Representations (ICLR).* 2017.

[114] S. Dong and M. Sebag. 'From graphs to DAGs: a low-complexity model and a scalable algorithm'. In: *ECML-PKDD 2022 - European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.* Grenoble, France, Sept. 2022. URL: https://inria.hal.science/hal-03798964.

[115] B. Donnot. 'Deep learning methods for predicting flows in power grids : novel architectures and algorithms'. Theses. Université Paris Saclay (COmUE), Feb. 2019. URL: https://tel.archives-ouvertes.fr/tel-02045873.

[116] D. Donoho. *Data Science at the Singularity.* 2023. arXiv: 2310.00865 [stat.OT].

[117] B. Donon. 'Deep statistical solvers & power systems applications'. Theses. Université Paris-Saclay, Mar. 2022. URL: https://theses.hal.science/tel-03624628.

[118] B. Donon, W. Liu, A. Marot, Z. Liu, I. Guyon and M. Schoenauer. 'Deep Statistical Solvers'. In: *NeurIPS 2020 - 34th Conference on Neural Information Processing Systems.* Vancouver / Virtuel, Canada, Dec. 2020. URL: https://hal.inria.fr/hal-02974541.

[119] R. Egele, R. Maulik, K. Raghavan, B. Lusch, I. Guyon and P. Balaprakash. 'AutoDEUQ: Automated Deep Ensemble with Uncertainty Quantification'. In: *26TH International Conference on Pattern Recognition.* Montréal, Canada: IEEE, Aug. 2022, pp. 1908–1914. URL: https://hal.science/hal-03518597.

[120] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos and H. Escalante. 'Multi-modal gesture recognition challenge 2013: dataset and results'. In: *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*. ICMI '13. Sydney, Australia: Association for Computing Machinery, 2013, pp. 445–452. DOI: 10.1145/2522848.2532595. URL: https://doi.org/10.1145/2522848.2532595.

[121] J. Frankle and M. Carbin. 'The lottery ticket hypothesis: Finding sparse, trainable neural networks'. In: *arXiv preprint arXiv:1803.03635* (2018).

[122] J. Giles. 'Computational social science: Making the links'. In: *Nature - News* 488.7412 (2012), pp. 448–450.

[123] A. Hyvarinen, H. Sasaki and R. Turner. 'Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning'. In: *Proc. AISTAT*. Ed. by K. Chaudhuri and M. Sugiyama. Vol. 89. PMLR. 2019, pp. 859–868. URL: https://proceedings.mlr.press/v89/hyvarinen19a.html.

[124] A. Jacot, F. Gabriel and C. Hongler. 'Neural Tangent Kernel: Convergence and Generalization in Neural Networks'. In: *NeurIPS*. Ed. by S. Bengio et al. Vol. 31. 2018.

[125] A. Karimi, G. Barthe, B. Schölkopf and I. Valera. 'A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations'. In: *ACM Comput. Surv.* 55.5 (2023), 95:1–95:29. DOI: 10.1145/3527848. URL: https://doi.org/10.1145/3527848.

[126] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, A. Pritzel, S. Ravuri, T. Ewalds, F. Alet, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, J. Stott, O. Vinyals, S. Mohamed and P. Battaglia. *GraphCast: Learning skillful medium-range global weather forecasting*. cite arxiv:2212.12794. 2022. URL: http://arxiv.org/abs/2212.12794.

[127] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy and M. Van Alstyne. 'Life in the network: the coming age of computational social science'. In: *Science* 323.5915 (2009), pp. 721–723.

[128] Y. Lecun, S. Chopra, R. Hadsell, M. Ranzato and F. Huang. 'A tutorial on energy-based learning'. In: *Predicting structured data*. Ed. by G. Bakir, T. Hofman, B. Scholkopt, A. Smola and B. Taskar. MIT Press, 2006.

[129] M. Leyli-Abadi, D. Danan, M. Yagoubi, S.-E. Attoui, A. Marot, J. Picault, B. Donnot, P. Dimitrov, C. Etienam and A. Farjallah. 'LIPS - Learning Industrial Physical Simulation benchmark suite'. In: *NeurIPS - Data & Benchmark Track*. New Orleans, United States, Nov. 2022. URL: https://hal.science/hal-03780816.

[130] W. Liu, M. Yagoubi, D. Danan and M. Schoenauer. 'Multi-Fidelity Transfer Learning for accurate data-based PDE approximation'. In: *NeurIPS 2022 - Workshop on Machine Learning and the Physical Sciences*. New Orleans, United States, Nov. 2022. URL: https://hal.science/hal-03878200.

[131] W. Liu, M. Yagoubi and M. Schoenauer. 'Multi-resolution Graph Neural Networks for PDE Approximation'. In: *ICANN 2021 - 30th International Conference on Artificial Neural Networks*. Ed. by I. Farkaš, P. Masulli, S. Otte and S. Wermter. Vol. LNCS-12893. Artificial Neural Networks and Machine Learning – ICANN 2021. Bratislava, Slovakia: Springer International Publishing, Sept. 2021, pp. 151–163. DOI: 10.1007/978-3-030-86365-4_13. URL: https://hal.science/hal-03448278.

