RESEARCH CENTRE

Inria Saclay Center at Université Paris-Saclay

IN PARTNERSHIP WITH:

CNRS, Université Paris-Saclay

2022 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 1

1 Team members, visitors, external collaborators 2

2 Overall objectives 3
- 2.1 Presentation 3

3 Research program 4
- 3.1 Toward Good AI 4
  - 3.1.1 Causal modeling and biases 4
  - 3.1.2 Robustness of Learned Models 4
- 3.2 Hybridizing numerical modeling and learning systems 5
- 3.3 Learning to learn 6
  - 3.3.1 Auto-* 7
  - 3.3.2 Information theory: adjusting model complexity and data fitting 8
  - 3.3.3 Analyzing and Learning Complex Systems 8
- 3.4 Organisation of Challenges 9

4 Application domains 9
- 4.1 Computational Social Sciences 9
- 4.2 Energy Management 10
- 4.3 Data-driven Numerical Modeling 10

5 Social and environmental responsibility 11
- 5.1 Footprint of research activities 11
- 5.2 Impact of research results 11

6 Highlights of the year 11
- 6.1 Awards 11

7 New software and platforms 12
- 7.1 New software 12
  - 7.1.1 Codalab 12
  - 7.1.2 Cartolabe 12
- 7.2 New platforms 13

8 New results 14
- 8.1 Toward Good AI 14
  - 8.1.1 Causal Modeling 14
  - 8.1.2 Explainability 14
  - 8.1.3 Robustness of AI Systems 15
- 8.2 Learning to Learn 15
  - 8.2.1 Auto-* 15
  - 8.2.2 Deep Learning: Practical and Theoretical Insights 16
  - 8.2.3 Analyzing and Learning Complex Systems 17
- 8.3 Computational Social Sciences 18
  - 8.3.1 Labor Studies 18
  - 8.3.2 Health and practices 19
  - 8.3.3 Scientific Information System and Visual Querying 19
- 8.4 Energy Management 20
  - 8.4.1 Power Grids Management 20
  - 8.4.2 Optimization of Local Grids 20
  - 8.4.3 Accelerating simulation codes 20
- 8.5 Data-driven Numerical Modelling 21
  - 8.5.1 Space Weather Forecasting 21

8.5.2 Genomic Data and Population Genetics . . . 21
8.5.3 Privacy and synthetic data generation . . . 22
8.5.4 Sampling molecular conformations . . . 23
8.5.5 Earthquake occurrence prediction . . . 23
8.5.6 Reduced order model correction . . . 24
8.5.7 Active Learning for chaotic systems . . . 24
8.5.8 Graph Neural Networks for Numerical Simulations . . . 25
8.6 Challenges . . . 25

9 Bilateral contracts and grants with industry **27**
9.1 Bilateral contracts with industry . . . 27

10 Partnerships and cooperations **27**
10.1 International initiatives . . . 27
10.1.1 STIC/MATH/CLIMAT AmSud projects . . . 27
10.1.2 Participation in other International Programs . . . 28
10.2 European initiatives . . . 28
10.2.1 Horizon Europe . . . 28
10.2.2 H2020 projects . . . 29
10.3 National initiatives . . . 30
10.3.1 ANR . . . 30
10.3.2 Others . . . 31
10.4 Regional initiatives . . . 31

11 Dissemination **31**
11.1 Promoting scientific activities . . . 31
11.1.1 Scientific events: organisation . . . 31
11.1.2 Scientific events: selection . . . 32
11.1.3 Journal . . . 32
11.1.4 Invited talks . . . 32
11.1.5 Leadership within the scientific community . . . 33
11.1.6 Scientific expertise . . . 33
11.1.7 Research administration . . . 33
11.2 Teaching - Supervision - Juries . . . 34
11.2.1 Teaching . . . 34
11.2.2 Supervision . . . 34
11.2.3 Juries . . . 36
11.3 Popularization . . . 36
11.3.1 Internal or external Inria responsibilities . . . 36
11.3.2 Interventions . . . 36

12 Scientific production **36**
12.1 Major publications . . . 36
12.2 Publications of the year . . . 37
12.3 Cited publications . . . 41

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]
- Cécile Germain [UNIV PARIS SACLAY, Emeritus, HDR]
- Flora Jay [CNRS, Researcher]
- Michèle Sebag [CNRS, Senior Researcher, HDR]
- Paola Tubaro [CNRS, Senior Researcher, until Sep 2022, HDR]

Faculty Members

- Philippe Caillou [UNIV PARIS SACLAY, Associate Professor]
- Sylvain Chevallier [UNIV. PARIS-SACLAY, Professor, from Sep 2022]
- Aurelien Decelle [UNIV PARIS SACLAY, Associate Professor, En disponiblité à Complutense Madrid]
- Isabelle Guyon [U. Paris-Saclay and INRIA, Professor, until Sep 2022, Inria Chair]
- François Landes [UNIV PARIS SACLAY, Associate Professor]

Post-Doctoral Fellows

- Jean Cury [UNIV PARIS SACLAY, until Feb 2022]
- Shuyu Dong [INRIA]
- Tamon Nakano [INRIA, until Jun 2022]
- Shiyang Yan [INRIA]

PhD Students

- Léonard Blier [Facebook (now Meta), CIFRE, until Apr 2022]
- Eva Boguslawski [RTE, CIFRE, from May 2022]
- Roman Bresson [Thalès, CIFRE, until Feb 2022]
- Balthazar Donon [RTE, CIFRE, until Mar 2022]
- Romain Egele [Argonne Lab., from Oct 2022]
- Loris Felardos Saint Jean [INRIA, until Nov 2022]
- Giancarlo Fissore [U. Paris-Saclay, until Mar 2022]
- Isabelle Hoxha [UNIV. PARIS-SACLAY]
- Badr Youbi Idrissi [Meta, CIFRE, from Sep 2022]
- Armand Lacombe [INRIA]

- Wenzhuo Liu [IRT SYSTEM X]
- Romain Lloria [UNIV. PARIS-SACLAY, from Nov 2022]
- Emmanuel Menier [IRT SYSTEM X]
- Thibault Monsel [UNIV PARIS SACLAY, from May 2022]
- Matthieu Nastorg [INRIA]
- Francesco Pezzicoli [UNIV PARIS SACLAY]
- Audrey Poinsot [EKIMETRICS, CIFRE, from Mar 2022]
- Arnaud Quelin [SORBONNE UNIVERSITE, from Oct 2022]
- Herilalaina Rakotoarison [INRIA, until Jun 2022]
- Cyriaque Rousselot [UNIV PARIS SACLAY, from Oct 2022]
- Théophile Sanchez [UNIV PARIS SACLAY, until Jan 2022]
- Vincenzo Schimmenti [CNRS]
- Nilo Schwencke [UNIV PARIS SACLAY]
- Haozhe Sun [UNIV PARIS SACLAY]
- Antoine Szatkownik [UNIV PARIS SACLAY, from Oct 2022]
- Marion Ullmo [CNRS, until Feb 2022]
- Manon Verbockhaven [UNIV PARIS SACLAY]
- Assia Wirth [UNIV PARIS SACLAY]
- Maria Sayu Yamamoto [UNIV. PARIS-SACLAY]

Technical Staff

- Adrien Pavao [UNIV PARIS SACLAY, Engineer]
- Sebastien Treguer [INRIA, Engineer, from Apr 2022]

2 Overall objectives

2.1 Presentation

Since its creation in 2003, TAO activities had constantly but slowly evolved, as old problems were being solved, and new applications arose, bringing new fundamental issues to tackle. But recent abrupt progresses in Machine Learning (and in particular in Deep Learning) have greatly accelerated these changes also within the team. It so happened that this change of slope also coincided with some more practical changes in TAO ecosystem: following Inria 12-years rule, the team definitely ended in December 2016. The new team TAU (for **TA**ckling the **U**nderspecified) has been proposed, and formally created in July 2019. At the same time important staff changes took place, that also justify even sharper changes in the team focus. During the year 2018, the second year of this new era for the (remaining) members of the team, our research topics have now stabilized around a final version of the TAU project.

Following the dramatic changes in TAU staff during the years 2016-2017 (see the 2017 activity report of the team for the details), the research around continuous optimization has definitely faded out in TAU (while the research axis on hyperparameter tuning has focused on Machine Learning algorithms), the Energy application domain has slightly changed direction under Isabelle Guyon's supervision (Section 4.2), after the completion of the work started by Olivier Teytaud, and a few new directions have emerged, around the robustness of ML systems (Section 3.1.2). The other research topics have been continued, as described below.

3 Research program

3.1 Toward Good AI

As discussed by [151], and in the recent collaborative survey paper [73], the topic of ethical AI was non-existent until 2010, was laughed at in 2016, and became a hot topic in 2017 as the AI disruptivity with respect to the fabric of life (travel, education, entertainment, social networks, politics, to name a few) became unavoidable [146], together with its expected impacts on the nature and amount of jobs. As of now, it seems that the risk of a new AI Winter might arise from legal[1] and societal[2] issues. While privacy is now recognized as a civil right in Europe, it is feared that the GAFAM, BATX and others can already capture a sufficient fraction of human preferences and their dynamics to achieve their commercial and other goals, and build a Brave New Big Brother (BNBB), a system that is openly beneficial to many, covertly nudging, and possibly dictatorial).

The ambition of TAU is to mitigate the BNBB risk along several intricated dimensions, and build i) causal and explainable models; ii) fair data and models; iii) provably robust models.

3.1.1 Causal modeling and biases

Participants: Isabelle Guyon, Michèle Sebag, Philippe Caillou, Paola Tubaro

The extraction of causal models, a long goal of AI [149, 126, 150], became a strategic issue as the usage of learned models gradually shifted from *prediction* to *prescription* in the last years. This evolution, following Auguste Comte's vision of science (*Savoir pour prévoir, afin de pouvoir*) indeed reflects the exuberant optimism about AI: Knowledge enables Prediction; Prediction enables Control. However, although predictive models can be based on correlations, prescriptions can only be based on causal models[3].

Among the research applications concerned with causal modeling, predictive modeling or collaborative filtering at TAU are all projects described in section 4.1 (see also Section 3.4), studying the relationships between: i) the educational background of persons and the job openings (FUI project JobAgile and DataIA project Vadore); ii) the quality of life at work and the economic performance indicators of the enterprises (ISN Lidex project Amiqap) [128] ; iii) the nutritional items bought by households (at the level of granularity of the barcode) and their health status, as approximated from their body-mass-index (IRS UPSaclay Nutriperso); iv) the actual offer of restaurants and their scores on online rating systems. In these projects, a wealth of data is available (though hardly sufficient for applications ii), iii and iv))) and there is little doubt that these data reflect the imbalances and biases of the world as is, ranging from gender to racial to economical prejudices. Preventing the learned models from perpetuating such biases is essential to deliver an AI endowed with common decency.

In some cases, the bias is known; for instance, the cohorts in the Nutriperso study are more well-off than the average French population, and the Kantar database includes explicit weights to address this bias through importance sampling. In other cases, the bias is only guessed; for instance, the companies for which Secafi data are available hardly correspond to a uniform sample as these data have been gathered upon the request of the company trade union.

Causal relationships are being identified using our recently published paper [18]. This work will be continued, as TAU is a partner of the PEPR-IA project Causalit-AI (local PI Michèle Sebag), starting next Spring.

3.1.2 Robustness of Learned Models

Participants: Guillaume Charpiat, Marc Schoenauer, Michèle Sebag

[1] For instance, the (fictitious) plea challenge proposed to law students in Oct. 2018 considered a chain reaction pileup occurred among autonomous and humanly operated vehicles on a highway.

[2] For instance related to information bubbles and nudge [116, 161].

[3] One can predict that it rains based on the presence of umbrellas in the street; but one cannot induce rainfall by going out with an umbrella. Likewise, the presence of books/tablets at home and the good scores of children at school are correlated; but offering books/tablets to all children might fail to improve their scores *per se*, if both good scores and books are explained by a so-called confounder variable, like the presence of adults versed in books/tablets at home.

Due to their outstanding performances, deep neural networks and more generally machine learning-based decision making systems, referred to as MLs in the following, have been raising hopes in the recent years to achieve breakthroughs in critical systems, ranging from autonomous vehicles to defense. The main pitfall for such applications lies in the lack of guarantees for MLs robustness.

Specifically, MLs are used when the mainstream software design process does not apply, that is, when no formal specification of the target software behavior is available and/or when the system is embedded in an open unpredictable world. The extensive body of knowledge developed to deliver guarantees about mainstream software – ranging from formal verification, model checking and abstract interpretation to testing, simulation and monitoring – thus does not directly apply either. Another weakness of MLs regards their dependency to the amount and quality of the training data, as their performances are sensitive to slight perturbations of the data distribution. Such perturbations can occur naturally due to domain or concept drift (e.g. due to a change in light intensity or a scratch on a camera lens); they can also result from intentional malicious attacks, a.k.a adversarial examples [162].

These downsides, currently preventing the dissemination of MLs in safety-critical systems (SCS), call for a considerable amount of research, in order to understand when and to which extent an MLs can be certified to provide the desired level of guarantees.

This activity has been put on hold in the team this year, but will be revived in the PEPR-IA SAIF project (starting next Spring) in which TAU is a partner (local PI Guillaume Charpiat).

3.2 Hybridizing numerical modeling and learning systems

Participants: Guillaume Charpiat, Cécile Germain, Isabelle Guyon, Marc Schoenauer, Michèle Sebag

In sciences and engineering, human knowledge is commonly expressed in closed form, through equations or mechanistic models characterizing how a natural or social phenomenon, or a physical device, will behave/evolve depending on its environment and external stimuli, under some assumptions and up to some approximations. The field of numerical engineering, and the simulators based on such mechanistic models, are at the core of most approaches to understand and analyze the world, from solid mechanics to computational fluid dynamics, from chemistry to molecular biology, from astronomy to population dynamics, from epidemiology and information propagation in social networks to economy and finance.

Most generally, numerical engineering supports the simulation, and when appropriate the optimization and control[4] of the phenomenons under study, although several sources of discrepancy might adversely affect the results, ranging from the underlying assumptions and simplifying hypotheses in the models, to systematic experiment errors to statistical measurement errors (not to mention numerical issues). This knowledge and know-how are materialized in millions of lines of code, capitalizing the expertise of academic and industrial labs. These softwares have been steadily extended over decades, modeling new and more fine-grained effects through layered extensions, making them increasingly harder to maintain, extend and master. Another difficulty is that complex systems most often resort to hybrid (pluridisciplinary) models, as they involve many components interacting along several time and space scales, hampering their numerical simulation.

At the other extreme, machine learning offers the opportunity to model phenomenons from scratch, using any available data gathered through experiments or simulations. Recent successes of machine learning in computer vision, natural language processing and games, to name a few, have demonstrated the power of such agnostic approaches and their efficiency in terms of prediction [132], inverse problem solving [147], and sequential decision making [164, 102], despite their lack of any "semantic" understanding of the universe. Even before these successes, Anderson's claim was that *the data deluge [might make] the scientific method obsolete* [91], as if a reasonable option might be to throw away the existing equational or software bodies of knowledge, and let Machine Learning rediscover all models from scratch. Such a claim is hampered among others by the fact that not all domains offer a wealth of data, as any academic involved in an industrial collaboration around data has discovered.

Another approach is considered in TAU, investigating how existing mechanistic models and related simulators can be partnered with ML algorithms: i) to achieve the same goals with the same methods

[4] Note that the causal nature of mechanistic models is established from prior knowledge and experimentations.

with a gain of accuracy or time; ii) to achieve new goals; iii) to achieve the same goals with new methods.

Toward more robust numerical engineering: In domains where satisfying mechanistic models and simulators are available, ML can contribute to improve their accuracy or usability. A first direction is to refine or extend the models and simulators to better fit the empirical evidence. The goal is to finely account for the different biases and uncertainties attached to the available knowledge and data, distinguishing the different types of *known unknowns*. Such *known unknowns* include the model hyper-parameters (coefficients), the systematic errors due to e.g., experiment imperfections, and the statistical errors due to e.g., measurement errors. A second approach is based on learning a surrogate model for the phenomenon under study that incorporate domain knowledge from the mechanistic model (or its simulation). See Section 8.5 for case studies.

A related direction, typically when considering black-box simulators, aims to learn a model of the error, or equivalently, a post-processor of the software. The discrepancy between simulated and empirical results, referred to as *reality gap* [137], can be tackled in terms of domain adaptation [95, 115]. Specifically, the source domain here corresponds to the simulated phenomenon, offering a wealth of inexpensive data, and the target domain corresponds to the actual phenomenon, with rare and expensive data; the goal is to devise accurate target models using the source data and models.

Extending numerical engineering: ML, using both experimental and numerical data, can also be used to tackle new goals, that are beyond the current state-of-the-art of standard approaches. Inverse problems are such goals, identifying the parameters or the initial conditions of phenomenons for which the model is not differentiable, or amenable to the adjoint state method.

A slightly different kind of inverse problem is that of recovering the ground truth when only noisy data is available. This problem can be formulated as a search for the simplest model explaining the data. The question then becomes to formulate and efficiently exploit such a simplicity criterion.

Another goal can be to model the distribution of given quantiles for some system: The challenge is to exploit available data to train a generative model, aimed at sampling the target quantiles.

Examples tackled in TAU are detailed in Section 8.5. Note that the "Cracking the Glass Problem", described in Section 8.2.3 is yet another instance of a similar problem.

Data-driven numerical engineering: Finally, ML can also be used to sidestep numerical engineering limitations in terms of scalability, or to build a simulator emulating the resolution of the (unknown) mechanistic model from data, or to revisit the formal background.

When the mechanistic model is known and sufficiently accurate, it can be used to train a deep network on an arbitrary set of (space,time) samples, resulting in a meshless numerical approximation of the model [158], supporting by construction *differentiable programming* [134].

When no mechanistic model is sufficiently efficient, the model must be identified from the data only. Genetic programming has been used to identify systems of ODEs [157], through the identification of invariant quantities from data, as well as for the direct identification of control commands of nonlinear complex systems, including some chaotic systems [109]. Another recent approach uses two deep neural networks, one for the state of the system, the other for the equation itself [152]. The critical issues for both approaches include the scalability, and the explainability of the resulting models. Such line of research will benefit from TAU unique mixed expertise in Genetic Programming and Deep Learning.

3.3 Learning to learn

According to Ali Rahimi's test of times award speech at NIPS 17, the current ML algorithms *have become a form of alchemy*. Competitive testing and empirical breakthroughs gradually become mandatory for a contribution to be acknowledged; an increasing part of the community adopts trials and errors as main scientific methodology, and theory is lagging behind practice. This style of progress is typical of technological and engineering revolutions for some; others ask for consolidated and well-understood theoretical advances, saving the time wasted in trying to build upon hardly reproducible results.

Basically, while practical achievements have often passed the expectations, there exist caveats along three dimensions. Firstly, excellent performances do not imply that the model has captured what was to learn, as shown by the phenomenon of adversarial examples. Following Ian Goodfellow, some well-performing models might be compared to *Clever Hans*, the horse that was able to solve mathematical exercizes using non verbal cues from its teacher [125]; it is the purpose of Pillar I. to alleviate the *Clever Hans* trap (section 3.1).
Secondly, some major advances, e.g. related to the celebrated adversarial learning [119, 115], establish proofs of concept more than a sound methodology, where the reproducibility is limited due to i) the computational power required for training (often beyond reach of academic labs); ii) the numerical instabilities (witnessed as random seeds happen to be found in the codes); iii) the insufficiently documented experimental settings. What works, why and when is still a matter of speculation, although better understanding the limitations of the current state of the art is acknowledged to be a priority. After Ali Rahimi again, *simple experiments, simple theorems are the building blocks that help us understand more complicated systems.* Along this line, [143] propose toy examples to demonstrate and understand the defaults of convergence of gradient descent adversarial learning.
Thirdly, and most importantly, the reported achievements rely on carefully tuned learning architectures and hyper-parameters. The sensitivity of the results to the selection and calibration of algorithms has been identified since the end 80s as a key ML bottleneck, and the field of automatic algorithm selection and calibration, referred to as AutoML or Auto-⋆ in the following, is at the ML forefront.

TAU aims to contribute to the ML evolution toward a more mature stage along three dimensions. In the short term, the research done in Auto-⋆ will be pursued (section 3.3.1). In the medium term, an information theoretic perspective will be adopted to capture the data structure and to calibrate the learning algorithm *depending on the nature and amount of the available data* (section 3.3.2). In the longer term, our goal is to leverage the methodologies forged in statistical physics to understand and control the trajectories of complex learning systems (section 3.3.3).

3.3.1 Auto-*

Participants: Isabelle Guyon, Marc Schoenauer, Michèle Sebag

The so-called Auto-⋆ task, concerned with selecting a (quasi) optimal algorithm and its hyper-parameters depending on the problem instance at hand, remained a key issue in ML for the last three decades [96], as well as in optimization at large [124], including combinatorial optimization and constraint satisfaction [131] and continuous optimization [93]. This issue, tackled by several European projects along the decades, governs the knowledge transfer to industry, due to the shortage of data scientists. It becomes even more crucial as models are more complex and their training requires more computational resources. This has motivated several international challenges devoted to Auto-ML [123] (see also Section 3.4), including the AutoDL challenge series [139] (see also Section 8.6). The latest one AutoGraph aims to bring Automated Machine Learning to Graph Learning, aiming to reduce the human effort and achieve generally top-performing Graph Neural Networks (GNN) [67]. The AutoML challenge series results pointed to the importance of meta-learning, which conducted us to pursue a new line of research on meta-learning from learning curves [28, 64] and cross-domain meta-learning [36]. The dataset that was developed for ths work was published in the NeurIPS datasets and benchmarks track [39]. This line of work led us to explore uses of reinforcement learning as a means to devise policies for meta-learning.

Several approaches have been used to tackle Auto-⋆ in the literature, and TAU has been particularly active in several of them. Meta-learning aims to build a surrogate performance model, estimating the performance of an algorithm configuration on *any* problem instance characterized from its meta-feature values [155, 93, 118]. Collaborative filtering, considering that a problem instance "likes better" an algorithm configuration yielding a better performance, learns to recommend good algorithms to problem instances [160, 144]. Bayesian optimization proceeds by alternatively building a surrogate model of algorithm performances on *the* problem instance at hand, and tackling it [111]. This last approach currently is the prominent one; as shown in [144], the meta-features developed for AutoML are hardly relevant, hampering both meta-learning and collaborative filtering. The design of better features is another long-term research direction, in which TAU has recently been [108], ans still is very active. more recent approach used in TAU [153] extends the Bayesian Optimization approach with a Multi-Armed

Bandit algorithm to generate the full Machine Learning pipeline, competing with the famed AutoSKLearn [111] (see Section 8.2.1).

3.3.2 Information theory: adjusting model complexity and data fitting

Participants: Guillaume Charpiat, Marc Schoenauer, Michèle Sebag

In the 60s, Kolmogorov and Solomonoff provided a well-grounded theory for building (probabilistic) models best explaining the available data [156, 120], that is, the shortest programs able to generate these data. Such programs can then be used to generate further data or to answer specific questions (interpreted as missing values in the data). Deep learning, from this viewpoint, efficiently explores a space of computation graphs, described from its hyperparameters (network structure) and parameters (weights). Network training amounts to optimizing these parameters, namely, navigating the space of computational graphs to find a network, as simple as possible, that explain the past observations well.

This vision is at the core of variational auto-encoders [130], directly optimizing a bound on the Kolmogorov complexity of the dataset. More generally variational methods provide quantitative criteria to identify superfluous elements (edges, units) in a neural network, that can potentially be used for structural optimization of the network (Leonard Blier's PhD, started Oct. 2018).

The same principles apply to unsupervised learning, aimed to find the maximum amount of structure hidden in the data, quantified using this information-theoretic criterion.

The known invariances in the data can be exploited to guide the model design (e.g. as translation invariance leads to convolutional structures, or LSTM is shown to enforce the invariance to time affine transformations of the data sequence [163]). Scattering transforms exploit similar principles [99]. A general theory of how to detect *unknown* invariances in the data, however, is currently lacking.

The view of information theory and Kolmogorov complexity suggests that key program operations (composition, recursivity, use of predefined routines) should intervene when searching for a good computation graph. One possible framework for exploring the space of computation graphs with such operations is that of Genetic Programming. It is interesting to see that evolutionary computation appeared in the last two years among the best candidates to explore the space of deep learning structures [154, 135]. Other approaches might proceed by combining simple models into more powerful ones, e.g. using "Context Tree Weighting" [166] or switch distributions [110]. Another option is to formulate neural architecture design as a reinforcement learning problem [94]; the value of the building blocks (predefined routines) might be defined using e.g., Monte-Carlo Tree Search. A key difficulty is the computational cost of retraining neural nets from scratch upon modifying their architecture; an option might be to use neutral initializations to support warm-restart.

3.3.3 Analyzing and Learning Complex Systems

Participants: Cyril Furtlehner, Aurélien Decelle, François Landes, Michèle Sebag

Methods and criteria from statistical physics have been widely used in ML. In early days, the capacity of Hopfield networks (associative memories defined by the attractors of an energy function) was investigated by using the replica formalism [90]. Restricted Boltzmann machines likewise define a generative model built upon an energy function trained from the data. Along the same lines, Variational Auto-Encoders can be interpreted as systems relating the free energy of the distribution, the information about the data and the entropy (the degree of ignorance about the micro-states of the system) [165]. A key promise of the statistical physics perspective and the Bayesian view of deep learning is to harness the tremendous growth of the model size (billions of weights in recent machine translation netwowrks), and make them sustainable through e.g. posterior drop-out [145], weight quantization and probabilistic binary networks [140]. Such "informational cooling" of a trained deep network can reduce its size by several orders of magnitude while preserving its performance.

Statistical physics is among the key expertises of TAU, originally only represented by Cyril Furtlehner, later strenghtened by Aurélien Decelle's and François Landes' arrivals in 2014 and 2018. On-going studies are conducted along several directions.

Generative models are most often expressed in terms of a Gibbs distributions $P[S] = \exp(-E[S])$, where energy E involves a sum of building blocks, modelling the interactions among variables. This formalization makes it natural to use mean-field methods of statistical physics and associated inference algorithms to both train and exploit such models. The difficulty is to find a good trade-off between the richness of the structure and the efficiency of mean-field approaches. One direction of research pursued in TAU, [113] in the context of traffic forecasting, is to account for the presence of cycles in the interaction graph, to adapt inference algorithms to such graphs with cycles, while constraining graphs to remain compatible with mean-field inference.

Another direction, explored in TAO/TAU in the recent years, is based on the definition and exploitation of self-consistency properties, enforcing principled divide-and-conquer resolutions. In the particular case of the message-passing Affinity Propagation algorithm for instance [167], self-consistency imposes the invariance of the solution when handled at different scales, thus enabling to characterize the critical value of the penalty and other hyper-parameters in closed form (in the case of simple data distributions) or empirically otherwise [114].

A more recent research direction examines the quantity of information in a (deep) neural net along the random matrix theory framework [101]. It is addressed in Giancarlo Fissore's PhD, and is detailed in Section 8.2.3.

Finally, we note the recent surge in using ML to address fundamental physics problems: from turbulence to high-energy physics and soft matter (with amorphous materials at its core) [74] or atrophysics/cosmology as well. TAU's dual expertise in Deep Networks and in statistical physics places it in an ideal position to significantly contribute to this domain and shape the methods that will be used by the physics community in the future. In that direction, the PhD thesis of Marion Ullmo and Tony Bonnaire applying statistical method coming either from deep learning or statistical physics to the task of inferring the structure of the cosmic web has show great succes with recents results discussed in Section 8.2.3. François Landes' recent arrival in the team makes TAU a unique place for such interdisciplinary research, thanks to his collaborators from the Simons Collaboration *Cracking the Glass Problem* (gathering 13 statistical physics teams at the international level). This project is detailed in Section 8.2.3.

Independently, François Landes is actively collaborating with statistical physicists (Alberto Rosso, LPTMS, Univ. Paris-Saclay) and physcists at the frontier with geophysics (Eugenio Lippiello, Second Univ. of Naples) [136, 75]. A CNRS grant (80Prime) finances a shared PhD (Vincenzo Schimmenti), at the frontier between seismicity and ML (Alberto Rosso, Marc Schoenauer and François Landes).

3.4 Organisation of Challenges

Participants: Cécile Germain, Isabelle Guyon, Marc Schoenauer, Michèle Sebag

Challenges have been an important drive for Machine Learning research for many years, and TAO members have played important roles in the organization of many such challenges: Michèle Sebag was head of the challenge programme in the Pascal European Network of Excellence (2005-2013); Isabelle Guyon, as mentioned, was the PI of many challenges ranging from causation challenges [121], to AutoML [122]. The Higgs challenge [89], most attended ever Kaggle challenge, was jointly organized by TAO (C. Germain), LAL-IN2P3 (D. Rousseau and B. Kegl) and I. Guyon (not yet at TAO), in collaboration with CERN and Imperial College.

Many challenges have been organized in the recent years on the Codalab platform, managed by Isabelle Guyon and maintained at LISN. See details in Section 8.6.

4 Application domains

4.1 Computational Social Sciences

Participants: Philippe Caillou, Isabelle Guyon, Michèle Sebag, Paola Tubaro
Collaboration: Jean-Pierre Nadal (EHESS); Marco Cuturi, Bruno Crépon (ENSAE); Thierry Weil (Mines); Jean-Luc Bazet (RITM)

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 [117] 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 [133]. The key technical difficulties related to data and model biases, and to self-fulfilling prophecies have been discussed in section 3.1. Thirdly, CSS does not only regard scientists: it is essential that the civil society participate in the science of society [159].

TAO was involved in CSS for the last five years, and its activities have been strengthened thanks to P. Tubaro's and I. Guyon's expertises respectively in sociology and economics, and in causal modeling. Their departures will impact the team activities in this domain, but many projects are tsill on-going and CSS will remain a domain of choice. Details are given in Section 8.3.

4.2 Energy Management

Participants: Isabelle Guyon, Marc Schoenauer, Michèle Sebag
Collaboration: Rémy Clément, Antoine Marot, Patrick Panciatici (RTE), Vincent Renault (Artelys), Thibault Faney (IFPEN)

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 3.2).

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 8.4 for more details).

Furthermore, predictive models of local grids are based on the estimated consumption of end-customers: Linky meters only provide coarse grain information due to privacy issues, and very few samples of fine-grained consumption are available (from volunteer customers). A first task is to transfer knowledge from small data to the whole domain of application. A second task is to directly predict the peaks of consumption based on the user cluster profiles and their representativity (see Section 8.4.2).

4.3 Data-driven Numerical Modeling

Participants: Guillaume Charpiat, Cécile Germain, Isabelle Guyon, Flora Jay, Marc Schoenauer, Michèle Sebag

As said (section 3.2), 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 [105] 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

Thanks to the pandemia, the impact of our activities regarding carbon footprint have decreased a lot, from our daily commute that have almost completely disappeared as we all switched to tele-working to the transformation of all conferences and workshops into virtual events. We all miss the informal discussions that took place during coffee breaks in the lab as well as during conferences. But when the pandemia vanishes, after the first moments of joy when actually meeting again physically with our colleagues, we will have to think of a new model for the way we work: we were indeed discussing before the pandemia about how to reduce the carbon footpring of the conferences, but now we know that there exist solutions, even though not perfect.

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.

6 Highlights of the year

Herilalaina Rakotoarison, Louisot Milijaona, Andry Rasoanaivo, Michèle Sebag, Marc Schoenauer, **Spotlight paper at ICLR** (top 5% submissions) for the paper [31] Learning Meta-features for AutoML. International Conference on Learning Representations, 2022.

Isabelle Guyon **Keynote, NeurIPS 2022** : *The Data-centric Era: How ML is becoming an experimental science.*

6.1 Awards

Herilalaina Rakotoarison, **First prize** ex æquo at the 2022 PhD Prize in Computer Science, awarded by Labex DigiCosme, the CS doctoral school of Universite‘ Paris Saclay and the doctoral school of IPP (i.e., the whole ”Plateau de Saclay” in CS).

7 New software and platforms

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/

News of the Year:

L2RPN The Learning to Run a Power Network competition track in collaboration with RTE France continues. The ICAPS 2021 competition allowed us to go one step further towards making the grid control with reinforcement learning more realistic, allowing adversarial attacks. A new open-source framework Grid2Operate was released.

AutoDL The Automated Deep Learning (AutoDL) challenge series evolved in the direction of meta learning (https://metalearning.chalearn.org/). We organized a competition fr NwurIPS 2021 sponsored by Google and Microsoft. The results, which will appear in PMLR, indicate that few shot learning (5 shots, 5 classes) is now within reach of the state of the art for small image object recognition, but heavily relies on pre-trained backbone networks, trained on large image datasets.

Industry challenges The first Ile de France industry challenge was organized on Codalab, in collaboration with Dassault aviation and the results were presented at ICMLA 2021. The goal was to predict sensor data indicating constrains on the fuselage. Surprisingly conventional methods based on ensembles of decision trees dominated this task and outperformed deep learnign methods.

World use of the platform In 2021, on average, 50 competitions per month were organized on Codalab by researchers from all over the world. Codalab is also used in education to organize code submission homework.

Codabench December 2021: Codabench (beta) is announced at NeurIPS 2021, see https://www.codabench.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 topics identification, visualisation 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 corpuses. 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

Partners: LRI - Laboratoire de Recherche en Informatique, CNRS

7.2 New platforms

Participants: Guillaume Charpiat, Isabelle Guyon, Flora Jay, Anne-Catherine Letournel, Adrien Pavao, Théophile Sanchez, Dinh Tran Tuan, Benjamin Maudet.

- **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 [65]. 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. 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 [23] 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.

- **DNA-DNA** (Deep Neural Architectures for Dna – https://mlgenetics.gitlab.io/dnadna/) is a package for deep learning inference in population genetics. 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. DNADNA should allow researchers to focus on their research project, be it the analysis of population genetic data or building new methods, without the need to focus on proper development methodology (unit test, continuous integration, documentation, etc.). Results will thus be more easily reproduced and shared. Having a common interface will also decrease the risk of bugs.

8 New results

8.1 Toward Good AI

8.1.1 Causal Modeling

Participants: Philippe Caillou, Isabelle Guyon, Michèle Sebag
PhDs: Armand Lacombe, Cyriaque Rousselot, Nicolas Atienza
Post-doc: Shuyu Dong, Shiyang Yan
Collaboration: Olivier Allais (INRAE); Julia Mink (Univ. Bonn); Jean-Pierre Nadal & Annick Vignes (CAMS, EHESS); David Lopez-Paz (Facebook).

This year, the long awaited journal version of SAM (Structural Agnostic Modelling), has been published in JMLR [18], long after Diviyan Kalainathan's PhD [127]. The causal modelling activity continues with three main directions in 2022. The first one is tackled in collaboration with INRAE (Cyriaque Rousselot's PhD), within the Horapest DataIA project. 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*. Besides getting the data[5] the difficulty lies in observational causal modelling from spatio-temporal data with hidden confounders. 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 [168] is of cubic complexity in the number p 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 [25]. Our current approach aims to distinguish the statistical and the geometrical errors of SEM identification, respectively related with the estimation of the inverse covariance matrix, and the projection of the associate causal graph on the DAG space [56]. A third direction is considered with Nicolas Atienza (PhD Cifre Thales), co-supervised with Johanne Cohen, LISN. The goal is to extend algorithmic recourse [129] 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.

Finally, causality is also at the core of TAU participation in the INRIA Challenge *OceanIA*, that started in 2021 [86]. Shiyang Yan's post-doc is dedicated 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, are handled using generalized contrastive losses and introducing fake outliers extracted from face images, or created as chimeras.

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

8.1.2 Explainability

Participants: Isabelle Guyon, François Landes, Alessandro Leite, Marc Schoenauer, Michèle Sebag
PhD: Cyriaque Rousselot
Collaboration: MyDataModels; Thales

[5] The IRB and access demand to the data have been accepted and the HDH data will be available in Sept. 2023.

In Roman Bresson's PhD [46], (coll. LISN-GALAC, Thales, U. Paderborn; pending patent Bresson-Labreuche-Sebag-Cohen), the goal was to adapt a neuronal architecture to yield an interpretable-by-design model. The extension of this approach is investigated to transform an accurate black-box into a hierarchical choquet integral.

The team is also involved in the proposal for the IPL HyAIAI (Hybrid Approaches for Interpretable AI), coordinated by the LACODAM team (Rennes) dedicated to the design of hybrid approaches that combine state of the art numeric models (e.g., deep neural networks) with explainable symbolic models, in order to be able to integrate high level (domain) constraints in ML models, to give model designers information on ill-performing parts of the model, to provide understandable explanations on its results. On-going collaboration with the Multispeech team in Nancy is concerned with co-supervision of G. Zervakis' PhD (to be defended March 2023), and concerns the use of background knowledge to improve the performances of foundational models in NLP [81] and an analogy based approach for solving target sense verification [34].

An original approach to DNN explainability might arise from the study of structural glasses (8.2.3), with a parallel to Graph Neural Networks (GNNs), that could become an excellent non-trivial example for developing explainability protocols, as we already suggest from results in [66].

Build on collaboration with Raymond Poincaré Hospital, the team is developing tools to increase the interpretability of medical data in applicative context. A first study published in [17] investigates how geometric methods could represent the evolution of patients' key indicators on a curved manifold to generate meaningful and interpretable representation. These representations could be generalized with minor modifications to temporal data.

Genetic Programming [92] is an Evolutionary Computing technique that evolves models as analytical expressions (Boolean formulae, functions, LISP-like code), that are hopefully easier to understand than black-box NNs with hundreds of thousands of weights. This idea has been picked up by the European FET project TRUST-AI (Transparent, Reliable and Unbiased Smart Tool for AI) that started in October 2020. Alessandro Leite joined the project (and the TAU team) in February 2021 on an ARP position. First work addressed explainable reinforcement learning using GP [33]. Current work, recently acepted to EuroGP 2023, concerns the adaptation of the Memetic Semantic Generic Programming [112] to the continuous case. Furthermore, a collaboration with O. Teytaud (Meta), around the follow-up is Mathurin's CIFRE PhD (started Oct. 2022).

8.1.3 Robustness of AI Systems

Participants: Guillaume Charpiat, Marc Schoenauer, Michèle Sebag
PhDs: Roman Bresson
Collaboration: Johanne Cohen (LISN-GALAC) and Christophe Labreuche (Thalès); Eyke Hullermeier (U. Paderborn, Germany).

Though several on-going activitires in this domain have been put on hold (see Section 3.1.2), new research lines have started to emerge, pertaining to robustness.

The first one, already described in section 8.1.2, concerns the indentifiability of the neural net implementing a hierarchical Choquet integral, in the large sample limit.

Another direction, part of A. Lacombe's on-going PhD, is concerned with privacy. Our primary motivation was to contribute to the understanding of the pandemy, with no former collaboration with hospitals, and therefore, no access to real data. An approach was developed to achieve excessively private learning through a differential-privacy compliant access to the only marginals of the data [82].

We have also explored relationships between theoretical guarantees provided by differential privacy and membership interence attacks [30], as described in Section 8.5.3.

8.2 Learning to Learn

8.2.1 Auto-*

Participants: Guillaume Charpiat, Isabelle Guyon, Marc Schoenauer, Michèle Sebag
PhDs: Léonard Blier, Adrien Pavao, Herilalaina Rakotoarison, Hoazhe Sun, Manon Verbockhaven, Romain Egele
Collaborations: Vincent Renault (SME Artelys); Yann Ollivier (Facebook); Wei-Wei Tu (4Paradigm, Chine);

André Elisseeff (Google Zurich); Prasanna Balaprakash (Argonne National labs), among others (for a full list see https://autodl.chalearn.org/ and https://metalearning.chalearn.org/)

Auto-⋆ studies at TAU investigate several research directions.

After proposing MOSAIC [153], that extends and adapts Monte-Carlo Tree Search to explore the structured space of pre-processing + learning algorithm configurations, and performs on par with AutoSklearn, the winner of Auto-⋆ international competitions in the last few years, Herilalaina Rakotoarison explored in the end of his PhD the learning of meta-features for tabular data, addressing the lack of expressiveness of the standard Hand-Crafted ones [50]. The idea is to use Optimal Transport to align the distribution of the datasets from the training meta-data with that of their best hyperparameter settings in the space of hyperparameter conigurations. The results were presented as a spotlight (top 5% submissions) at ICLR 2022 [31].

Heri also contributed to a large benchmarking effort together with Olivier Teytaud, former member of the team, now with Facebook AI Research [19].

In a second direction, with the internship and starting PhD thesis of Manon Verbockhaven, we adopt a functional analysis viewpoint in order to adapt on the fly the architecture of neural networks that are being trained. This allows to start training neural networks with very few neurons and layers, and add them where they are needed, instead of training huge architectures and then pruning them, a common practice in deep learning, for optimization reasons. For this, we quantify the lack of expressivity of a neural network being trained, by analyzing the difference between how the backpropagation would like the activations to change and what the tangent space of the parameters offers as possible activation variations. We can then localize the lacks of expressivity, and add neurons accordingly. It turns out that the optimal weights of the added neurons can be computed in closed form.

A last direction of investigation concerns the design of challenges, that contribute to the collective advance of research in the Auto-⋆ direction. The team has been very active in the series of AutoML and AutoDL, which has been extended to Meta-Learning, with support from Microsoft, Google, 4Paradigm and ChaLearn, to Meta-Learning, namely meta-learning from learning curves [28, 64] and cross-domain meta-learning [36].

Sef-supervised learning seems to be an avenue with great future, allowing to train reprsentations without costly human labeling. A new challenge accepted as part of the WCCI competition program 2022 is currently running. Another challenge on Neural Architecture Search (NAS) has been run together with a workshop at the CVPR 2021 conference. Preliminary results on NAS have been produced by one of our interns (Romain Egele [77]). Further developments have led to effective algorithms to conduct simultaneously NAS and hyper-parameter selection [27, 57]. More details on challenges are found in Section 8.6).

8.2.2 Deep Learning: Practical and Theoretical Insights

Participants: Guillaume Charpiat, Isabelle Guyon, Marc Schoenauer, Michèle Sebag
PhDs: Léonard Blier, Zhengying Liu, Adrien Pavao, Haozhe Sun, Romain Egele
Collaboration: Yann Ollivier (Facebook AI Research, Paris)

Although a comprehensive mathematical theory of deep learning is yet to come, theoretical insights from information theory or from dynamical systems can deliver principled improvements to deep learning and/or explain the empirical successes of some architectures compared to others.

During his CIFRE PhD with Facebook AI Research Paris, co-supervised by Yann Ollivier (former TAU member) [45] , Léonard Blier has properly formalized the concepts of *successor states* and *multi-goal functions* [87], in particular in the case of continuous state spaces. This allowed him to define unbiased algorithms with finite variance to learn such ojects, including the continuous case thanks to approximation functions. In the case of finite environments, new convergence bounds have been obtained for the learning of the value function. These new algorithms capable of learning successor states in turn lead to define and learn new representations for the state space.

The AutoDL challenges, co-organized in TAU (in particular by Isabelle Guyon), also contribute to a better understanding of Deep Learning. It is interesting to note that no Neural Architecture Search algorithm was proposed to solve the different challenges in AutoDL (corresponding to different data types). See section 8.6 for more details.

Our PhD student Haozhe Sun is continuing to work on the problem of modularity in Deep Learning. He wrote a survey under revision and recently submitted a paper on novel algorithms for low-cost AI exploiting modularity. 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 heath-care applications, to name a few) DL models are still notoriously hard to deploy in new applications. In particular, the require massive numbers of training examples, hours of GPU training, and highly qualified engineers to hand-tune their architectures. This thesis will contribute to reduce the barrier of entry in using DL models for new applications, a step towards "democratizing AI".

Romain Egele in his PhD in collaboration with Argonne National Labs (USA), is been 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 frameworks [27]. 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. In 2022 he published two papers on hyper-parameter optimization. In [57], he proposes to exploit the parallelism in large compute clusters to speed up Bayesian hyper-parameter search in an effective way. In [26] he extends Bayesan optimization algorithms to perform accurate estimations of uncertainties.

8.2.3 Analyzing and Learning Complex Systems

Participants: Cyril Furtlehner, Aurélien Decelle, François Landes, Guillaume Charpiat
PhDs: Giancarlo Fissore, Marion Ullmo
Collaboration: Jacopo Rocchi (LPTMS Paris Sud); the Simons team: Rahul Chako (post-doc), Andrea Liu (UPenn), David Reichman (Columbia), Giulio Biroli (ENS), Olivier Dauchot (ESPCI).; Clément Vignax (EPFL); Yufei Han (Symantec), Nabila Aghanim (Institut d'Astrophysique Spatiale), Tony Bonnaire (ENS Paris).

Generative models constitute an important piece of unsupervised ML techniques which is still under rapid developpment. In this context insights from statistical physics are relevent in particular for energy based models like restricted Boltzmann machines. The information content of a trained restricted Boltzmann machine (RBM) and its learning dynamics can be analyzed precisely with help of ensemble averaging techniques [103, 104]. More insight can be obtained by looking at data of low intrinsic dimension, where exact solutions of the RBM can be obtained [70], thanks to a convex relaxation. In particular we have found a 1st order transition mechanisms that may plague the learning in a more advanced part of the learning.In [11] we investigate further this question and show that sampling the equilibrium distribution using 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 computation of the log-likelihood gradient during training, leading to dramatic improvements in 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 PCD recipes. Learning dynamics has also been adressed in a different context of feature learning processes in [59] where closed form expressions are obtained for train and test errors via random matrix theory, 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. We have also investigated models of decision-making with approaches grounded in statistical physics, that are able to predict experimental observed data [68]. A last point concerns road traffic forecasting, a long standing application of mean-field inference methods based on probabilistic modelling. In [16] we wrap up some of the techniques developed in past works and perform comprehensive experimental tests on various real world Urban traffic dataset, thanks to PTV-SISTeMA, showing the effectiveness of our method.

As mentioned earlier, the use of ML to address fundamental physics problems is quickly growing. In that direction in the context of T. Bonnaire's PhD [97], is undertaken in [12] the first comprehensive and quantitative real-space analysis of the cosmological information content in the environments of the cosmic web (voids, filaments, walls, and nodes) up to non-linear scales, k=0.5 h/Mpc, using a method based on the Gaussian mixture model with a prior forcing the centers to "live" on a tree-graph [13] [98, 71]. This method has been further developed for handling in particular possible outliers and put into a general formalism [72].

Another place where ML can help address fundamental physics questions is the domain of glasses [40] (how the structure of glasses is related to their dynamics), which is of the major problems in modern theoretical physics (glasses are a key part of the carrer of Giorgio Parisi, a 2021 nobel prize laureate). The idea is to let ML models automatically find the hidden structures (features) that control the flowing or non-flowing state of matter, discriminating liquid from solid states. These models can then help identifying "computational order parameters", that would advance the understanding of physical phenomena [74, 71], on the one hand, and support the development of more complex models, on the other hand. Attacking the problem of amorphous condensed matter by novel Graph Neural Networks (GNN) architectures is a very promising lead, regardless of the precise quantity one may want to predict. Currently GNNs are engineered to deal with molecular systems and/or crystals, but not to deal with amorphous matter. This strategy is currently being investigated by Francesco Pezzicoli (PhD student), who has already demonstrated the generalizing abilities of rotation-equivariant GNNs [66].

8.3 Computational Social Sciences

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, the labor of human annotators for AI data, quality of life and economic performance), about nutrition (health, food, and socio-demographic issues), around Cartolabe, a platform for scientific information system and visual querying.

8.3.1 Labor Studies

Participants: Philippe Caillou, Isabelle Guyon, Michèle Sebag, Paola Tubaro
PhDs: Guillaume Bied, Armand Lacombe, Assia Wirth
Engineers: Victor Alfonso Naya
Collaboration: Jean-Pierre Nadal (EHESS); Bruno Crépon (ENSAE); Antonio Casilli, Ulrich Laitenberger (Telecom Paris); Odile Chagny (IRES); Francesca Musiani, Mélanie Dulong de Rosnay (CNRS); José Luis Molina (Universitat Autònoma de Barcelona); Antonio Ortega (Universitat de València); Julian Posada (University of Toronto)

A first area of activity of TAU in Computational Social Sciences is the study of labor, from the functioning of the job market, to the rise of new, atypical forms of work in the networked society of internet platforms, and the quality of life at work.

Job markets The DATAIA project Vadore (partners ENSAE and Pôle Emploi) benefits from the sustained cooperation and from the wealth of data gathered by Pôle Emploi. 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). Primary online testing (beta-test campaigns) have assessed the suitability of the recommendations. A second round of testing at the region Rhone-Alpes scale will take place in 2023.

The learned models are inspected w.r.t. several criteria and requirements. A first criterion regards the robustness of the recommender performances under non-stationary distributions, e.g. due to the Covid pandemy [35]. Another 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 were investigated in [83, 85], using optimal transport; and this direction will be pursued in S. Nathan's PhD.

A third criterion regards the fairness of the recommendation model. A gender-related gap in several utilities (wages, types of contract, distance-to-job) is observed by contrasting the jobs recommended to men and women, everything else being equal. Most interestingly, this gap parallels the gap among the jobs actually occupied by men and women (everything else being equal). Several directions of research are considered based on this fact, depending on the regulations to be enforced (in French or European public services). The first direction consists in integrating a risk-avoidance sub-model in the recommendation model, to decouple the prejudice effects (from the recruiters' decisions) and the social conditioning (from the job seekers' preferences). The second one aims at defining new population-based and individual-based performance criteria for a job recommender system.

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.

The human labor behind AI

We look at business-to-business platform labour [55] and more specifically at the data "micro-workers" who perform essential, yet marginalized and poorly paid tasks such as labeling objects in a photograph, translating or transcribing short texts, or recording utterances. Micro-workers are recruited through specialist intermediaries across supply chains that span the globe and reproduce inherited North-South outsourcing relationships [44]. Further observed inequalities are gender-based [22]. Despite the opportunity to telework, the COVID-19 pandemic has adversely affected these workers, widening the gap that separates them from the formally employed [21].

Current work extends this research to workers' skills, competencies and workplace learning practices in an environment in which they support machine learning [20], and to the resilience of these emerging labour markets [43]

8.3.2 Health and practices

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

Continuing our former partnership with INRAE (in the context of the *Initiative de Recherche Stratégique* Nutriperso; [78]), we proposed the HORAPEST 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 Sept. 2023, and contacts have been taken with the CHU Toulouse for cooperation on complementary data.

8.3.3 Scientific Information System and Visual Querying

Participants: Philippe Caillou, Michèle Sebag
Engineers: Anne-Catherine Letournel, Victor Alfonso Naya
Collaboration: 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 [69].

8.4 Energy Management

8.4.1 Power Grids Management

Participants: Isabelle Guyon, Marc Schoenauer
PhDs: Balthazar Donon, Wenzhuo Liu
Collaboration: Rémi Clément, Patrick Panciatici (RTE)

Our collaboration with RTE, during Benjamin Donnot's (2016-2019) [106] and Balthazar Donon's [47] CIFRE PhDs, is centered on the maintainance of the national French Power Grid. In order to maintain the so-called "(n-1) safety" (see Section 4.2), fast simulations of the electrical flows on the grid are mandatory, that the home-brewed simulator HADES is too slow to provide. The main difficulty of using Deep Neural Networks surrogate models is that the topology of the grid (a graph) should be taken into account, and because all topologies cannot be included in the training set, this requires out-of-sample generalization capabilities of the learned models.

Balthazar Donon developped in his PhD [47] an approach based on Graph Neural Networks (GNNs). From a Power Grid perspective, GNNs can be viewed as including the topology in the heart of the structure of the neural network, and learning some generic transfer function amongst nodes that will perform well on any topology. His work uses a loss that directly aims to minimize Kirshhoff's law on all lines. Theoretical results as well as a generalization of the approach to other optimization problems had been originaly published at NeurIPS 2021 [107].

Eva Boguslawski's CIFRE PhD, that started in Sept. 2022, addresses the problem of global monitoring of the grid through decentralized decision process (aka multi-agent Reinforcement Learning), in the line of the LR2PN challenge (see Section 8.6) that she contributed to organize during a previous internship [32].

8.4.2 Optimization of Local Grids

Participants: Isabelle Guyon, Marc Schoenauer, Michèle Sebag
PhDs: Herilalaina Rakotoarison
Collaboration: Vincent Renault (Artelys).

One of the goals of the ADEME Next project, in collaboration with SME Artelys (see also Section 4.2), is the sizing and capacity design of regional power grids. Though smaller than the national grid, regional and urban grids nevertheless raise scaling issues, in particular because many more fine-grained information must be taken into account for their design and predictive growth.

Regarding the design of such grids, and provided accurate predictions of consumption are available (see below), off-the-shelf graph optimization algorithms can be used. However, they require a careful tuning of their hyperparameters, and this was the motiaton of funding Herilalaina Rakotoarison's PhD, that tackles the automatic tuning of such hyer-parameters (see Section 8.2.1); both the Mosaic algorithm [153] and the Metabu algorithm to learn meta-features are being used for Artelys' home optimizer Knitro, and compared to the state-of-the-art in parameter tuning (confidential deliverable).

8.4.3 Accelerating simulation codes

Participants: Guillaume Charpiat, Marc Schoenauer, Michèle Sebag
PhDs: Matthieu Nastorg
Post-doc: Tamon Nakano

Collaboration: Alessandro Bucci (Safran Tech, former member of the team); Thilbault Faney et Jean-Marc Gratien (IFPEN).

During the 2.5 years that he spent at TAU, funded by the bilateral project with IFPEN, Alessandro Bucci worked on several use case of IFPEN, with the goal of accelerating some softwares that IFPEN uses daily. This IFPEN/TAU collaboration lead to a successful application to a DATAIA program with the ML4CFD project. Direct follow-up of the previous collaboration , a prominent result was obtained on the simulation of diphasic fluid flow in a distillation column: one of the main timeconsuming step in the simulation is the tracking of the interface between the bubbles of the gaz and the liquid they circulate in within the Volume-of-Fluid numerical method: this critical step was replaced with a Graph Neural Network model directly working on the unstrutured mesh, making the industrial application possible [63]. ML4CFD is also funding Matthieu Nastorg's PhD, who significantly accelerated [38] 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 Statistical Solvers [107]. Note that this resul is more general than its application to energy problems, but was made possible only because of the collaboration with IFPEN.

8.5 Data-driven Numerical Modelling

8.5.1 Space Weather Forecasting

Participants: Cyril Furtlehner, Michèle Sebag
Post-doc: Olivier Bui
Collaboration: Jannis Teunissen (CWI)

Space Weather is broadly defined as the study of the relationships between the variable conditions on the Sun and the space environment surrounding Earth. Aside from its scientific interest from the point of view of fundamental space physics phenomena, Space Weather plays an increasingly important role on our technology-dependent society. In particular, it focuses on events that can affect the performance and reliability of space-borne and ground-based technological systems, such as satellite and electric networks that can be damaged by an enhanced flux of energetic particles interacting with electronic circuits.[6]

Since 2016, in the context of the Inria-CWI partnership, a collaboration between TAU and the Multi-scale Dynamics Group of CWI aims to long-term Space Weather forecasting. The goal is to take advantage of the data produced everyday by satellites surveying the sun and the magnetosphere, and more particularly to relate solar images and the quantities (e.g., electron flux, proton flux, solar wind speed) measured on the L1 libration point between the Earth and the Sun (about 1,500,000 km and 1 hour time forward of Earth). A challenge is to formulate such goals in terms of supervised learning problem, while the "labels" associated to solar images are recorded at L1 (thus with a varying and unknown time lag). In essence, while typical ML models aim to answer the question *What*, our goal here is to answer both questions *What* and *When*. This project has been articulated around Mandar Chandorkar's Phd thesis [100] which has been defended this year in Eindhoven. The continuation of this collaboration is inseured by the hiring of Olivier Bui as a post-doc who's work has consisting in extending preliminary results on solar wind forecasting based on auto-encoded solar magnetograms on a longer period of data corresponding to 2 solar cycles. Negative results have incited us to dig more into physical models of solar wind propagation and try to combine them with ML models in a systematic way.

8.5.2 Genomic Data and Population Genetics

Participants: Guillaume Charpiat, Flora Jay, Aurélien Decelle, Cyril Furtlehner
PhD: Théophile Sanchez, Jérémy Guez
PostDoc: Jean Cury, 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)

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.

[6] After a recent survey conducted by the insurance company Lloyd's, an extreme Space Weather event could produce up to $2.6 trillion in financial damage.

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).In Theophile Sanchez' PhD [51], we propose an alternative to summary statistics, based on the automatic extraction of relevant information using deep learning techniques. Specifically, we design artificial neural networks (ANNs) that take as input single nucleotide polymorphic sites (SNPs) found in individuals sampled from a single population and infer the past effective population size history. First, we provide guidelines to construct artificial neural networks that comply with the intrinsic properties of SNP data such as invariance to permutation of haplotypes, long scale interactions between SNPs and variable genomic length. Thanks to a Bayesian hyperparameter optimization procedure, we evaluate the performance of multiple networks and compare them to well established methods like Approximate Bayesian Computation (ABC). Even without the expert knowledge of summary statistics, our approach compares fairly well to an ABC based on handcrafted features. Furthermore we show that combining deep learning and ABC can improve performance while taking advantage of both frameworks. Later, we experimented with other types of permutation invariance, based on similar architectures, and achieved a significative performance gain with respect to the state of the art, including w.r.t. ABC on summary statistics (20% gap), which means that we extract information from raw data that is not present in summary statistics. The question is now how to express this information in a human-friendly way.

In the short-term these architectures can be used for demographic inference [60] or selection inference in bacterial populations (ongoing work with a postdoctoral researcher, J Cury, collab: Pasteur Institute, for ancient DNA: UNAM and U Brown); the longer-term goal is to integrate them in various systems handling genetic data or other biological sequence data. Regarding the bacterial populations, we already implemented a flexible simulator that will allow researchers to investigate complex evolutionary scenarios (e.g. dynamics of antibiotic resistance in 2D space through time) with realistic biological processes (bacterial recombination), which was impossible before (collab. U Cornell, MNHN) [14].

In collaboration with the Institute of Genomics of Tartu, we leveraged 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 [76]. 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. Ongoing work concerns scaling up to the full genome and developing new privacy scores.

We released dnadna, a flexible open-source python-based software for deep learning inference in population genetics[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[88].

8.5.3 Privacy and synthetic data generation

Participants: Isabelle Guyon
PhD: Adrien Pavao
Collaboration: Kristin Bennett and Joe Pedersen (RPI, NY, USA), Wei-Wei Tu (4Paradigm, Chine), Pablo Piantanida (Centrale-Supelec)

While theoretical criteria of privacy preservation, such as “differencial privacy” are important to gain insights into how to protect privacy, they are often impractical, because they put forward pessimistic

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

bounds and impose degrading data and/or model to a point that hampers utility. Additionally, for all practical purposes, data owners seek to obtain guarantees that no privite information is leaked in the form of an empirical statistical test, rather than a more elusive theoretical guarantee. To that end, we have set to work on evaluating the effectiveness of privacy protection agains specific attacks, such as membership inference or attribute inference. We devised an evaluation apparatus called "LTU-attacker" [30], in collaboration with Kristin Bennett, Joe Pedersen, and Wei-Wei Tu and with 2 interns (Rafel Monos-Gomez and Jiangna Huang) have obtained interesting preliminary results demonstrating lack of privacy preservation of most scikit-learn algorithms under membership inference attacks. New directions currently explored in collabotation with Pablo Piantanida include defining a degree of "privacy exposure" of particular individual involving information theoretic arguments.

With Master student Alice Lacan, we have been investigating the modelization of the Covid-19 epidemic propogation using compartimental models, following earlier work by former master student Martin Cepeda. A group of students including Alice entered the "Pandemic response" XPrize and qualified for the final phase. This work was follwed by a paper on estimating uncertainty in time series, in application to prediciting the evolution of the number of Covid cases presented at the BayLearn 2022 conference[84]. Alice was invited to give a presentation of this work at the WIDS 2023 conference.

Last but not least regarding Covid-19, F. Landes participated to Inria Saclay collaborative effort to monitor and optimize the emergency bed occupancy in East of France [24].

8.5.4 Sampling molecular conformations

Participants: Guillaume Charpiat
PhD: Loris Felardos
Collaboration: Jérôme Hénin (IBPC), Bruno Raffin (InriAlpes)

Numerical simulations on massively parallel architectures, routinely used to study the dynamics of biomolecules at the atomic scale, produce large amounts of data representing the time trajectories of molecular configurations, with the goal of exploring and sampling all possible configuration basins of given molecules. The configuration space is high-dimensional (10,000+), hindering the use of standard data analytics approaches. The use of advanced data analytics to identify intrinsic configuration patterns could be transformative for the field.

The high-dimensional data produced by molecular simulations live on low-dimensional manifolds; the extraction of these manifolds will enable to drive detailed large-scale simulations further in the configuration space. We study how to bypass simulations by directly predicting, given a molecule formula, its possible configurations. This is done using Graph Neural Networks [58] in a generative way, producing 3D configurations, and constitutes the main part of Loris Felardos' PhD [48], funded by the Inria Challenge HPC/Big Data. The goal is to sample all possible configurations, and with the right probability. This year we studied various normalizing flow architectures as well as varied training criteria suitable for distributions (Kullback-Leibler divergence in latent or sample space, in one direction or the other one, as it is not symmetric, but also pairwise distances, optimal transport, etc.). It turns out that mode collapse is frequently observed in most cases, even on simple tasks. Further analysis identified several causes for this, from which we built remedies.

8.5.5 Earthquake occurrence prediction

Participants: François Landes, Marc Schoenauer
PhD: Vincenzo Schimmenti
Collaboration: Alberto Rosso (LPTMS)

Earthquakes occur in brittle regions of the Crust typically located at the depth of 5-15 km and characterized by a solid friction, which is at the origin of the stick-slip behaviour. Their magnitude distribution displays thecelebrated Gutenberg-Richter law and a significant increase of the seismic rate is observed after large events (called main shocks). The occurrence of the subsequent earthquakes in the same region, the aftershocks, obeys well established empirical laws that demand to be understood. A change in the seismic rate also happens before a main shock, with an excess of small events compared to the expected rate of aftershocks related to the previous main shock in that region. These additional events are defined as foreshocks of the coming main shock, however they are scarce so that defininig them is a very difficult task. For this reason their statistical fingerprint, so important for human secutiry, remains elusive.

The treatment of rare events by Machine Learning is a challenging yet rapidly evolving domain. At TAU we have a great expertise in data modeling, in particular Bayesian models and Restricted Boltzman Machines (RBMs) have been built to model space weather forecast data (Section 8.5.1). These techniques, inspired from statistical physics, are both based on a probabilistic description of latent variables, allowing the modelling of a large span of data correlations. This kind of models can be extended to study spatially resolved earthquakes, the latent variable here being the local stress within the fault and in the ductile regions. Our goal is to characterize the statistical properties of a sequence of events (foreshocks, main shock and aftershocks) and predict its following history. We will first study the sequences obtained from simulations of the physical model [75]. We will answer the following question: given a short sequence of foreshocks, can we predict the future of the sequence? How big will be the main shock? When will it occur? In a second step we will use also the data coming from real sequences, where events are unlabeled. These sequences are public and available (The most accurate catalog is for Southern California, a catalog with 1.81 million earthquakes. It is available at https://scedc.caltech.edu/research-tools/QTMcatalog.html). Concretely, the data consists in the earthquakes' magnitude, occurrence time and hypocenter locations.

Two parallel directions are being explored, with our PhD Student, Vincenzo Schimmenti:

- The available data can be used to tune the parameters of the new model to improve its accuracy and generalization properties. We will adjust the parameters of the elastic and friction coefficients in order to produce earthquakes with realistic magnitudes. This will allow us to have information about the physical condition in the fault and in the ductile regions.

- We will use our understanding of foreshocks statistics to perform classification of earthquakes with respect to their nature: foreshock, main shock or after shock, and alignment (assignment of the earthquake to a sequence). These labels are known in the synthetic data and unknown in the catalogs, so this would be an instance of semi-supervised learning. Our final goal is real data completion: presented with an incomplete catalog, the machine is asked to complete it with the missing points.

8.5.6 Reduced order model correction

Partecipants: Michele Alessandro Bucci, Marc Schoenauer
PhD: Emmanuel Menier
Collaboration: Mouadh Yagoubi (IRT-SystemX), Lionel Mathelin (DATAFLOT team, LISN)

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 dynamic 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. We tested this solution on benchmarked cases where it is well known that transient dynamics are poorly represented by ROMs. The results [62] show how the correction term improves the prediction while preserving the guarantees of the ROM. Furthermore, the approach was generalized, and the extension was validated on Michelin use case of rubber calendering process [61].

8.5.7 Active Learning for chaotic systems

Participants: Michele Alessandro Bucci (now with SafranTech)
Collaboration: Lionel Mathelin (LISN), Onofrio Semeraro (LISN), Sergio Chibbaro (UPMC), Alexandre Allauzen (ESPCI)

The inference of a data driven model aiming at reproducing chaotic systems is challenging even for the most performing Neural Network architectures. According to the ergodic theory, the amount of data required to converge to the invariant measure of a chaotic system goes exponentially with its intrinsic dimension. Recent work [54] analyzes the amount of data that is sufficient for a priori guaranteeing a faithful model of the physical system.

8.5.8 Graph Neural Networks for Numerical Simulations

Participants: Guillaume Charpiat, Marc Schoenauer, Michèle Sebag
PhDs: Balthazar Donon, Loris Felardos, Wenzhuo Liu, Matthieu Nastorg
Post-doc: Tamon Nakano
Collaboration: Mouahd, Yagoubi (IRT SystemX), Lionel Mathelin (LISN), Alessandro Bucci (Safran Tech, former member of the team); Thilbault Faney et Jean-Marc Gratien (IFPEN).

Many of the works that have been introduced earlier featured the use of Graph Neural Networks to learn how to solve numerical problems invoving data on graphs: Balthazar Donon's PhD [47] simulated the French Power Grid, Loris Felardos' PhD [48] learnt the distribution of 3D molecule conformations, Matthieu Nastorg accelerate the numerical resolution of Poisson's equation on any unstructured mesh with GNNs [38], Tamon Nakano also handled unstructured meshes with GNNs to track the interface between both fluids in multi-phasic flow simulations [63].

But the use of GNNs to approximate the numerical solutions of PDEs on any unstructured mesh, rather than using grid meshes to be able to use the CNNs and the whold zoology of Deep Neural Networks designed for image processing was systematicall studies in Wenzhuo Liu's PhD: After porting ideas from multi-grid approaches to Finite Elements, and comparing the CNN and GNN approchaes [138], she tackled the poor Out-of-Distribution generalization issue using Meta-Learning [37], improving the OoD learning on CFD simulations of air flow around an airfoil by considering each airfoil shape as a separate task. he is now completing her PhD (defense in March) by applying Transfer Learning to decrease the amount of data to learn accurate simulation on fine meshes using numerous costless simulations on coarse meshes (submitted).

8.6 Challenges

Participants: Cécile Germain, Isabelle Guyon, Adrien Pavao, Anne-Catherine Letournel, Marc Schoenauer, Michèle Sebag
PhD: Eva Boguslawski, Balthazar Donon, Adrien Pavao, Haozhe Sun, Romain Egele
Engineer: Sébastien Tréguer.
Collaborations: D. Rousseau (LAL), André Elisseeff (Google Zurich), Jean-Roch Vilmant (CERN), Antoine Marot and Benjamin Donnot (RTE), Kristin Bennett (RPI), Magali Richard (Université de Grenoble), Wei-Wei Tu (4Paradigm, Chine), Sergio Escalera (U. Barcelona, Espagne).

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 will also receive the support of the Chaire Nationale d'Intelligence Artificielle of Isabelle Guyon (2020-2024).

Adrien Pavao has also set to work on the theoretical rationalization of judging competitions. A first work built ties between this problem and the theory of social choice [80]. 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.

Following the highly successful ChaLearn **AutoML** Challenges (NIPS 2015 – ICML 2016 [122] – PKDD 2018 [123])and **AutoDL** [139] was run in 2019 (see http://autodl.chalearn.org), that pointed to the importance of meta-learning, we opened a new line of research on meta-learning from learning curves [28, 64] and cross-domain meta-learning [36]. This led us to explore uses of reinforcement learning as a means to devise policies for meta-learning (on-going). In parallel, a new challenge on automated reinforcement learning (AutoRL) is currently under design.

A new challenge series in Reinforcement Learning for Power Grid control was started in 2021 with the company RTE France on the theme "Learning to run a power network" [142] (**L2RPN**, http://l2rpn.chalearn.org). 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 first edition was run in Spring 2019, and aimed at demonstrating the feasibility of applying Reinforcement Learning for controlling electrical flows on a power grid. The 2020 edition [141] introduced a realistically-sized grid environment along with two fundamental real-life properties of power grid systems to reconsider while shifting towards a sustainable world: robustness and adaptability, and the 2022 edition was concerned with changing topology, and was co-organized with RTE and TAILOR challenge task force (TAU team is responsible of the organization of challenges within the European project TAILOR). The analysis paper is under review.

In preparation, with the sponsorship of Paris-Région Ile de France, a competition between statups in the AI challenge for Industry series is being organized, in collaboration with RTE [32]. The competition is assorted with a 1 millon Euro prize pool. The objective is to device 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.

Paris Ile-de-France region selected in 2021 Codalab and the TAU team to organize the industry machine learning challenge seris of the Paris Region. Adrien Pavao, who was the project leader, organized with Dassault aviation a project of "numerical twins", aiming at performing predictive maintenance on airplanes. The Paris Region offered 500K Euros to the winner, a startup, which would then collaborate with Dassault to productize the solution. The challenge took place from February 2021 to May 2021. The results have indicated that, on such problems of time series regression, ensembles of decision trees such as XGBoost dominate over DL methods. This result, which came somewhat as a surprise, but stem from the massive amount of data that had to be processed. Despite the significant compute power made avaliable (10 GPUs for 2 days), search for optimal architectures was difficult. Results of detailed analyses conducted by a consortium of organizers and participants have been published [79]. This challenge has demonstrated that Codalab is now "industry grade", and has paved the way to organizing other AI for Industry challenges. We have currently in preparation a challenge targeting carbon-neutrality by 2025, in collaboration with RTE-France.

It is important to introduce challenges in ML teaching. This has been done (and is on-going) in I. Guyon's Master courses [148] : some assignments to Master students are to design small challenges, which are then given to other students in labs, and both types of students seem to love it. Codalab has also been used to implement reinforcement learning homework in the form of challenges by Victor Berger and Heri Rakotoarison for the class of Michèle Sebag. New directions being explored by students in 2021 include takling fairness and bias in data.

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 Springer series on challenges in machine learning, see http://www.chalearn.org/books.html. Challenge organizaton is now better grounded in theory, with such effort. The thesis of Adrien Pavao will include 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" [29].

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 3 on-going CIFRE PhDs, with the corresponding side-contracts with the industrial supervisor, one bilateral contract with IFPEN, one recently started bilateral contract with Fujitsu (within the national "accord-cadre" Inria/Fujitsu), plus at least two new CIFRE PhDs, one with our long-lasting partner RTE, and one with Ekimetrics company, with whom we have never worked before), that will start in 2022.

- **IFPEN** (Institut Français du Pétrole Energies Nouvelles) 2019-2023 (300 kEuros), to hire an Inria Starting Research Position (Alessandro Bucci) to work in all topics mentioned in Section 3.2 relevant to IFPEN activity.
  Coordinator: Marc Schoenauer
  Participants: Alessandro Bucci, Guillaume Charpiat
- **Fujitsu**, 2021-2022 renewed 2022-2023 (200k€ per year), *Causal discovery in high dimensions*
  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
  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: Isabelle Guyon (now Michèle Sebag) and David Lopez-Paz (Meta)
  Participants: Badr Youbi

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 STIC/MATH/CLIMAT AmSud projects

Green AI

Participants: Marc Schoenauer, Michèle Sebag.

Title: Towards an ecologically viable machine learning

Program: CLIMAT-AmSud

Duration: January 1, 2021 – December 31, 2022

Local supervisor: Marc Schoenauer

Partners:

- *Nayat Sanchez-Pi* (Inria Chili)
- Universidad Nacional de Asuncion

Inria contact: Marc Schoenauer

Summary: The Green AI project's main goal is to conceive a systemic and multi-component approach to the problem of the Artificial Intelligence's ecological impact. Thus, it focuses on cloud and mobile computing, transfer learning, model reuse, active learning and evolutionary computing, among others.

10.1.2 Participation in other International Programs

HFSP RGY0075/2019

Participants: Flora Jay.

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

Partner Institution(s):

- Human Frontier Science Program (funded by)
- Emilia Huerta-Sanchez (U Brown, USA), coordinator
- M Avila-Arcos (UNAM, Mexico)

Date/Duration: 2019-2024

10.2 European initiatives

10.2.1 Horizon Europe

Adra-e

Participants: Marc Schoenauer.

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 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: Artificial intelligence, data and robotics (ADR): these are three domains that are closely connected. The rise of artificial intelligence was made possible due to the availability of data, and advances in robotics made it possible to increase the number of sensors that each robot can have. In this context, the EU-funded Adra-e project will work to boost Europe's excellent research centres, innovative start-ups, a world-leading position in robotics and competitive manufacturing and services sectors. With a consortium composed of leading industry and research organisations in all three domains, the project will create the conditions for an inclusive, sustainable, effective, multilayered and coherent European ADR ecosystem. The expected result will be the increased trust and adoption of ADR.

10.2.2 H2020 projects

VISION

Participants: Marc Schoenauer.

TAILOR project on cordis.europa.eu

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

Duration: From September 1, 2020 to August 31, 2023 (+1 year extension)

Partners:

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- LINKOPINGS UNIVERSITET (LIU), Sweden
- and 51 other partners (see web site)

Inria contact: Marc Schoenauer

Coordinator: Fredrik Heintz (U. Linköpings)

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.

TAILOR

Participants: Isabelle Guyon, Marc Schoenauer.

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, 2023 (+1 year extension)

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 SRL (INTELLERA CONSULTING), Italy
- UNIVERSITY COLLEGE CORK - NATIONAL UNIVERSITY OF IRELAND, CORK (UCC), Ireland

Inria contact: Jozef Geurts

Coordinator: Holger Hoos (U. Leiden)

Summary: Artificial intelligence (AI) is an area of strategic importance and a key driver of economic development, bringing solutions to many societal challenges ranging from treating diseases to minimising the environmental impact of farming. The EU is focussing on connecting and strengthening AI research centres across Europe and supporting the development of AI applications in key sectors. To ensure Europe stays at the forefront of AI developments, the EU-funded VISION project will build on Europe's world-class community of researchers. The project will also build on the success and organisation of CLAIRE (the Confederation of Laboratories for AI Research in Europe) as well as of AI4EU, which was established to set up the first European Artificial Intelligence On-Demand Platform and Ecosystem.

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.

- **HUSH** 2020-2023 (348k euros), *HUman Supply cHain behind smart technologies.*
  Coordinator : Antonio A. Casilli (Telecom Paris)
  Participants: Paola Tubaro

- **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.

- **RoDAPoG** 2021-2025 (302k€) *Robust Deep learning for Artificial genomics and Population Genetics*
Coordinator:Flora Jay,
Participants: Cyril Furtlehner, Guillaume Charpiat.

10.3.2 Others

- **ADEME NEXT** 2017-2021, extended 2023 (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)

- **IPL HPC-BigData** 2018-2022 (100 kEuros) High Performance Computing and Big Data (Section 8.5.4)
Coordinator: Bruno Raffin (Inria Grenoble)
Participants: Guillaume Charpiat, Loris Felardos (PhD)

- **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

- **Les vraies voix de l'Intelligence Artificielle**, 2021-2023 (29k euros), funded by Maison des Sciences de l'Homme Paris-Saclay.
Coordinator : Paola Tubaro
Participants: A.A. Casilli (Telecom Paris); I. Vasilescu, L. Lamel, Gilles Adda (CNRS-LISN); J.L. Molina (UAB Barcelona); J.A. Ortega (Univ. València)

- **Inria Challenge OceanAI** 2021-2025, *AI, Data, Models for a Blue Economy*
Coordinator: Nayat Sanchez Pi (Inria Chile)
Participants: Marc Schoenauer, Michèle Sebag and Shiyang Yan

10.4 Regional initiatives

- **DATAIA ML4CFD** 2020-2022 (105 kEuros) *Machine Learning for Computational Fluid Dynamics.*
Coordinator: Michele Alessandro Bucci
Participants: Guillaume Charpiat, Marc Schoenauer
Collaboration: IFPEN (Jean-Marc Gratien and Thibault Faney)

- **DATAIA YARN** 2022-2025 (240 kEuros) *Automatic Processing of Messy Brain Data with Robust Methods and Transfer Learning.*
Coordinator: Sylvain Chevallier, Florent Bouchard (L2S)
Collaboration: Raymond Poincare' Hospital (France), FCAI (Aalto University, Finland), Fre'de'ric Pascal (L2S), Alexandre Gramfort (Meta)

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

Member of the organizing committees

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

- Cecile Germain - Steering committee of the Learning to Discover program of Institut Pascal (originally 2020, postponed to 2022)
- Flora Jay - Organizer of Thematic School “Graph as models in life sciences: Machine learning and integrative approaches” (supported by Digicosme)

11.1.2 Scientific events: selection

Chair of conference program committees

- Marc Schoenauer, Senior Area Chair, IJCAI 2022
- Michele Sebag, Senior Area Chair ICML 2022

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); series editor, Springer series Challenges in Machine Learning (CiML).
- Marc Schoenauer - Advisory Board, Evolutionary Computation Journal, MIT Press, and Genetic Programming and Evolutionary Machines, Springer Verlag; Action editor, Journal of Machine Learning Research (JMLR); Editorial Board, ACM Transaction on Evolutionary Learning and Optimization (TELO).
- Michèle Sebag - Editorial Board, Machine Learning, Springer Verlag; ACM Transactions on Evolutionary Learning and Optimization.
- Paola Tubaro: Sociology, Revue française de sociologie, Journal of Economic Methodology, Lecturas de Economia.

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, *Apprentissage profond pour le recalage, la segmentation et la polygonisation d’images satellitaires*, lors du "Séminaire IA: imagerie aérienne et report au plan" de la DGFIP (Bercy) le 29 juin.
- Guillaume Charpiat, invited keynote and co-chair of the session "Trustworthy and Explainable AI" of the day "Franco-German Research and Innovation Network on AI", Rocquencourt, June 14th.
- Flora Jay,
- Marc Schoenauer,
- Michele Sebag, keynote speaker IDA 2022; keynote MIT-France Symposium; keynote French-German symposium;
- Paola Tubaro,
- François Landes, *SE(3)-Equivariant GNNs for Machine Learning Glasses*, at the CNRS+CFM workshop "Machine Learning Glasses", Paris, November 11, 2022
- Isabelle Guyon, NeurIPS’22 keynote The Data-Centric Era: How ML is Becoming an Experimental Science.

11.1.5 Leadership within the scientific community

- Guillaume Charpiat: creation and co-animation of 2 DigiCosme working groups on the Saclay plateau and beyond: vrAI (verification and robustness of AI) and SNAP (simulations numériques et apprentissage)
- Isabelle Guyon: Member of the board, NeurIPS; Member of the Board, JEDI, Joint European Disruptive Initiative; President and co-founder, ChaLearn, non-for-profit organization dedicated to the organization of challenges.
- 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: Executive Committee, Institut de Convergence DataIA; Member of IRSN Scientific Council; Member of scientific council of the AMIES Labex;

11.1.6 Scientific expertise

- Guillaume Charpiat: member of the Commission Scientifique (CS) at INRIA Saclay (PhD/post-docs grant allocations)
- Guillaume Charpiat: Jean Zay (GENCI/IDRIS) committee member for resource allocation (GPU) demand expertise
- Guillaume Charpiat: PhD grant allocation: UdopIA jury + DigiCosme Labex jury
- Guillaume Charpiat: jury of "Prix Doctorants STIC du Plateau de Saclay"
- Flora Jay, CR hiring committee, INRAE Toulouse
- Flora Jay, MdC hiring committee, LIX
- Marc Schoenauer, Scientific Advisory Board, BCAM, Bilbao, Spain
- Marc Schoenauer, "Conseil Scientifique", IFPEN
- Marc Schoenauer, "Conseil Scientifique", Mines Paritech
- Marc Schoenauer, "Commission Recherche", Université Paris-Diderot
- Michele Sebag, UDOPIA jury (PhDs)
- Michele Sebag, FNRS (PhDs and Post-docs)
- Michele Sebag, professorship hiring committee, INSA Rouen
- Paola Tubaro, MdC hiring committee, University of Lille
- Paola Tubaro, professorship hiring committee, Sorbonne Université
- Paola Tubaro, associate professorship hiring committee, University of Greenwich (UK)
- Paola Tubaro, assistant professorship hiring committee, University of Insubria (IT)
- Isabelle Guyon, Advisory board Kaggle competitons.

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
- Paola Tubaro, member of Local Committee of Institut Pascal, Université Paris-Saclay

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, 51h, L2, Univ. Paris-Sud.
- Licence : François Landes, Introduction to Statistical Learning, 88h, 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-Sud.
- Master : François Landes, A first look inside the ML black box, 25h, M1 Recherche (AI track), U. Paris-Sud.
- Master : François Landes, Machine Learning, 28h, M2 Univ. Paris-sud, physics department (PCS international Master)
- Master : Guillaume Charpiat, Deep Learning in Practice, 21h, M2 Recherche, MVA / Centrale-Supelec / DSBA.
- Master : Guillaume Charpiat, Information Theory, 14h, M1 IA Paris-Sud.
- Master : Guillaume Charpiat, Introduction to Deep Learning, 1h30, Eugloh.
- Master : Guillaume Charpiat, Deep Learning for Physics, 3h, M2 IASD, Paris-Dauphine university
- 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 : Paola Tubaro, Sociology of social networks, 24h, M2, EHESS/ENS.
- Master : Paola Tubaro, Social and economic network science, 24h, M2, ENSAE.
- Master: Paola Tubaro, Ethics of social and digital data, 12h, Université de Toulouse Jean Jaurès
- 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.
- Fall school : Flora Jay, Inference using full genome data, 7h, TUM, Germany.

11.2.2 Supervision

- PhD - Leonard BLIER, *Some Principled Methods for Deep Reinforcement Learning* [45], 28/04/22, Yann Ollivier (Facebook AI Research, Paris) and Marc Schoenauer
- PhD - R. BRESSON, *Neural learning and validation of hierarchical multi-criteria decision aiding models with interacting criteria* [46], 2/02/22, Johanne Cohen (Galac, LISN) and Michèle Sebag
- PhD - Balthazar DONON, *Deep Statistical Solvers and Power Systems Applications* [47], 16/03/22, Isabelle Guyon, Marc Schoenauer, and Rémy Clément (RTE)
- PhD - Loris FELARDOS, *Neural networks for molecular dynamics simulations* [48], 2/12/22, Guillaume Charpiat, Jérôme Hénin (IBPC) and Bruno Raffin (InriAlpes)

- PhD - Giancarlo FISSORE, *Statistical physics analysis of generative models* [49], 9/03/22, Aurélien Decelle and Cyril Furtlehner
- PhD - Herilalaina RAKOTOARISON, *Automatic Algorithm Configuration for Power Grid Optimization* [50], 27/06/22, Marc Schoenauer and Michèle Sebag
- PhD - Théophile SANCHEZ, *Reconstructing the past: deep learning for population genetics* [51], 18/03/22, Guillaume Charpiat and Flora Jay
- PhD - Marion ULLMO, *Emulation and prediction of cosmic web simulations through deep learning* [52], from 1/02/22, Nabila Aghanim (Institut d'Astrophysique Spatiale) and Aurélien Decelle
- 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 - 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 - Wenzhuo LIU, *Graph Neural Networks for Numerical Simulation of PDEs*, from 1/11/2019, Mouadh Yagoubi (IRT SystemX) and Marc Schoenauer
- 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, *Complementary Deep Reduced Order Model*, from 1/9/2020, Michele Alessandro Bucci and Marc Schoenauer
- 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, *Machine Learning enhanced resolution of Navier-Stokes equations on general unstructured grids*, 4/1/2021, Guillaume Charpiat and Michele Alessandro Bucci.
- PhD in progress - Adrien PAVAO, *Theory and practice of challenge organization*, from 1/03/2020, Isabelle Guyon.
- 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
- PhD in progress - Vincenzo SCHIMMENTI, *Eartquake Predictions: Machine Learned Features using Expert Models Simulations*, from 1/11/2020, François Landes and Alberto Rosso (LPTMS)
- 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 - Manon VERBOCKHAVEN, *Spotting and fixing expressivity bottlenecks*, from 11/2021, Sylvain Chevallier and Guillaume Charpiat
- PhD in progress - Assia WIRTH, *Coloniality of the production of facial recognition technologies*, started 01/04/2021, Paola Tubaro
- 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)

11.2.3 Juries

- Marc Schoenauer, PhD jury member, Paul Defossé, 22/12/22, CMAP Ecole Polytechnique, Palaiseau; PhD jury member, Ekhi Ajuria, 5/12/2022, Cerfacs, Toulouse; PhD Committee (2nd year), Kaitlin Maile, 16/12/2022, IRIT, Toulouse.
- Guillaume Charpiat, PhD jury member, Arthur Ouaknine, 04/03/22, Telecom, IPP, Palaiseau
- Michèle Sebag: HdR jury member: Ievgen Redko; Thomas Lampert; Raphael Féraud; PhD jury member: Teo Sanchez; Naoufal Acharki; Marwa Kechaou; Fatoumata Dama

11.3 Popularization

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 CEA and CNRS)
- Marc Schoenauer, scientific coordinator of ICT49 CSA Adra-e (coordinated by Inria)

11.3.2 Interventions

- Michèle Sebag; France Culture; RadioLibertaire; Exposé invité Sinclair Lab; Exposé invité, Labex Nano-Saclay; participation table ronde, Expertise sur la régulation numérique, ministère des Finances.

12 Scientific production

12.1 Major publications

[1] 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 (Dec. 2015). DOI: 10.1088/1742-6596/664/7/072015. URL: https://hal.inria.fr/hal-01745998.

[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] L. Da Costa, Á. Fialho, M. Schoenauer and M. Sebag. 'Adaptive Operator Selection with Dynamic Multi-Armed Bandits'. In: *Proc. Genetic and Evolutionary Computation Conference (GECCO)*. ACM-SIGEVO 10-years Impact Award. ACM. 2008, pp. 913–920. DOI: 10.1145/1389095.1389272. URL: https://hal.inria.fr/inria-00278542.

[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] C. Furtlehner and A. Decelle. 'Cycle-based Cluster Variational Method for Direct and Inverse Inference'. In: *Journal of Statistical Physics* 164.3 (Aug. 2016), pp. 531–574. URL: https://hal.inria.fr/hal-01214155.

[6] S. Gelly, M. Schoenauer, M. Sebag, O. Teytaud, L. Kocsis, D. Silver and C. Szepesvari. 'The Grand Challenge of Computer Go: Monte Carlo Tree Search and Extensions'. In: *Communications- ACM* 55.3 (2012), pp. 106–113. URL: https://hal.inria.fr/hal-00695370.

[7] O. Goudet, D. Kalainathan, P. Caillou, D. Lopez-Paz, I. Guyon and M. Sebag. 'Learning Functional Causal Models with Generative Neural Networks'. In: *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer Series on Challenges in Machine Learning. https://arxiv.org/abs/1709.05321. Springer International Publishing, 2018. DOI: 10.1007/978-3-319-98131-4. URL: https://hal.archives-ouvertes.fr/hal-01649153.

[8] M. Mısır and M. Sebag. 'Alors: An algorithm recommender system'. In: *Artificial Intelligence* 244 (2017). Published on-line Dec. 2016, pp. 291–314. URL: https://hal.inria.fr/hal-01419874.

[9] Y. Ollivier, L. Arnold, A. Auger and N. Hansen. 'Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles'. In: *Journal of Machine Learning Research* 18.18 (2017), pp. 1–65. URL: https://hal.inria.fr/hal-01515898.

[10] 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

[11] N. Béreux, A. Decelle, C. Furtlehner and B. Seoane. 'Learning a Restricted Boltzmann Machine using biased Monte Carlo sampling'. In: *SciPost Physics* (2022). URL: https://hal.inria.fr/hal-03795598.

[12] T. Bonnaire, N. Aghanim, J. Kuruvilla and A. Decelle. 'Cosmology with cosmic web environments I. Real-space power spectra'. In: *Astronomy and Astrophysics - A&A* 661.A146 (24th May 2022). DOI: 10.1051/0004-6361/202142852. URL: https://hal.science/hal-03477764.

[13] T. Bonnaire, N. Aghanim, J. Kuruvilla and A. Decelle. 'Cosmology with cosmic web environments: I. Real-space power spectra'. In: *Astronomy and Astrophysics - A&A* 661 (May 2022), A146. DOI: 10.1051/0004-6361/202142852. URL: https://hal.science/hal-03677762.

[14] J. Cury, B. C. Haller, G. Achaz and F. Jay. 'Simulation of bacterial populations with SLiM'. In: *Peer Community Journal* (13th Jan. 2022). DOI: 10.24072/pcjournal.72. URL: https://hal.science/hal-03152153.

[15] A. Decelle. 'Fundamental problems in Statistical Physics XV: Lecture on Machine Learning'. In: *Physica A: Statistical Mechanics and its Applications* (5th Sept. 2022), p. 128154. DOI: 10.1016/j.physa.2022.128154. URL: https://hal.inria.fr/hal-03795587.

[16] C. Furtlehner, J.-M. Lasgouttes, A. Attanasi, M. Pezzulla and G. Gentile. ‘Short-term Forecasting of Urban Traffic using Spatio-Temporal Markov Field’. In: *IEEE Transactions on Intelligent Transportation Systems* 23.8 (2022), pp. 10858–10867. DOI: 10.1109/TITS.2021.3096798. URL: https://hal.inria.fr/hal-03285664.

[17] S. Hartley, G. Bao, M. Zagdoun, S. Chevallier, F. Lofaso, A. Leotard and E. Azabou. ‘Noninvasive Vagus Nerve Stimulation: A New Therapeutic Approach for Pharmacoresistant Restless Legs Syndrome’. In: *Neuromodulation* (Nov. 2022). DOI: 10.1016/j.neurom.2022.10.046. URL: https://hal-universite-paris-saclay.archives-ouvertes.fr/hal-03921967.

[18] 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.

[19] L. Meunier, H. Rakotoarison, P. K. Wong, B. Roziere, J. Rapin, O. Teytaud, A. Moreau and C. Doerr. ‘Black-Box Optimization Revisited: Improving Algorithm Selection Wizards through Massive Benchmarking’. In: *IEEE Transactions on Evolutionary Computation* 26.3 (2022). DOI: 10.1109/TEVC.2021.3108185. URL: https://hal.inria.fr/hal-03154019.

[20] P. Tubaro. ‘Learners in the loop: hidden human skills in machine intelligence’. In: *Sociologia del Lavoro* 163 (2022), pp. 110–129. DOI: 10.3280/SL2022-163006. URL: https://hal.inria.fr/hal-03787017.

[21] P. Tubaro and A. A. Casilli. ‘Who bears the burden of a pandemic? COVID-19 and the transfer of risk to digital platform workers’. In: *American Behavioral Scientist* (15th Jan. 2022). DOI: 10.1177/00027642211066027. URL: https://hal.science/hal-03369291.

[22] P. Tubaro, M. Coville, C. Le Ludec and A. A. Casilli. ‘Hidden inequalities: the gendered labour of women on micro-tasking platforms’. In: *Internet Policy Review*. The gender of the platform economy 11.1 (2022). DOI: 10.14763/2022.1.1623. URL: https://hal.inria.fr/hal-03551747.

[23] 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* (8th July 2022). URL: https://hal.science/hal-03374222.

National journals

[24] C. Icubam, L. Bonnasse-Gahot, M. Dénès, G. Dulac-Arnold, S. Girgin, F. Husson, V. Iovene, J. Josse, A. Kimmoun, F. Landes, J.-P. Nadal, R. Primet, F. Quintao, P. G. Raverdy, V. Rouvreau, O. Teboul and R. Yurchak. ‘ICU Bed Availability Monitoring and analysis in the Grand Est region of France during the COVID-19 epidemic’. In: *Statistique et Société* (17th Mar. 2022). DOI: 10.1101/2020.05.18.20091264. URL: https://hal.science/hal-02620018.

International peer-reviewed conferences

[25] 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, 19th Sept. 2022. URL: https://hal.inria.fr/hal-03798964.

[26] M. Dorier, R. Egele, P. Balaprakash, J. Koo, S. Madireddy, S. Ramesh, A. Malony and R. Ross. ‘HPC Storage Service Autotuning Using Variational- Autoencoder -Guided Asynchronous Bayesian Optimization’. In: CLUSTER 2022 - IEEE International Conference on Cluster Computing (CLUSTER). Heidelberg, Germany: IEEE, 6th Sept. 2022, pp. 381–393. DOI: 10.1109/CLUSTER51413.2022.00049. URL: https://hal.science/hal-03864478.

[27] 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, 21st Aug. 2022, pp. 1908–1914. URL: https://hal.science/hal-03518597.

[28] M. H. Nguyen, L. Sun-Hosoya, N. Grinsztajn and I. Guyon. 'Meta-learning from Learning Curves: Challenge Design and Baseline Results'. In: IJCNN 2022 - International Joint Conference on Neural Networks. Padua, Italy: IEEE, 18th July 2022, pp. 1–8. URL: https://hal.science/hal-03740118.

[29] 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, 5th Oct. 2022. URL: https://hal.inria.fr/hal-03869648.

[30] J. Pedersen, R. Muñoz Gómez, J. Huang, H. Sun, W.-W. Tu and I. Guyon. 'LTU Attacker for Membership Inference'. In: Third AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI-22). virtual, France, 2022. URL: https://hal.science/hal-03522633.

[31] 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.

[32] 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. Singapour, Singapore, 4th Dec. 2022. URL: https://hal.science/hal-03726294.

[33] M. Videau, A. Leite, O. Teytaud and M. Schoenauer. 'Multi-Objective Genetic Programming for Explainable Reinforcement Learning'. In: Proc. EuroGP 2022. Vol. 13223. Lecture Notes in Computer Science. Madrid, France: Springer International Publishing, 13th Apr. 2022, pp. 278–293. DOI: 10.1007/978-3-031-02056-8_18. URL: https://hal.inria.fr/hal-03886307.

[34] G. Zervakis, E. Vincent, M. Couceiro, M. Schoenauer and E. Marquer. 'An analogy based approach for solving target sense verification'. In: NLPIR 2022 - 6th International Conference on Natural Language Processing and Information Retrieval. Bangkok, Thailand, 16th Dec. 2022. URL: https://hal.inria.fr/hal-03792071.

Conferences without proceedings

[35] G. Bied, S. Nathan, E. Pérennes, V. A. Naya, P. Caillou, B. Crépon, C. Gaillac and M. Sebag. 'Recommender system in a non-stationary context: recommending job ads in pandemic times'. In: FEAST workshop ECML-PKDD 2022. Grenoble, France, 19th Sept. 2022. URL: https://hal.inria.fr/hal-03831247.

[36] D. Carrión-Ojeda, H. Chen, A. El Baz, S. Escalera, C. Guan, I. Guyon, I. Ullah, X. Wang and W. Zhu. 'NeurIPS'22 Cross-Domain MetaDL competition: Design and baseline results'. In: Meta-Knowledge Transfer/Communication in Different Systems. Grenoble, France, 23rd Sept. 2022. URL: https://hal.science/hal-03765151.

[37] 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, 19th Nov. 2022. URL: https://hal.science/hal-03878200.

[38] 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, 3rd Dec. 2022. URL: https://hal.inria.fr/hal-03864015.

[39] I. Ullah, D. Carrión-Ojeda, S. Escalera, I. Guyon, M. Huisman, F. Mohr, J. N. van Rijn, H. Sun, J. Vanschoren and P. A. Vu. 'Meta-Album: Multi-domain Meta-Dataset for Few-Shot Image Classification'. In: 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks. New Orleans, United States, 2022. URL: https://hal.inria.fr/hal-03991982.

Scientific book chapters

[40] 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)*. 19th Mar. 2022. DOI: 10.1007/978-3-642-27737-5_248-2. URL: https://hal.inria.fr/hal-02942375.

[41] E. Aurell, J. Barbier, A. Decelle and R. Mulet. 'The mighty force: statistical inference and high-dimensional statistics'. In: *Spin Glass Theory & Far Beyond - Replica Symmetry Breaking after 40 Years*. World Scientific, 2022. DOI: 10.48550/arXiv.2205.00750. URL: https://hal.inria.fr/hal-03795572.

[42] A. A. Pol, G. Cerminara, C. Germain and M. Pierini. 'Data Quality Monitoring Anomaly Detection'. In: *Artificial Intelligence for High Energy Physics*. World Scientific, Mar. 2022. DOI: 10.1142/9789811234033_0004. URL: https://hal.inria.fr/hal-03159873.

[43] P. Tubaro. 'Social networks and resilience in emerging labor markets'. In: *A Research Agenda for Social Networks and Social Resilience*. Edward Elgar, Nov. 2022, pp. 45–57. URL: https://hal.science/hal-03850444.

[44] P. Tubaro and A. A. Casilli. 'Human Listeners and Virtual Assistants: Privacy and Labor Arbitrage in the Production of Smart Technologies'. In: *Digital Work in the Planetary Market*. The MIT Press, 2022. DOI: 10.7551/mitpress/13835.003.0014. URL: https://hal.inria.fr/hal-03688430.

Doctoral dissertations and habilitation theses

[45] L. Blier. 'Some Principled Methods for Deep Reinforcement Learning'. Université Paris-Saclay, 28th Apr. 2022. URL: https://theses.hal.science/tel-03829500.

[46] R. Bresson. 'Neural learning and validation of hierarchical multi-criteria decision aiding models with interacting criteria'. Université Paris-Saclay, 2nd Feb. 2022. URL: https://theses.hal.science/tel-03596964.

[47] B. Donon. 'Deep statistical solvers & power systems applications'. Université Paris-Saclay, 16th Mar. 2022. URL: https://theses.hal.science/tel-03624628.

[48] L. Felardos. 'Data-free Generation of Molecular Configurations with Normalizing Flows'. Université Grenoble-Alpes, Dec. 2022.

[49] G. Fissore. 'Generative modeling : statistical physics of Restricted Boltzmann Machines, learning with missing information and scalable training of Linear Flows'. Université Paris-Saclay, 9th Mar. 2022. URL: https://theses.hal.science/tel-03710286.

[50] H. Rakotoarison. 'Some contributions to AutoML : hyper-parameter optimization and meta-learning'. Université Paris-Saclay, 27th June 2022. URL: https://theses.hal.science/tel-03783610.

[51] T. Sanchez. 'Reconstructing our past : deep learning for population genetics'. Université Paris-Saclay, 18th Mar. 2022. URL: https://theses.hal.science/tel-03701132.

[52] M. Ullmo. 'Emulation and prediction of cosmic web simulations through deep learning'. Université Paris-Saclay, 1st Feb. 2022. URL: https://theses.hal.science/tel-03663099.

Reports & preprints

[53] A. Barra, G. Catania, A. Decelle and B. Seoane. *Thermodynamics of bidirectional associative memories*. 18th Nov. 2022. URL: https://hal.science/hal-03859276.

[54] M. A. Bucci, O. Semeraro, A. Allauzen, S. Chibbaro and L. Mathelin. *Leveraging the structure of dynamical systems for data-driven modeling*. 9th Nov. 2022. URL: https://hal.science/hal-03846125.

[55] O. Chagny, A. A. Casilli, D. Chartier, T. d'Avezac, F. Pailler, P. Rothey, P. Tubaro, M. Cornet, M. Joulin and S. Kagambega. *Les Nouveaux Intermédiaires du Travail B2B: Comparer les modèles d'affaires dans l'économie numérique collaborative*. 27. DARES - Direction de l'animation de la recherche, des études et des statistiques du Ministère du travail, de l'emploi et de l'insertion, 15th Mar. 2022. URL: https://hal.inria.fr/hal-03615806.

[56] S. Dong, K. Uemura, A. Fujii, S. Chang, Y. Koyanagi, K. Maruhashi and M. Sebag. *High-Dimensional Causal Discovery: Learning from Inverse Covariance via Independence-based Decomposition*. 6th Dec. 2022. URL: https://hal.inria.fr/hal-03885791.

[57] R. Egele, J. Gouneau, V. Vishwanath, I. Guyon and P. Balaprakash. *Asynchronous Distributed Bayesian Optimization at HPC Scale*. 21st Nov. 2022. URL: https://hal.science/hal-03864504.

[58] 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.

[59] C. Furtlehner. *Free Dynamics of Feature Learning Processes*. 19th Oct. 2022. URL: https://hal.inria.fr/hal-03878500.

[60] 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*. 2022. DOI: 10.1101/2022.05.25.493366. URL: https://hal.science/hal-03875721.

[61] E. Menier, M. A. Bucci, M. Yagoubi, L. Mathelin, T. Dairay, R. Meunier and M. Schoenauer. *Continuous Methods : Adaptively intrusive reduced order model closure*. 23rd July 2022. URL: https://hal.inria.fr/hal-03879332.

[62] E. Menier, M. A. Bucci, M. Yagoubi, L. Mathelin and M. Schoenauer. *CD-ROM: Complementary Deep-Reduced Order Model*. 9th Nov. 2022. URL: https://hal.science/hal-03846122.

[63] T. Nakano, M. A. Bucci, J.-M. Gratien, T. Faney and G. Charpiat. *Machine learning model for gas-liquid interface reconstruction in CFD numerical simulations*. 12th July 2022. URL: https://hal.science/hal-03721729.

[64] M. H. Nguyen, L. Sun, N. Grinsztajn and I. Guyon. *Meta-learning from Learning Curves Challenge: Lessons learned from the First Round and Design of the Second Round*. 3rd Aug. 2022. URL: https://hal.science/hal-03725313.

[65] 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*. Université Paris-Saclay, FRA., 4th Apr. 2022. URL: https://hal.inria.fr/hal-03629462.

[66] F. S. Pezzicoli, G. Charpiat and F. P. Landes. *SE(3)-equivariant Graph Neural Networks for Learning Glassy Liquids Representations*. 2022. URL: https://hal.inria.fr/hal-03868206.

[67] Z. Xu, L. Wei, H. Zhao, R. Ying, Q. Yao, W.-W. Tu and I. Guyon. *BRIDGING THE GAP OF AUTOGRAPH BETWEEN ACADEMIA AND INDUSTRY: ANALYSING AUTOGRAPH CHALLENGE AT KDD CUP 2020*. 4th Apr. 2022. URL: https://hal.science/hal-03629994.

Other scientific publications

[68] I. Hoxha, C. Bäker, S. Chevallier, S. Glasauer and M.-A. Amorim. 'Generating stimulus anticipation from stimulus and prediction history'. In: BC 2022 - Bernstein Conference. Berlin, Germany, 13th Sept. 2022. DOI: 10.12751/nncn.bc2022.049. URL: https://hal.science/hal-03786876.

12.3 Cited publications

[69] 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.

[70] 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.

[71] A. Decelle, T. Bonnaire and N. Aghanim. 'Cascade of phase transitions for multiscale clustering'. In: *Physical Review E* 103.1 (Jan. 2021). DOI: 10.1103/PhysRevE.103.012105. URL: https://hal.archives-ouvertes.fr/hal-03477663.

[72] A. Decelle, T. Bonnaire and N. Aghanim. 'Regularization of Mixture Models for Robust Principal Graph Learning'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (13th Nov. 2021), pp. 1–1. DOI: 10.1109/TPAMI.2021.3124973. URL: https://hal.archives-ouvertes.fr/hal-03477742.

[73] R. Kusters, D. Misevic, H. Berry, A. Cully, Y. Le Cunff, L. Dandoy, N. Díaz-Rodríguez, M. Ficher, J. Grizou, A. Othmani, T. Palpanas, M. Komorowski, P. Loiseau, C. Moulin-Frier, S. Nanini, D. Quercia, M. Sebag, F. Soulié Fogelman, S. Taleb, L. Tupikina, V. Sahu, J.-J. Vie and F. Wehbi. 'Interdisciplinary Research in Artificial Intelligence: Challenges and Opportunities'. In: *Frontiers in Big Data* 3 (23rd Nov. 2020). DOI: 10.3389/fdata.2020.577974. URL: https://hal.inria.fr/hal-03111148.

[74] F. P. Landes, G. Biroli, O. Dauchot, A. Liu and D. Reichman. 'Attractive versus truncated repulsive supercooled liquids: The dynamics is encoded in the pair correlation function'. In: *Physical Review E* 101.1 (13th Jan. 2020). DOI: 10.1103/PhysRevE.101.010602. URL: https://hal.inria.fr/hal-02439295.

[75] G. Petrillo, E. Lippiello, F. P. Landes and A. Rosso. 'The influence of the brittle-ductile transition zone on aftershock and foreshock occurrence'. In: *Nature Communications* 11.1 (15th June 2020). DOI: 10.1038/s41467-020-16811-7. URL: https://hal.inria.fr/hal-02942365.

[76] 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.

[77] R. Egele, P. Balaprakash, V. Vishwanath, I. Guyon and Z. Liu. 'AgEBO-Tabular: Joint Neural Architecture and Hyperparameter Search with Autotuned Data-Parallel Training for Tabular Data'. In: SC '21: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. St. Louis, Missouri, United States, Nov. 2021. DOI: 10.1145/3458817.3476203. URL: https://hal.archives-ouvertes.fr/hal-02973288.

[78] 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.

[79] A. Pavao, I. Guyon, N. Stéphane, F. Lebeau, M. Ghienne, L. Platon, T. Barbagelata, P. Escamilla, S. Mzali, M. Liao, S. Lassonde, A. Braun, S. B. Amor, L. Cucu-Grosjean, M. Wehaiba, A. Bar-Hen, A. Gogonel, A. B. Cheikh, M. Duda, J. Laugel, M. Marauri, M. Souissi, T. Lecerf, M. Elion, S. Tabti, J. Budynek, P. Le Bouteiller, A. Penon, R.-D. Lasseri, J. Ripoche and T. Epalle. 'Aircraft Numerical "Twin": A Time Series Regression Competition'. In: ICMLA 2021 - 20th IEEE International Conference on Machine Learning and Applications. Pasadena / Virtual, United States, 13th Dec. 2021. URL: https://hal.inria.fr/hal-03463307.

[80] A. Pavao, M. Vaccaro and I. Guyon. 'Judging competitions and benchmarks: a candidate election approach'. In: ESANN 2021 - 29th European Symposium on Artificial Neural Networks. Bruges/Virtual, Belgium, 6th Oct. 2021. URL: https://hal.inria.fr/hal-03367857.

[81] G. Zervakis, E. Vincent, M. Couceiro and M. Schoenauer. 'On Refining BERT Contextualized Embeddings using Semantic Lexicons'. In: ECML PKDD 2021 - Machine Learning with Symbolic Methods and Knowledge Graphs co-located with European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. http://ceur-ws.org/Vol-2997/paper4.pdf. Online, Spain, 1st Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03318571.

[82] A. Lacombe, S. Jetley and M. Sebag. 'EXtremely PRIvate supervised Learning'. In: Conférence d'APprentssage - CAP. St-Etienne, France, 2021. URL: https://hal.inria.fr/hal-03620873.

[83] G. Bied, E. Perennes, V. A. Naya, P. Caillou, B. Crépon, C. Gaillac and M. Sebag. 'Congestion-Avoiding Job Recommendation with Optimal Transport'. In: FEAST workshop ECML-PKDD 2021. Bilbao, Spain, 17th Sept. 2021. URL: https://hal.inria.fr/hal-03540316.

[84] A. Lacan and I. Guyon. 'ML-CI: Machine Learning Confidence Intervals for Covid-19 forecasts'. In: BayLearn - Machine Learning Symposium 2021. San Francisco, United States, 28th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03501101.

[85] V. A. Naya, G. Bied, P. Caillou, B. Crépon, C. Gaillac, E. Pérennes and M. Sebag. 'Designing labor market recommender systems: the importance of job seeker preferences and competition'. In: 4. IDSC of IZA Workshop: Matching Workers and Jobs Online - New Developments and Opportunities for Social Science and Practice. Online, France, 8th Oct. 2021. URL: https://hal.inria.fr/hal-03540319.

[86] 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.

[87] L. Blier, C. Tallec and Y. Ollivier. *Learning Successor States and Goal-Dependent Values: A Mathematical Viewpoint*. 19th Jan. 2021. URL: https://hal.inria.fr/hal-03151901.

[88] T. Sanchez, E. Madison Bray, P. Jobic, J. Guez, A.-C. Letournel, G. Charpiat, J. Cury and F. Jay. *dnadna: DEEP NEURAL ARCHITECTURES FOR DNA - A DEEP LEARNING FRAMEWORK FOR POPULATION GENETIC INFERENCE*. 17th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03352910.

[89] 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.

[90] D. J. Amit, H. Gutfreund and H. Sompolinsky. 'Statistical Mechanics of Neural Networks Near Saturation'. In: *Annals of Physics* 173 (1987), pp. 30–67.

[91] C. Anderson. *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. Wired Magazine. 2008. URL: https://www.wired.com/2008/06/pb-theory/.

[92] W. Banzhaf, F. D. Francone, R. E. Keller and P. Nordin. *Genetic Programming: An Introduction: On the Automatic Evolution of Computer Programs and Its Applications*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998.

[93] N. Belkhir. 'Per Instance Algorithm Configuration for Continuous Black Box Optimization'. Theses. Université Paris-Saclay, Nov. 2017.

[94] I. Bello, B. Zoph, V. Vasudevan and Q. V. Le. 'Neural Optimizer Search with Reinforcement Learning'. In: *34th ICML*. 2017, pp. 459–468.

[95] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira and J. W. Vaughan. 'A theory of learning from different domains'. In: *Machine Learning* 79.1 (2018), pp. 151–175. DOI: 10.1007/s10994-009-5152-4.

[96] J. Bergstra, R. Bardenet, Y. Bengio and B. Kégl. 'Algorithms for Hyper-Parameter Optimization'. In: *NIPS 25*. 2011, pp. 2546–2554.

[97] T. Bonnaire. 'Cosmic web environments : identification, characterisation, and quantification of cosmological information'. Theses. Université Paris-Saclay, Sept. 2021. URL: https://tel.archives-ouvertes.fr/tel-03389995.

[98] T. Bonnaire, N. Aghanim, A. Decelle and M. Douspis. 'T-ReX: a graph-based filament detection method'. In: *Astronomy and Astrophysics - A&A* 637 (May 2020), A18. DOI: 10.1051/0004-6361/201936859. URL: https://hal.archives-ouvertes.fr/hal-02903856.

[99] J. Bruna and S. Mallat. 'Invariant Scattering Convolution Networks'. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 35.8 (2013), pp. 1872–1886.

[100] M. Chandorkar. 'Machine Learning in Space Weather'. Theses. Université of Eindhoven, Nov. 2019. URL: https://hal.inria.fr/tel-02430788.

[101] R. Couillet and M. Debbah. *Random matrix methods for wireless communications*. Cambridge University Press, 2011.

[102] David Silver et al. 'Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm'. In: *CoRR* abs/1712.01815 (2017). URL: http://arxiv.org/abs/1712.01815.

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

[104] 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.

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

[106] 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.

[107] 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.

[108] G. F. Doquet and M. Sebag. 'Agnostic feature selection'. In: *ECML PKDD 2019 - European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Würzburg, Germany, Sept. 2019. URL: https://hal.archives-ouvertes.fr/hal-02436824.

[109] T. Duriez, S. L. Brunton and B. R. Noack. *Machine Learning Control – Taming Nonlinear Dynamics and Turbulence*. Springer International Publishing, 2017.

[110] T. v. Erven, P. Grunwald and S. de Rooij. 'Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC-BIC dilemma: Catching up Faster'. In: *J. Royal Statistical Society: B* 74.3 (2012), pp. 361–417. DOI: 10.1111/j.1467-9868.2011.01025.x. (Visited on 08/11/2018).

[111] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum and F. Hutter. 'Efficient and Robust Automated Machine Learning'. In: *NIPS 28*. 2015, pp. 2962–2970.

[112] R. Ffrancon and M. Schoenauer. 'Memetic Semantic Genetic Programming'. In: *Genetic and Evolutionary Computation COnference (GECCO 2015)*. Ed. by A. Esparcia and S. Silva. Best Paper Award in the GP track - see http://www.sigevo.org/gecco-2015/papers.html. Madrid, Spain: ACM, July 2015, pp. 1023–1030. URL: https://hal.inria.fr/hal-01169074.

[113] C. Furtlehner and A. Decelle. 'Cycle-Based Cluster Variational Method for Direct and Inverse Inference'. In: *Journal of Statistical Physics* 164.3 (2016), pp. 531–574.

[114] C. Furtlehner, M. Sebag and X. Zhang. 'Scaling analysis of affinity propagation'. In: *Physical Review E* 81.6 (2010), p. 066102.

[115] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand and V. Lempitsky. 'Domain-Adversarial Training of Neural Networks'. In: *Journal of Machine Learning Research* 17.59 (2016), pp. 1–35.

[116] K. Garimella, G. D. F. Morales, A. Gionis and M. Mathioudakis. 'Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship'. In: *WWW*. ACM, 2018, pp. 913–922.

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

[118] F. Gonard. 'Cold-start recommendation : from Algorithm Portfolios to Job Applicant Matching'. Theses. Université Paris-Saclay, May 2018. URL: https://tel.archives-ouvertes.fr/tel-01825220.

[119] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio. 'Generative Adversarial Nets'. In: *NIPS 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger. Curran Associates, Inc., 2014, pp. 2672–2680.

[120] P. Grunwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[121] I. Guyon, C. F. Aliferis, G. F. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes and A. R. Statnikov. 'Design and Analysis of the Causation and Prediction Challenge'. In: *WCCI Causation and Prediction Challenge*. JMLR W-CP. 2008, pp. 1–33.

[122] I. Guyon, K. Bennett, G. Cawley, H. J. Escalante, S. Escalera, T. K. Ho, N. Macia, B. Ray, M. Saeed and A. Statnikov. 'Design of the 2015 ChaLearn AutoML challenge'. In: *Proc. IJCNN*. IEEE. 2015, pp. 1–8.

[123] I. Guyon, L. Sun-Hosoya, M. Boullé, H. J. Escalante, S. Escalera, Z. Liu, D. Jajetic, B. Ray, M. Saeed, M. Sebag, A. Statnikov, W.-W. Tu and E. Viegas. 'Analysis of the AutoML Challenge series 2015-2018'. In: *AutoML: Methods, Systems, Challenges*. Ed. by F. Hutter, L. Kotthoff and J. Vanschore. The Springer Series on Challenges in Machine Learning. Springer Verlag, 2018. URL: https://hal.archives-ouvertes.fr/hal-01906197.

[124] H. H. Hoos. 'Programming by Optimization'. In: *Commun. ACM* 55.2 (2012), pp. 70–80.

[125] D. Hothersall. *History of Psychology*. McGraw-Hill, 2004.

[126] G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[127] D. Kalainathan. 'Generative Neural networks to infer Causal Mechanisms: Algorithms and applications'. Theses. Université Paris Sud (Paris 11) - Université Paris Saclay, Dec. 2019. URL: https://hal.inria.fr/tel-02435986.

[128] D. Kalainathan, O. Goudet, P. Caillou, M. Sebag and P. Tubaro. *Conditions objectives de travail et ressenti des individus : le rôle du management*. Synthèse n. 14 de "La Fabrique de l'Industrie". 2017. URL: https://hal.inria.fr/hal-01742592.

[129] A.-H. 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.

[130] D. Kingma and M. Welling. 'Auto-encoding variational Bayes.' In: *Int. Conf. on Learning Representations (ICLR)*. 2014.

[131] L. Kotthoff. 'Algorithm Selection for Combinatorial Search Problems: A Survey'. In: *Data Mining and Constraint Programming: Foundations of a Cross-Disciplinary Approach*. Ed. by Ch. Bessiere et al. Cham: Springer Verlag, 2016, pp. 149–190.

[132] A. Krizhevsky, I. Sutskever and G. E. Hinton. 'ImageNet Classification with Deep Convolutional Neural Networks'. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'12. 2012, pp. 1097–1105.

[133] 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.

[134] Y. LeCun. *Deep Learning est mort. Vive Differentiable Programming!* Facebook Post. 2018. URL: https://www.facebook.com/yann.lecun/posts/10155003011462143.

[135] J. Z. Liang, E. Meyerson and R. Miikkulainen. 'Evolutionary architecture search for deep multitask networks'. In: *Proc. ACM-GECCO*. Ed. by H. E. Aguirre and K. Takadama. ACM, 2018, pp. 466–473. DOI: 10.1145/3205455.3205489. URL: http://doi.acm.org/10.1145/3205455.3205489.

[136] E. Lippiello, G. Petrillo, F. Landes and A. Rosso. 'Fault Heterogeneity and the Connection between Aftershocks and Afterslip'. In: *Bulletin of the Seismological Society of America* 109.3 (Apr. 2019), pp. 1156–1163. DOI: 10.1785/0120180244. URL: https://hal.inria.fr/hal-02156407.

[137] H. Lipson and J. B. Pollack. 'Automatic design and manufacture of robotic lifeforms'. In: *Nature Letters* 406 (2000), pp. 974–978.

[138] 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.

[139] Z. Liu, O. Bousquet, A. Elisseeff, S. Escalera, I. Guyon, J. Jacques, A. Pavao, D. Silver, L. Sun-Hosoya, S. Treguer, W.-W. Tu, J. Wang and Q. Yao. 'AutoDL Challenge Design and Beta Tests-Towards automatic deep learning'. In: *CiML workshop @ NIPS2018*. Montreal, Canada, Dec. 2018. URL: https://hal.archives-ouvertes.fr/hal-01906226.

[140] C. Louizos, M. Reisser, T. Blankevoort, E. Gavves and M. Welling. 'Relaxed Quantization for Discretized Neural Networks'. In: *ICLR*. 2019.

[141] A. Marot, B. Donnot, C. Romero, L. Veyrin-Forrer, M. Lerousseau, B. Donon and I. Guyon. 'Learning to run a power network challenge for training topology controllers'. In: *Electric Power Systems Research* 189 (Dec. 2020), p. 106635. DOI: 10.1016/j.epsr.2020.106635. URL: https://hal.inria.fr/hal-03159692.

[142] 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.

[143] L. Mescheder, A. Geiger and S. Nowozin. 'Which Training Methods for GANs do actually Converge?' In: *35th ICML*. International Conference on Machine Learning. 2018, pp. 3481–3490.

[144] M. Mısır and M. Sebag. 'Alors: An algorithm recommender system'. In: *Artificial Intelligence* 244 (2017). Published on-line Dec. 2016, pp. 291–314. URL: https://hal.inria.fr/hal-01419874.

[145] D. Molchanov, A. Ashukha and D. Vetrov. 'Variational Dropout Sparsifies Deep Neural Networks'. In: *ArXiv e-prints* (Jan. 2017). arXiv: 1701.05369 [stat.ML].

[146] C. O'Neill. *Weapons of Math Destruction*. Crown Books, 2016.

[147] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu. 'WaveNet: A Generative Model for Raw Audio'. In: *CoRR* abs/1609.03499 (2016). URL: http://arxiv.org/abs/1609.03499.

[148] A. Pavao, D. Kalainathan, L. Sun-Hosoya, K. Bennett and I. Guyon. 'Design and Analysis of Experiments: A Challenge Approach in Teaching'. In: *NeurIPS 2019 - 33th Annual Conference on Neural Information Processing Systems*. Vancouver, Canada, Dec. 2019. URL: https://hal.inria.fr/hal-02415639.

[149] J. Pearl. *Causality: Models, Reasoning, and Inference (2nd edition)*. Cambridge University Press, 2009.

[150] J. Pearl. 'Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution'. In: *11th ACM WSDM*. 2018, p. 3.

[151] G. Pleiss, M. Raghavan, F. Wu, J. M. Kleinberg and K. Q. Weinberger. 'On Fairness and Calibration'. In: *NIPS*. 2017, pp. 5684–5693.

[152] M. Raissi. 'Deep Hidden Physics Models: Deep Learning of Nonlinear Partial Differential Equations'. In: *JMLR* 19 (2018), pp. 1–24.

[153] H. Rakotoarison, M. Schoenauer and M. Sebag. 'Automated Machine Learning with Monte-Carlo Tree Search'. In: *IJCAI-19 - 28th International Joint Conference on Artificial Intelligence*. Macau, China: International Joint Conferences on Artificial Intelligence Organization, Aug. 2019, pp. 3296–3303. DOI: 10.24963/ijcai.2019/457. URL: https://hal.inria.fr/hal-02300884.

[154] E. Real, S. Moore, A. Selle, S. Saxena, Y. L. Suematsu, J. T. Q. Le and A. Kurakin. 'Large-Scale Evolution of Image Classifiers'. In: *34th ICML*. 2017. arXiv: 1703.01041.

[155] J. R. Rice. 'The Algorithm Selection Problem'. In: *Advances in Computers* 15 (1976). Ed. by M. Rubinoff and M. C. Yovits, pp. 65–118. DOI: 10.1016/S0065-2458(08)60520-3. URL: http://www.sciencedirect.com/science/article/pii/S0065245808605203.

[156] J. Rissanen. *Information and Complexity in Statistical Modeling*. Information Science and Statistics. Springer-Verlag, 2007.

[157] M. Schmidt and H. Lipson. 'Distilling free-form natural laws from experimental data'. In: *Science* 324.5923 (2009), pp. 81–85.

[158] J. Sirignano and K. Spiliopoulos. 'DGM: A deep learning algorithm for solving partial differential equations'. In: *Journal of Computational Physics* 375 (2018), pp. 1339–1364. DOI: https://doi.org/10.1016/j.jcp.2018.08.029.

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

[160] D. H. Stern, R. Herbrich and T. Graepel. 'Matchbox: large scale online bayesian recommendations'. In: *WWW*. Madrid, Spain: ACM Press, 2010, p. 111. DOI: 10.1145/1526709.1526725.

[161] C. Sunstein and R. Thaler. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, 2008.

[162] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow and R. Fergus. 'Intriguing properties of neural networks'. In: *CoRR* abs/1312.6199 (2013). URL: http://arxiv.org/abs/1312.6199.

[163] C. Tallec and Y. Ollivier. 'Unbiased online recurrent optimization'. In: *International Conference On Learning Representation*. Vancouver, Canada, Apr. 2018. URL: https://hal.archives-ouvertes.fr/hal-01972587.

[164] Volodymyr Mnih et al. 'Human-level control through deep reinforcement learning'. In: *Nature* 518.7540 (2015), pp. 529–533. URL: https://doi.org/10.1038/nature14236.

[165] M. Welling. *Intelligence per Kilowatthour*. Keynote talk @ICML. 2018. URL: https://icml.cc/Conferences/2018/Schedule?showEvent=1866.

[166] F. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens. 'The context-tree weighting method: basic properties'. In: *IEEE Transactions on Information Theory* 41.3 (1995), pp. 653–664.

[167] X. Zhang, C. Furtlehner, C. Germain-Renaud and M. Sebag. 'Data Stream Clustering with Affinity Propagation'. In: *IEEE Transactions on Knowledge and Data Engineering* 26.7 (2014), p. 1. URL: https://hal.inria.fr/hal-00862941.

[168] 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.