[132] Z. Liu and M. Tegmark. 'Machine Learning Hidden Symmetries'. In: *Phys. Rev. Lett.* 128 (18 2022), p. 180201.

[133] C. Louizos, K. Ullrich and M. Welling. 'Bayesian compression for deep learning'. In: *Advances in neural information processing systems* 30 (2017).

[134] A. Marot, B. Donon, I. Guyon and B. Donnot. 'Learning To Run a Power Network Competition'. In: *CiML Workshop, NeurIPS*. Montréal, Canada, Dec. 2018. URL: https://hal.archives-ouvertes.fr/hal-01968295.

[135] E. Menier, M. A. Bucci, M. Yagoubi, L. Mathelin, T. Dairay, R. Meunier and M. Schoenauer. 'Continuous Methods : Adaptively intrusive reduced order model closure'. In: *ICML 2022 - Workshop Continuous time methods for machine learning*. Baltimore, United States, July 2022. URL: https://inria.hal.science/hal-03879332.

[136] M. Nastorg, M. Schoenauer, G. Charpiat, T. Faney, J.-M. Gratien and M. A. Bucci. 'DS-GPS : A Deep Statistical Graph Poisson Solver'. In: *NeurIPS 2022 - Machine Learning and the Physical Sciences, workshop*. New-Orleans, United States, Dec. 2022. URL: https://inria.hal.science/hal-03864015.

[137] A. Pavao, I. Guyon, A.-C. Letournel, X. Baró, H. Escalante, S. Escalera, T. Thomas and Z. Xu. *CodaLab Competitions: An open source platform to organize scientific challenges*. Technical Report. Université Paris-Saclay, FRA., Apr. 2022. URL: https://inria.hal.science/hal-03629462.

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

[139] F. S. Pezzicoli, G. Charpiat and F. P. Landes. 'SE (3)-equivariant graph neural networks for learning glassy liquids representations'. In: *arXiv preprint arXiv:2211.03226* (2022).

[140] H. Rakotoarison. 'Some contributions to AutoML : hyper-parameter optimization and meta-learning'. Theses. Université Paris-Saclay, June 2022. URL: https://theses.hal.science/tel-03783610.

[141] H. Rakotoarison, L. Milijaona, A. Rasoanaivo, M. Sebag and M. Schoenauer. 'Learning Meta-features for AutoML'. In: *ICLR 2022 - International Conference on Learning Representations (spotlight)*. Virtual, United States, Apr. 2022. URL: https://inria.hal.science/hal-03583789.

[142] T. Sanchez. 'Reconstructing our past : deep learning for population genetics'. Theses. Université Paris-Saclay, Mar. 2022. URL: https://theses.hal.science/tel-03701132.

[143] G. Serré, E. Boguslawski, B. Donnot, A. Pavão, I. Guyon and A. Marot. 'Reinforcement learning for Energies of the future and carbon neutrality: a Challenge Design'. In: *SSCI 2022 - IEEE Symposium Series on Computational Intelligence*. IEEE. Singapour, Singapore, Dec. 2022. URL: https://hal.science/hal-03726294.

[144] C. Shi, D. M. Blei and V. Veitch. 'Adapting Neural Networks for the Estimation of Treatment Effects'. In: *NeurIPS 2019*. 2019, pp. 2503–2513. URL: https://proceedings.neurips.cc/paper/2019/hash/8fb5f8be2aa9d6c64a04e3ab9f63feee-Abstract.html.

[145] I. Stengers. *Another Science Is Possible*. Open Humanities Press, 2013.

[146] Y. Wang and D. Blei. 'A Proxy Variable View of Shared Confounding'. In: *Proc. 38th ICML*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 10697–10707. URL: https://proceedings.mlr.press/v139/wang21c.html.

[147] Z. Xu, S. Escalera, A. Pavao, M. Richard, W.-W. Tu, Q. Yao, H. Zhao and I. Guyon. 'Codabench: Flexible, Easy-to-Use and Reproducible Meta-Benchmark Platform'. In: *Patterns* (July 2022). URL: https://hal.science/hal-03374222.

[148] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao and K. P. Bennett. 'Generation and Evaluation of Privacy Preserving Synthetic Health Data'. In: *Neurocomputing* 416 (Nov. 2020), pp. 244–255. DOI: 10.1016/j.neucom.2019.12.136. URL: https://hal.inria.fr/hal-03158544.

[149] X. Zheng, B. Aragam, P. K. Ravikumar and E. P. Xing. 'DAGs with NO TEARS: Continuous Optimization for Structure Learning'. In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018. URL: https://proceedings.neurips.cc/paper/2018/file/%20e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf.