RESEARCH CENTRE

**Lille - Nord Europe**

**IN PARTNERSHIP WITH:**
**CNRS, Université de Lille**

2020
# ACTIVITY REPORT

Project-Team
# MAGNET

## Machine Learning in Information Networks

**IN COLLABORATION WITH: Centre de Recherche en Informatique, Signal et Automatique de Lille**

**DOMAIN**

**Perception, Cognition and Interaction**

**THEME**

**Data and Knowledge Representation and Processing**

# Contents

# Project-Team MAGNET

*Creation of the Team: 2013 January 01, updated into Project-Team: 2016 May 01*

## Keywords

### Computer sciences and digital sciences

A3.1.3. – Distributed data

A3.1.4. – Uncertain data

A3.4. – Machine learning and statistics

A3.4.1. – Supervised learning

A3.4.2. – Unsupervised learning

A3.4.4. – Optimization and learning

A3.4.6. – Neural networks

A4.8. – Privacy-enhancing technologies

A9.4. – Natural language processing

### Other research topics and application domains

B9.5.1. – Computer science

B9.5.6. – Data science

B9.6.8. – Linguistics

B9.6.10. – Digital humanities

B9.9. – Ethics

B9.10. – Privacy

# 1 Team members, visitors, external collaborators

## Research Scientists

- Aurelien Bellet [Inria, Researcher]
- Pascal Denis [Inria, Researcher]
- Michael Perrot [Inria, from Oct 2020, Starting Faculty Position]
- Jan Ramon [Inria, Senior Researcher, HDR]

## Faculty Members

- Marc Tommasi [Team leader, Université de Lille, Professor, HDR]
- Mikaela Keller [Université de Lille, Associate Professor]
- Fabio Vitale [Université de Lille, Associate Professor, until Aug 2020]

## Post-Doctoral Fellow

- Mohamed Maouche [Inria]

## PhD Students

- Mahsa Asadi [Inria]
- Moitree Basu [Inria]
- Gaurav Maheshwari [Inria, from Nov 2020]
- Paul Mangold [Inria, from Oct 2020]
- Onkar Pandit [Inria]
- Arijus Pleska [Inria]
- César Sabater [Inria]
- Ali Shahin Shamsabadi [Inria, from Oct 2020]
- Brij Mohan Lal Srivastava [Inria]
- Mariana Vargas Vieyra [Inria]

## Technical Staff

- Yannick Bouillard [Inria, Engineer, from Nov 2020]
- Pradipta Deb [Inria, Engineer]
- Joseph Renner [Inria, Engineer, from Oct 2020]
- Sophie Villerot [Inria, Engineer, from Nov 2020]

## Interns and Apprentices

- Edwige Cyffers [École Normale Supérieure de Lyon, from Apr 2020 until Sep 2020]
- Paul Mangold [École Normale Supérieure de Lyon, until Jul 2020]
- Gabriel Rudloff [Inria, until Mar 2020]

### Administrative Assistant

- Julie Jonas [Inria]

### External Collaborator

- Remi Gilleron [Université de Lille, HDR]

## 2 Overall objectives

The main objective of MAGNET is to develop original machine learning methods for networked data. We consider information networks in which the data consist of feature vectors or texts. We model such networks as graphs wherein nodes correspond to entities (documents, spans of text, users, datasets, learners etc.) and edges correspond to relations between entities (similarity, answer, co-authoring, friendship etc.). In *Mining and Learning in Graphs*, our main research goal is to efficiently search for the best hidden graph structure to be generated for solving a given learning task which exploits the relationships between entities. In *Machine Learning for Natural Language Processing* the objective is to go beyond vectorial classification to solve tasks like coreference resolution and entity linking, temporal structure prediction, and discourse parsing. In *Decentralized Machine Learning* we address the problem of learning in a private, fair and energy efficient way when data are naturally distributed in a network.

The challenges are the dimensionality of the input space, possibly the dimensionality of the output space, the high level of dependencies between the data, the inherent ambiguity of textual data and the limited amount of human labeling. We are interested in making machine learning approaches more acceptable to society. Privacy, sobriety and fairness are important issues that pertain to this research line, and we are interested in the empowerment of end users in the machine learning processes.

## 3 Research program

The research program of MAGNET is structured along three main axes.

**Axis 1: Mining and Learning in Graphs** This axis is the backbone of the team. Most of the techniques and algorithms developed in this axis are known by the team members and have impact on the two other axes. We address the following questions and objectives:

*How to adaptively build graphs with respect to the given tasks?* We study adaptive graph construction along several directions. The first one is to learn the best similarity measure for the graph construction. The second one is to combine different views over the data in the graph construction and learn good representations. We also study weak forms of supervision like comparisons.

*How to design methods able to achieve a good trade-off between predictive accuracy and computational complexity?* We develop new algorithms for efficient graph-based learning (for instance node prediction or link prediction). In order to deal with scalability issues, our approach is based on optimization, graph sparsification techniques and graph sampling methods.

*How to find patterns in graphs based on efficient computations of some statistics?* We develop graph mining algorithms and statistics in the context of correlated data.

**Axis 2: Machine Learning for Natural Language Processing** In this axis, we address the general question that relates graph-based learning and Natural Language Processing (NLP): *How to go beyond vectorial classification models in NLP tasks?* We study the combination of learning representation, structured prediction and graph-based learning methods. Data sobriety and fairness are major constraints we want to deal with. The targeted NLP tasks are coreference resolution and entity linking, temporal structure prediction, and discourse parsing.

**Axis 3: Decentralized Machine Learning and Privacy** In this axis, we study *How to design private by design machine learning algorithms?* Taking as an opportunity the fact that data collection is now decentralized on smart devices, we propose alternatives to large data centers where data are gathered by developing collaborative and personalized learning.

Contrary to many machine learning approaches where data points and tasks are considered in isolation, we think that a key point of this research is to be able to leverage the relationships between data and learning objectives. Therefore, using graphs as an abstraction of information networks is a major playground for MAGNET. Research related to graph data is a transversal axis, describing a layer of work supporting two other axes on Natural Language Processing and decentralized learning. The machine learning and mining in graphs communities have evolved, for instance taking into account data streams, dynamics but maybe more importantly, focusing on deep learning. Deep neural nets are here to stay, and they are useful tools to tackle difficult problems so we embrace them at different places in the three axes.

MAGNET conducts research along the three axis described above but will put more emphasis on social issues of machine learning. In the context of the recent deployment of artificial intelligence into our daily lives, we are interested in making machine learning approaches more acceptable to society. Privacy, sobriety and fairness are important issues that pertain to this research line, but more generally we are interested in the empowerment of end users in the machine learning processes. Reducing the need of one central authority and pushing more the data processing on the user side, that is decentralization, also participates to this effort. Reducing resources means reducing costs and energy and contributes to building more accessible technologies for companies and users. By considering learning tasks in a more personalized way, but increasing collaboration, we think that we can design solutions that work in low resources regime, with less data or supervision.

In MAGNET we emphasize a different approach than blindly brute-forcing tasks with loads of data. Applications to social sciences for instance have different needs and constraints that motivate data sobriety, fairness and privacy. We are interested in weaker supervision, by leveraging structural properties described in graphs of data, relying on transfer and multi-task learning when faced with graphs of tasks and users. Algorithmic and statistical challenges related to the graph structure of the data still contain open questions. On the statistical side, examples are to take dependencies into account, for instance to compute a mean, to reduce the need of sampling by exploiting known correlations. For the algorithmic point of view, going beyond unlabeled undirected graphs, in particular considering attributed graphs containing text or other information and addressing the case of distributed graphs while maintaining formal guarantees are getting more attention.

In the second axis devoted to NLP, we focus our research on graph-based and representation learning into several directions, all aiming at learning *richer, more robust, and more transferable linguistic representations*. This research program will attempt to bring about strong cross-fertilizations with the other axes, addressing problems in graph, privacy and fairness and making links with decentralized learning. At the intersection between graph-based and representation learning, we will first develop graph embedding algorithms for deriving linguistic representations which are able to capture higher-level semantic and world-knowledge information which eludes strictly distributional models. As an initial step, we envision leveraging pre-existing ontologies (e.g., WordNet, DBpedia), from which one can easily derive interesting similarity graphs between words or noun phrases. We also plan to investigate innovative ways of articulating graph-based semi-supervised learning algorithms and word embedding techniques. A second direction involves learning representations that are more robust to bias, privacy attacks and adversarial examples. Thus, we intend to leverage recent adversarial training strategies, in which an adversary attempts to recover sensitive attributes (e.g., gender, race) from the learned representations, to be able to neutralize bias or to remove sensitive features. An application domain for this line of research is for instance speech data. The study of learning private representation with its link to fairness in the decentralized setting is another important research topic for the team. In this context of fairness, we also intend to develop similar algorithms for detecting slants, and ultimately for generating de-biased or "re-biased" versions of text embeddings. An illustration is on political slant in written texts (e.g., political speeches and manifestos). Thirdly, we intend to learn linguistic representations that can transfer more easily across languages and domains, in particular in the context of structured prediction problems for low-resource languages. For instance, we first propose to jointly learn model parameters for each language (and/or domains) in a multi-task setting, and leverage a (pre-existing or learned) graph encoding structural similarities between languages (and/or domains). This type of approach would nicely tie in with our previous work on multilingual dependency parsing and on learning personalized models. Furthermore, we will also study how to combine and adapt some neural architectures recently introduced for sequence-to-sequence problems in order to enable transfer of language representations.

In terms of technological transfer, we maintain collaborations with researchers in the humanities and the social sciences, helping them to leverage state-of-the-art NLP techniques to develop new insights to their research by extracting relevant information from large amounts of texts.

The third axis is on distributed and decentralized learning and privacy preserving machine learning. Recent years have seen the evolution of information systems towards ubiquitous computing, smart objects and applications fueled by artificial intelligence. Data are collected on smart devices like smartphones, watches, home devices etc. They include texts, locations, social relationships. Many sensitive data —race, gender, health conditions, tastes etc— can be inferred. Others are just recorded like activities, social relationships but also biometric data like voice and measurements from sensor data. The main tendency is to transfer data into central servers mostly owned by a few tier parties. The situation generates high privacy risks for the users for many reasons: loss of data control, unique entry point for data access, unsolicited data usage etc. But it also increases monopolistic situations and tends to develop oversized infrastructures. The centralized paradigm also has limits when data are too huge such as in the case of multiple videos and sensor data collected for autonomous driving. Partially or fully decentralized systems provide an alternative, to emphasis data exploitation rather than data sharing. For MAGNET, they are source of many new research directions in machine learning at two scales: at the algorithmic level and at a systemic level.

At the algorithmic level the question is to develop new privacy preserving algorithms in the context of decentralized systems. In this context, data remains where it has been collected and learning or statistical queries are processed at the local level. An important question we study is to take into account and measure the impact of collaboration. We also aim at developing methods in the online setting where data arrives continuously or participants join and leave the collaboration network. The granularity of exchanges, the communication cost and the dynamic scenarios, are also studied. On the privacy side, decentralization is not sufficient to establish privacy guarantees because learned models together with the dynamics of collaborative learning may reveal private training data if the models are published or if the communications are observed. But, although it has not been yet well established, decentralization can naturally increase privacy-utility ratio. A direction of research is to formally prove the privacy gain when randomized decentralized protocols are used during learning. In some situations, for instance when part of the data is not sensitive or when trusted servers can be used, a combination between a fully decentralized and a centralized approach is very relevant. In this setting, the question is to find a good trade-off between local versus global computations.

At the systemic layer, in MAGNET we feel that there is a need for research on a global and holistic level, that is to consider full processes involving learning, interacting, predicting, reasoning, repeating etc. rather than studying the privacy of isolated learning algorithms. Our objective is to design languages for describing processes (workflows), data (database schema, background knowledge), population statistics, privacy properties of algorithms, privacy requirements and other relevant information. This is fully aligned with recent trends that aim at giving to statistical learning a more higher level of formal specifications and illustrates our objective for more acceptable and transparent machine learning. We also work towards more robust privacy-friendly systems, being able to handle a wider range of malicious behavior such as collusion to obtain information or inputting incorrect data to obtain information or to influence the result of collaborative computations. From the transfer point of view, we plan to apply transparent, privacy-friendly in significant application domains, such as medicine, surveying, demand prediction and recommendation. In this context, we are interested to understand the appreciation of humans of transparency, verifiability, fairness, privacy-preserving and other trust-increasing aspects of our technologies.

# 4 Application domains

Our application domain cover health, mobility, social sciences and voice technologies. VVVV

**Health** Privacy is of major importance in the health domain. We contribute to develop methods to give access to the use of data in a private way rather than to the data itself centralized in vulnerable single locations. As an example, we are working with hospitals to develop the means of multicentric studies with privacy guarantees. A second example is personalized medicine where personal devices collect private and highly sensitive data. Potential applications of our research allow to keep data on device and to privately compute statistics.

**Social sciences** Our NLP research activities are rooted in linguistics, but learning unbiased representations of texts for instance or simply identifying unfair representations also have impacts in political sciences and history.

**Voice technologies** We develop methods for privacy in speech that can be embedded in software suites dedicated to voice-based interaction systems.

# 5 Highlights of the year

AURÉLIEN BELLET is co-organizing the Federated Learning One World webinar (700+ registered attendees)

## 5.1 Awards

MATHIEU DEHOUCK has received an award from the French association on Natural Language Processing (ATALA) for his PhD "Multi-Lingual Dependency Parsing: Word Representation and Joint Training for Syntactic Analysis".

# 6 New software and platforms

## 6.1 New software

### 6.1.1 CoRTeX

**Name:** Python library for noun phrase COreference Resolution in natural language TEXts

**Keyword:** Natural language processing

**Functional Description:** CoRTex is a LGPL-licensed Python library for Noun Phrase coreference resolution in natural language texts. This library contains implementations of various state-of-the-art coreference resolution algorithms, including those developed in our research. In addition, it provides a set of APIs and utilities for text pre-processing, reading the CONLL2012 and CONLLU annotation formats, and performing evaluation, notably based on the main evaluation metrics (MUC, B-CUBED, and CEAF). As such, CoRTex provides benchmarks for researchers working on coreference resolution, but it is also of interest for developers who want to integrate a coreference resolution within a larger platform. It currently supports use of the English or French language.

**Contact:** Pascal Denis

**Participant:** Pascal Denis

**Partner:** Orange Labs

### 6.1.2 Mangoes

**Name:** MAgnet liNGuistic wOrd vEctorS

**Keywords:** Word embeddings, NLP

**Functional Description:** Mangoes is a toolbox for constructing and evaluating static and contextual token vector representations (aka embeddings). The main functionalities are:

- Contextual embeddings: Access a large collection of pretrained transformer-based language models, Pre-train a BERT language model on a corpus, Fine-tune a BERT language model for a number of extrinsic tasks, Extract features/predictions from pretrained language models.

- Static embeddings: Process textual data and compute vocabularies and co-occurrence matrices. Input data should be raw text or annotated text, Compute static word embeddings with different state-of-the art unsupervised methods, Propose statistical and intrinsic evaluation methods, as well as some visualization tools, Generate context dependent embeddings from a pretrained language model.

Future releases will include methods for injecting lexical and semantic knowledge into token and multi-model embeddings, and interfaces into common external knowledge resources.

**URL:** https://gitlab.inria.fr/magnet/mangoes

**Contact:** Nathalie Vauquier

### 6.1.3 metric-learn

**Keywords:** Machine learning, Python, Metric learning

**Functional Description:** Distance metrics are widely used in the machine learning literature. Traditionally, practicioners would choose a standard distance metric (Euclidean, City-Block, Cosine, etc.) using a priori knowledge of the domain. Distance metric learning (or simply, metric learning) is the sub-field of machine learning dedicated to automatically constructing optimal distance metrics.

This package contains efficient Python implementations of several popular metric learning algorithms.

**URL:** https://github.com/scikit-learn-contrib/metric-learn

**Contact:** Aurélien Bellet

**Partner:** Parietal

### 6.1.4 MyLocalInfo

**Keywords:** Privacy, Machine learning, Statistics

**Functional Description:** Decentralized algorithms for machine learning and inference tasks which (1) perform as much computation as possible locally and (2) ensure privacy and security by avoiding that personal data leaves devices.

**Contact:** Nathalie Vauquier

### 6.1.5 COMPRISE Voice Transformer

**Name:** COMPRISE Voice Transformer

**Keywords:** Speech, Privacy

**Functional Description:** COMPRISE Voice Transformer is an open source tool that increases the privacy of users of voice interfaces by converting their voice into another person's voice without modifying the spoken message. It ensures that any information extracted from the transformed voice can hardly be traced back to the original speaker, as validated through state-of-the-art biometric protocols, and it preserves the phonetic information required for human labelling and training of speech-to-text models.

**Release Contributions:** This version gives access to the 2 generations of tools that have been used to transform the voice, as part of the COMPRISE project (https://www.compriseh2020.eu/). The first one is a python library that implements 2 basic voice conversion methods, both using VLTN. The second one implements an anonymization method using x-vectors and neural waveform models.

**URL:** https://gitlab.inria.fr/comprise/voice_transformation

**Contact:** Marc Tommasi

**Participants:** Nathalie Vauquier, Brij Mohan Lal Srivastava, Marc Tommasi, Emmanuel Vincent, Md Sahidullah

# 7 New results

## 7.1 Natural Language Processing

**Integrating knowledge graph embeddings to improve mention representation for bridging anaphora resolution** Lexical semantics and world knowledge are crucial for interpreting bridging anaphora[1]. Yet, existing computational methods for acquiring and injecting this type of information into bridging resolution systems suffer important limitations. Based on explicit querying of external knowledge bases, earlier approaches are computationally expensive (hence, hardly scalable) and they map the data to be processed into high-dimensional spaces (careful handling of the curse of dimensionality and overfitting has to be in order). In [24], we take a different and principled approach which naturally addresses these issues. Specifically, we convert the external knowledge source (in this case, WordNet) into a graph, and learn embeddings of the graph nodes of low dimension to capture the crucial features of the graph topology and, at the same time, rich semantic information. Once properly identified from the mention text spans, these low dimensional graph node embeddings are combined with distributional text-based embeddings to provide enhanced mention representations. We illustrate the effectiveness of our approach by evaluating it on commonly used datasets, namely ISNotes [36] and BASHI [37]. Our enhanced mention representations yield significant accuracy improvements on both datasets when compared to different standalone text-based mention representations.

## 7.2 Decentralized Learning

**Fully Decentralized Joint Learning of Personalized Models and Collaboration Graphs** In [23], we consider the fully decentralized machine learning scenario where many users with personal datasets collaborate to learn models through local peer-to-peer exchanges, without a central coordinator. We propose to train personalized models that leverage a collaboration graph describing the relationships between user personal tasks, which we learn jointly with the models. Our fully decentralized optimization procedure alternates between training nonlinear models given the graph in a greedy boosting manner, and updating the collaboration graph (with controlled sparsity) given the models. Throughout the process, users exchange messages only with a small number of peers (their direct neighbors when updating the models, and a few random users when updating the graph), ensuring that the procedure naturally scales with the number of users. Overall, our approach is communication-efficient and avoids exchanging personal data. We provide an extensive analysis of the convergence rate, memory and communication complexity of our approach, and demonstrate its benefits compared to competing techniques on synthetic and real datasets.

## 7.3 Privacy and Machine Learning

**Who started this rumor? Quantifying the natural differential privacy guarantees of gossip protocols** Gossip protocols are widely used to disseminate information in massive peer-to-peer networks. These protocols are often claimed to guarantee privacy because of the uncertainty they introduce on the node that started the dissemination. But is that claim really true? Can the source of a gossip safely hide in the crowd? In [1] we examine, for the first time, gossip protocols through a rigorous mathematical framework based on differential privacy to determine the extent to which the source of a gossip can be traceable. Considering the case of a complete graph in which a subset of the nodes are curious, we study a family of gossip protocols parameterized by a “muting” parameter $s$: nodes stop emitting after each communication with a fixed probability $1-s$. We first prove that the standard push protocol, corresponding to the case $s=1$, does not satisfy differential privacy for large graphs. In contrast, the protocol with $s=0$ achieves optimal privacy guarantees but at the cost of a drastic increase in the spreading time compared

[1] An anaphor is an expression whose interpretation depends upon a previous expression in the discourse, an antecedent. A Bridging anaphor is a special type of anaphor where there is non-identical or associative relation with its antecedent.

to standard push, revealing an interesting tension between privacy and spreading time. Yet, surprisingly, we show that some choices of the muting parameter $s$ lead to protocols that achieve an optimal order of magnitude in both privacy and speed. We also confirm empirically that, with appropriate choices of $s$, we indeed obtain protocols that are very robust against concrete source location attacks while spreading the information almost as fast as the standard (and non-private) push protocol.

**Private Protocols for U-Statistics in the Local Model and Beyond** In [15], we study the problem of computing U-statistics of degree 2, i.e., quantities that come in the form of averages over pairs of data points, in the local model of differential privacy (LDP). The class of U-statistics covers many statistical estimates of interest, including Gini mean difference, Kendall's tau coefficient and Area under the ROC Curve (AUC), as well as empirical risk measures for machine learning problems such as ranking, clustering and metric learning. We first introduce an LDP protocol based on quantizing the data into bins and applying randomized response, which guarantees an $\epsilon$-LDP estimate with a Mean Squared Error (MSE) of $O(1/\sqrt{n}\epsilon)$ under regularity assumptions on the U-statistic or the data distribution. We then propose a specialized protocol for AUC based on a novel use of hierarchical histograms that achieves MSE of $O(\alpha^3/n\epsilon^2)$ for arbitrary data distribution on a domain with $2^\alpha$ elements. We also show that 2-party secure computation allows to design a protocol with MSE of $O(1/n\epsilon^2)$, without any assumption on the kernel function or data distribution and with total communication linear in the number of users $n$. Finally, we evaluate the performance of our protocols through experiments on synthetic and real datasets.

**Interpretable privacy with optimizable utility** In a position paper ([25]), we discuss the problem of specifying privacy requirements for machine learning based systems, in an interpretable yet operational way. Explaining privacy-improving technology is a challenging problem, especially when the goal is to construct a system which at the same time is interpretable and has a high performance. In order to address this challenge, we propose to specify privacy requirements as constraints, leaving several options for the concrete implementation of the system open, followed by a constraint optimization approach to achieve an efficient implementation also, next to the interpretable privacy guarantees.

**Reconstructing Genotypes in Private Genomic Databases from Genetic Risk Scores** Some organisations like 23andMe and the UK Biobank have large genomic databases that they re-use for multiple different genomewide association studies (GWAS). Even research studies that compile smaller genomic databases often utilise these databases to investigate many related traits. It is common for the study to report a genetic risk score (GRS) model for each trait within the publication. In [18] we show that under some circumstances, these GRS models can be used to recover the genetic variants of individuals in these genomic databases—a reconstruction attack. In particular, if two GRS models are trained using a largely overlapping set of participants, then it is often possible to determine the genotype for each of the individuals who were used to train one GRS model, but not the other. We demonstrate this theoretically and experimentally by analysing the Cornell Dog Genome database. The accuracy of our reconstruction attack depends on how accurately we can estimate the rate of co-occurrence of pairs of single nucleotide polymorphisms (SNP) within the private database, so if this aggregate information is ever released, it would drastically reduce the security of a private genomic database. Caution should be applied when using the same database for multiple analysis, especially when a small number of individuals are included or excluded from one part of the study.

**A Decentralized Framework for Biostatistics and Privacy Concerns** The paper [13] is the first result of the new collaboration engaged with the Lille Hospital (CHRU) and its INCLUDE team. Biostatistics and machine learning have been the cornerstone of a variety of recent developments in medicine. In order to gather large enough datasets, it is often necessary to set up multi-centric studies; yet, centralization of measurements can be difficult, either for practical, legal or ethical reasons. As an alternative, federated learning enables leveraging multiple centers' data without actually collating them. While existing works generally require a center to act as a leader and coordinate computations, we propose a fully decentralized framework where each center plays the same role. In this paper, we apply this framework to logistic regression, including confidence intervals computation. We test our algorithm on two distinct clinical datasets split among different centers, and show that it matches results from the centralized framework.

In addition, we discuss possible privacy leaks and potential protection mechanisms, paving the way towards further research.

**Distributed Differentially Private Averaging with Improved Utility and Robustness to Malicious Parties** Learning from data owned by several parties, as in federated learning, raises challenges regarding the privacy guarantees provided to participants and the correctness of the computation in the presence of malicious parties. In [32, 33], we tackle these challenges in the context of distributed averaging, an essential building block of distributed and federated learning. Our first contribution is a novel distributed differentially private protocol which can match the accuracy of the trusted curator model even when each party communicates with only a logarithmic number of other parties chosen at random. Our second contribution is to enable users to prove the correctness of their computations without compromising the efficiency and privacy guarantees of the protocol. Our construction relies on standard cryptographic primitives like commitment schemes and zero knowledge proofs.

## 7.4 Learning in Graphs

**Joint Learning of the Graph and the Data Representation for Graph-Based Semi-Supervised Learning** Graph-based semi-supervised learning is appealing when labels are scarce but large amounts of unlabeled data are available. These methods typically use a heuristic strategy to construct the graph based on some fixed data representation, independently of the available labels. In [22], we propose to jointly learn a data representation and a graph from both labeled and unlabeled data such that (i) the learned representation indirectly encodes the label information injected into the graph, and (ii) the graph provides a smooth topology with respect to the transformed data. Plugging the resulting graph and representation into existing graph-based semi-supervised learning algorithms like label spreading and graph convolutional networks, we show that our approach outperforms standard graph construction methods on both synthetic data and real datasets.

**Skill Rating for Multiplayer Games Introducing Hypernode Graphs and their Spectral Theory** This result corresponds to an extended version [8] of the results obtained on learning with hypergraphs. We consider the skill rating problem for multiplayer games, that is how to infer player skills from game outcomes in multiplayer games. We formulate the problem as a minimization problem $\operatorname{argmin}_s s^\top \Delta s$ where $\Delta$ is a positive semidefinite matrix and $s$ a real-valued function, of which some entries are the skill values to be inferred and other entries are constrained by the game outcomes. We leverage graph-based semi-supervised learning (SSL) algorithms for this problem. We apply our algorithms on several data sets of multiplayer games and obtain very promising results compared to Elo Duelling [34] and TrueSkill [35]. As we leverage graph-based SSL algorithms and because games can be seen as relations between sets of players, we then generalize the approach. For this aim, we introduce a new finite model, called hypernode graph, defined to be a set of weighted binary relations between sets of nodes. We define Laplacians of hypernode graphs. Then, we show that the skill rating problem for multiplayer games can be formulated as $\operatorname{argmin}_s s^\top \Delta s$ where $\Delta$ is the Laplacian of a hypernode graph constructed from a set of games. From a fundamental perspective, we show that hypernode graph Laplacians are symmetric positive semidefinite matrices with constant functions in their null space. We show that problems on hypernode graphs cannot be solved with graph constructions and graph kernels. We relate hypernode graphs to signed graphs showing that positive relations between groups can lead to negative relations between individuals.

## 7.5 Metric Learning

**metric-learn: Metric Learning Algorithms in Python** The paper [12] describes `metric-learn`, an open source Python package implementing supervised and weakly supervised distance metric learning algorithms. As part of `scikit-learn-contrib`, it provides a unified interface compatible with `scikit-learn` which allows to easily perform cross-validation, model selection, and pipelining with other machine learning estimators. `metric-learn` is thoroughly tested and available on `PyPi` under the MIT license.

## 7.6 Learning and Speech Recognition

**Introducing the VoicePrivacy initiative** The VoicePrivacy initiative aims to promote the development of privacy preservation tools for speech technology by gathering a new community to define the tasks of interest and the evaluation methodology, and benchmarking solutions through a series of challenges. In [21], we formulate the voice anonymization task selected for the VoicePrivacy 2020 Challenge and describe the datasets used for system development and evaluation. We also present the attack models and the associated objective and subjective evaluation metrics. We introduce two anonymization baselines and report objective evaluation results.

**Design Choices for X-vector Based Speaker Anonymization** The recently proposed x-vector based anonymization scheme converts any input voice into that of a random pseudo-speaker. In [19], we present a flexible pseudo-speaker selection technique as a baseline for the first VoicePrivacy Challenge. We explore several design choices for the distance metric between speakers, the region of x-vector space where the pseudo-speaker is picked, and gender selection. To assess the strength of anonymization achieved, we consider attackers using an x-vector based speaker verification system who may use original or anonymized speech for enrollment, depending on their knowledge of the anonymization scheme. The Equal Error Rate (EER) achieved by the attackers and the decoding Word Error Rate (WER) over anonymized data are reported as the measures of privacy and utility. Experiments are performed using datasets derived from LibriSpeech to find the optimal combination of design choices in terms of privacy and utility.

**A comparative study of speech anonymization metrics** Speech anonymization techniques have recently been proposed for preserving speakers' privacy. They aim at concealing speak-ers' identities while preserving the spoken content. In [17], we compare three metrics proposed in the literature to assess the level of privacy achieved. We exhibit through simulation the differences and blindspots of some metrics. In addition, we conduct experiments on real data and state-of-the-art anonymiza-tion techniques to study how they behave in a practical scenario. We show that the application-independent log-likelihood-ratio cost function $C_{\text{llr}}^{\text{min}}$ provides a more robust evaluation of privacy than the equal error rate (EER), and that detection-based metrics provide different information from linkability metrics. Interestingly, the results on real data indicate that current anonymization design choices do not induce a regime where the differences between those metrics become apparent.

**Evaluating Voice Conversion-based Privacy Protection against Informed Attackers** Speech data conveys sensitive speaker attributes like identity or accent. With a small amount of found data, such attributes can be inferred and exploited for malicious purposes: voice cloning, spoofing etc. Anonymization aims to make the data unlinkable, i.e., ensure that no utterance can be linked to its original speaker. In [9], we investigate anonymization methods based on voice conversion. In contrast to prior work, we argue that various linkage attacks can be designed depending on the attackers' knowledge about the anonymization scheme. We compare two frequency warping-based conversion methods and a deep learning based method in three attack scenarios. The utility of converted speech is measured via the word error rate achieved by automatic speech recognition, while privacy protection is assessed by the increase in equal error rate achieved by state-of-the-art i-vector or x-vector based speaker verification. Our results show that voice conversion schemes are unable to effectively protect against an attacker that has extensive knowledge of the type of conversion and how it has been applied, but may provide some protection against less knowledgeable attackers.

# 8 Partnerships and cooperations

## 8.1 International initiatives

### 8.1.1 Inria International Labs

**LEGO — LEarning GOod representations for natural language processing**

**Participants** Aurélien Bellet *(coordinator & contact person)*, Pascal Denis.

**Duration:** 2019 – 2021

**Partners:** Theoretical and Empirical Data Science (TEDS) research group Department of Computer Science, University of Southern California (United States)

**Summary:** LEGO lies in the intersection of Machine Learning and Natural Language Processing (NLP). Its goal is to address the following challenges: what are the right representations for text data and how to learn them in a robust and transferable way? How to apply such representations to solve real-world NLP tasks, specifically in scenarios where linguistic resources are scarce? The past years have seen an increasing interest in learning continuous vectorial embeddings, which can be trained together with the prediction model in an end-to-end fashion, as in recent sequence-to-sequence neural models. However, they are unsuitable to low-resource languages as they require massive amounts of data to train. They are also very prone to overfitting, which makes them very brittle, and sensitive to bias present in the original text as well as to confounding factors such as author attributes. LEGO strongly relies on the complementary expertise of the two partners in areas such as representation learning, structured prediction, graph-based learning, multi-task/transfer learning, and statistical NLP to offer a novel alternative to existing techniques. Specifically, we propose to investigate the following two research directions: (a) optimize the representations to make them robust to bias and adversarial examples, and (b) learn transferable representations across languages and domains, in particular in the context of structured prediction problems for low-resource languages. We will demonstrate the usefulness of the proposed methods on several NLP tasks, including multilingual dependency parsing, machine translation, question answering and text summarization.

### 8.1.2 Inria associate team not involved in an IIL

**North-European Associate Team**

- Project Acronym: PAD-ML
- Project title: Privacy-Aware Distributed Machine Learning.
- International Partner: the PPDA team at the Alan Turing Institute.
- Start year: 2018
- In the context of increasing legislation on data protection (e.g., the recent GDPR), an important challenge is to develop privacy-preserving algorithms to learn from datasets distributed across multiple data owners who do not want to share their data. The goal of this joint team is to devise novel privacy-preserving, distributed machine learning algorithms and to assess their performance and guarantees in both theoretical and practical terms.

### 8.1.3 Participation in other international programs

**Program: Bilateral ANR project with Luxembourg**

**Participants** Pascal Denis *(contact person)*, Aurélien Bellet, Mikaela Keller, Gaurav Maheshwari.

- Project acronym: SLANT
- Project title: Spin and Bias in Language Analyzed in News and Texts
- Duration: December 2019 – June 2023.
- Coordinator: Philippe Muller, IRIT, Toulouse
- Other partners: IRIT (Toulouse), SnT (Luxembourg)
- Abstract: There is a growing concern about misinformation or biased information in public communication, whether in traditional media or social forums. While automating fact-checking has received a lot of attention, the problem of fair information is much larger and includes more insidious forms like biased presentation of events and discussion. The SLANT project aims at characterizing bias in textual data, either intended, in public reporting, or unintended in writing aiming at neutrality. An abstract model of biased interpretation using work on discourse structure, semantics and interpretation will be complemented and concretized by finding relevant lexical, syntactic, stylistic or rhetorical differences through an automated but explainable comparison of texts with different biases on the same subject, based on a dataset of news media coverage from a diverse set of sources. We will also explore how our results can help alter bias in texts or remove it from automated representations of texts.

**Program: Bilateral ANR project with Switerland**

**Participants** Pascal Denis *(contact person)*, Mathieu Dehouck.

- Project acronym: REM
- Project title: Re-thinking English Modal Constructions: From feature-based paradigms to usage-based probabilistic representations
- Duration: 2016-2021
- Coordinator: Ilse Depreatere (Université de Lille) and Martin Hilpert (University of Neuchâtel, Switzerland)
- Other partners: STL (Université de Lille), (University of Neuchâtel, Switzerland)
- Abstract: One of the central features of human language is that speakers can verbalize states of affairs that are not factual, but that rather should, might, or could be the case. Non-factual ideas can be expressed through words and constructions that belong to the grammatical domain of modality. The main question of this project relates modality to human cognition and the mental representation of language: *How are modal expressions mentally represented?* More specifically, this project proposes to go beyond the standard categorical analysis of modal verbs and instead analyze them using an exemplar-based and probabilistic approach. We thus view speakers' knowledge of modal expression not as a discrete one-to-one mapping between a form and a list of semantic features, but rather as knowledge of the probability that a given form will convey a certain meaning in a certain context.

## 8.2 European initiatives

### 8.2.1 FP7 & H2020 Projects

**COMPRISE — Cost-effective, Multilingual, Privacy-driven voice-enabled Services**

**Participants** Aurélien Bellet, Marc Tommasi *(WP2 leader)*, Brij Mohan Lal Srivastava.

- Duration: 2018 – 2021
- Coordinator: Emmanuel Vincent, Inria Nancy – Grand Est
- Partners: Ascora GMBH (Germany), Netfective Technology SA, Rooter Analysis SL (Spain), Tilde SIA (Latvia), Saarland University (Germany), Université de Lorraine, Université de Lille
- Inria contact: Akira Campbell, Inria Nancy – Grand Est
- Summary: Besides visual and tactile, the Next Generation Internet will rely more and more on voice interaction. This technology requires huge amounts of speech and language data in every language to reach state-of-the-art performance. The standard today is to store the voices of end users in the cloud and label them manually. This approach raises critical privacy concerns, it limits the number of deployed languages, and it has led to market and data concentration in the hands of big non-European companies such as Google, Facebook etc.

  COMPRISE defines a fully private-by-design methodology and tools that will reduce the cost and increase the inclusiveness of voice interaction technology through research advances on privacy-driven data transformations, personalised learning, automatic labelling, and integrated translation. This leads to a holistic easy-to-use software development kit interoperatingwith a cloud-based resource platform. The sustainability of this new ecosystem will be demonstrated for three sectors with high commercial impact: smart consumer apps, e-commerce, and e-health. COMPRISE will address the mission-oriented challenges of privacy-by-design, inclusiveness, and cost-effectiveness in a sector-agnostic way; allow virtually unlimited collection of real-life non-private quality speech and language data; enable businesses in the Digital Single Market to quickly develop multilingual voice-enabled services in many languages; allow all citizens to transparently access contents and services available in other languages by voice interaction in their own language; result in cost savings for both technology providers and users.

  COMPRISE will find application in many sectors beyond those demonstrated, e.g., e-government, e-justice, e-learning, tourism, culture, media etc. It will have a huge societal impact in terms of unprecedented verifiable privacy guarantees, service to speakers of under-resourced languages or accented speakers, and overall user experience.

### 8.2.2 Collaborations in European programs, except FP7 and H2020

**Program: Bilateral Inria-DFKI project**

**Participants** Pascal Denis *(Inria coordinator)*, Priyansh Trivedi.

- Project acronym: IMPRESS
- Project title: Improving Embeddings with Semantic Knowledge
- Duration: October 2020 – September 2023
- DFKI coordinator: Ivana Kruijff-Korbayova
- Other partners: Sémagramme (Inria Nancy), DFKI (Germany)

- Abstract: Virtually all NLP systems nowadays use vector representations of words, a.k.a. word embeddings. Similarly, the processing of language combined with vision or other sensory modalities employs multimodal embeddings. While embeddings do embody some form of semantic relatedness, the exact nature of the latter remains unclear. This loss of precise semantic information can affect downstream tasks. Furthermore, while there is a growing body of NLP research on languages other than English, most research on multimodal embeddings is still done on English. The goals of IMPRESS are to investigate the integration of semantic knowledge into embeddings and its impact on selected downstream tasks, to extend this approach to multimodal and mildly multilingual settings, and to develop open source software and lexical resources, focusing on video activity recognition as a practical testbed.

## 8.3 National initiatives

**ANR Pamela (2016–2022)**

**Participants** Marc Tommasi *(contact person)*, Aurélien Bellet, Rémi Gilleron, Jan Ramon, Mahsa Asadi.

The Pamela project aims at developing machine learning theories and algorithms in order to learn local and personalized models from data distributed over networked infrastructures. Our project seeks to provide first answers to modern information systems built by interconnecting many personal devices holding private user data in the search of personalized suggestions and recommendations. More precisely, we will focus on learning in a collaborative way with the help of neighbors in a network. We aim to lay the first blocks of a scientific foundation for these new types of systems, in effect moving from graphs of data to graphs of data and learned models. We argue that this shift is necessary in order to address the new constraints arising from the decentralization of information that is inherent to the emergence of big data. We will in particular focus on the question of learning under communication and privacy constraints. A significant asset of the project is the quality of its industrial partners, Snips and Mediego, who bring in their expertise in privacy protection and distributed computing as well as use cases and datasets. They will contribute to translate this fundamental research effort into concrete outcomes by developing personalized and privacy-aware assistants able to provide contextualized recommendations on small devices and smartphones.

https://project.inria.fr/pamela/

**ANR JCJC GRASP (2016–2022)**

**Participants** Pascal Denis *(contact person)*, Aurélien Bellet, Rémi Gilleron, Mikaela Keller, Marc Tommasi.

The GRASP project aims at designing new graph-based Machine Learning algorithms that are better tailored to Natural Language Processing structured output problems. Focusing on semi-supervised learning scenarios, we will extend current graph-based learning approaches along two main directions: (i) the use of structured outputs during inference, and (ii) a graph construction mechanism that is more dependent on the task objective and more closely related to label inference. Combined, these two research strands will provide an important step towards delivering more adaptive (to new domains and languages), more accurate, and ultimately more useful language technologies. We will target semantic and pragmatic tasks such as coreference resolution, temporal chronology prediction, and discourse parsing for which proper Machine Learning solutions are still lacking.

https://project.inria.fr/grasp/

**ANR DEEP-Privacy (2019–2023)**

**Participants** Marc Tommasi *(contact person)*, Aurélien Bellet, Pascal Denis, Jan Ramon, Brij Mohan Lal Srivastava.

DEEP-PRIVACY proposes a new paradigm based on a distributed, personalized, and privacy-preserving approach for speech processing, with a focus on machine learning algorithms for speech recognition. To this end, we propose to rely on a hybrid approach: the device of each user does not share its raw speech data and runs some private computations locally, while some cross-user computations are done by communicating through a server (or a peer-to-peer network). To satisfy privacy requirements at the acoustic level, the information communicated to the server should not expose sensitive speaker information.

### 8.3.1 ANR-JCJC PRIDE (2020–2024)

**Participants** Aurélien Bellet *(contact person)*, Paul Mangold.

Machine learning (ML) is ubiquitous in AI-based services and data-oriented scientific fields but raises serious privacy concerns when training on personal data. The starting point of PRIDE is that personal data should belong to the individual who produces it. This requires to revisit ML algorithms to learn from many decentralized personal datasets while preventing the reconstruction of raw data. Differential Privacy (DP) provides a strong notion of protection, but current decentralized ML algorithms are not able to learn useful models under DP. The goal of PRIDE is to develop theoretical and algorithmic tools that enable differentially-private ML methods operating on decentralized datasets, through two complementary objectives: (1) prove that gossip protocols naturally reinforce DP guarantees; (2) propose algorithms at the intersection of decentralized ML and secure multi-party computation.

### 8.3.2 ANR PMR (2020-2024)

**Participants** Jan Ramon *(contact person)*.

Given the growing awareness of privacy risks of data processing, there is an increasing interest in privacy-preserving learning. However, shortcomings in the state of the art limit the applicability of the privacy-preserving learning paradigm. First, most approaches assume too optimistically a honest-but-curious setting. Second, most approaches consider one learning task in isolation, not accounting for the context where querying is a recurring activity. We will investigate new algorithms and models that address these shortcomings. Among others, (i) our algorithms will combine privacy-preserving properties of differential privacy with security offered by cryptography and (ii) based on models of information flows in integrated data handling processes, we will build more refined models analyzing the implications of repeated querying. We will demonstrate the utility of our new theory and algorithms by proposing strategies to realistically apply them in significant real-world problems illustrated through use cases in the medical domain

## 8.4 Regional initiatives

At the regional level, we participate to the *Data Advanced data science and technologies* project (CPER Data). This project is organized following three axes: internet of things, data science, high performance computing. MAGNET is involved in the data science axis to develop machine learning algorithms for big data, structured data and heterogeneous data. The project MyLocalInfo is an open API for privacy-friendly collaborative computing in the internet of things. MAGNET will also participate to the CPER Cornelia.

MAGNET also has various collaborations with research groups in linguistics and psycholinguistics at Université de Lille, in particular UMR STL (with an ongoing joint ANR project) and UMR SCALab (co-supervision of students).

MAGNET has also started collaborations with Centre Hospitalier Universitaire de Lille (CHU) in the context of the Inria "Action Exploratoire" FLAMED (Federated Learning and Analytics on Medical Data) led by AURÉLIEN BELLET.

# 9 Dissemination

## 9.1 Promoting scientific activities

### 9.1.1 Scientific events: organisation

- AURÉLIEN BELLET co-organized the Privacy Preserving Machine Learning (PPML'20) workshop at NeurIPS'20.[2]
- Since May 2020, AURÉLIEN BELLET is co-organizing the Federated Learning One World webinar[3] (700+ registered attendees)

### 9.1.2 Scientific events: selection

**Member of the conference program committees**

- AURÉLIEN BELLET served as Area Chair for ICML'20 and NeurIPS'20 and as PC member for AISTATS'21, TrustworthyML@ICLR'20, FL@IJCAI'20, PPAI@AAAI'21, CAp'20.
- RÉMI GILLERON served as PC member for ICML'20, ICLR'20& '21, AISTATS'20& '21, CAp'20, and NeurIPS'20.
- JAN RAMON served as PC member for AAAI'20, AISTATS'20 & '21, Bigdata'20, DS'20, ECML/PKDD'20, GEM@ECML'20, IEEE-ICDM'20, ICML'20, IJCAI'20, ILP'20, LOD'20, MLG'19, MIDI'20, NeurIPS'20, PPML@NeurIPS'20, SDM'20 & '21.
- JAN RAMON was member and action editor (since 10/2020) of the editorial board of Data mining and knowledge discovery (DMKD), member of the editorial board of Machine learning journal (MLJ) and member of the guest editorial board for the journal track of the ECML/PKDD conference.
- MICHAËL PERROT served as PC member for ICML'20, IJCAI'20 and NeurIPS'20.
- MARC TOMMASI served as PC member for ICML'20, CAp'20.
- PASCAL DENIS served as Area Chair at ACL'20 and TALN'20, as well PC member for COLING'20, EMNLP'20, CODI@COLING'20. As of 2020, PASCAL DENIS is standing reviewer for Transactions of the Association for Computational Linguistics (TACL).
- MIKAELA KELLER served as PC member for CAp'20

### 9.1.3 Invited talks

- AURÉLIEN BELLET gave invited talks at the Federated Learning Winter School,[4] the Google Workshop on Federated Learning and Analytics, the Applied Machine Learning Days,[5] the CIRM Conference on Optimization for Machine Learning,[6], the International Workshop on Distributed Cloud Computing[7] and the International Workshop on Privacy and Personal Data Protection [8].

[2] https://ppml-workshop.github.io/ppml20/

[3] https://sites.google.com/view/one-world-seminar-series-flow/

[4] https://sites.google.com/view/federatedlearning-workshop

[5] https://appliedmldays.org/events/amld-epfl-2020

[6] https://conferences.cirm-math.fr/2133.html

[7] http://dcc2020.ec.tuwien.ac.at/

[8] https://ict4v.org/es/workshop-2020

- FABIO VITALE gave an invited talk at ISI Foundation (Turin, Italy). Title: “Compressing graph information for binary node classification”. February, 2020.
- FABIO VITALE gave an invited talk at Technical University of Vienna (Austria). Title: “Fast Clustering through Pairwise Similarity Information”. September, 2020.

### 9.1.4 Scientific expertise

- AURÉLIEN BELLET was a member of the jury for the Gilles-Kahn PhD award of the French Society of Computer Science (SIF), sponsored by the French Academy of Sciences.[9]
- AURÉLIEN BELLET was a memberof the recruitment committee of an associate professor at Ecole Normale Supérieur de Lyon.
- JAN RAMON was a member of the jury for the European Young Researchers Award (EYRA) of Euroscience.
- PASCAL DENIS served as CoCNRS representative for the evaluation panel of Haut Conseil de l’évaluation de la recherche et de l’enseignement supérieur, HCERES[10].
- MARC TOMMASI was a member of the recruitment committee of full professors at Marseille and Saint-Etienne.
- MARC TOMMASI served as an expert for ANRT
- PASCAL DENIS acted as vice-president (before resigning) of the Inria Lille Search committee for Junior Researchers CRCN/IFSP. He was also member of search committee for a MdC position at Université Paris-Diderot, Linguistics Department.

### 9.1.5 Research administration

- MARC TOMMASI is co-head of the DatInG group (4 teams, about 100 persons), member of the Conseil Scientifique du laboratoire CRIStAL and member of the Commission mixte CRIStAL/Faculty of Science, Université de Lille.
- MARC TOMMASI is Directeur d’Études for the Master of Computer Science, Opt. Machine Learning.
- PASCAL DENIS served as an elected member of the Comité National du CNRS, section 34 (Sciences du Langage).
- PASCAL DENIS is a member of the CNRS GDR NLP Group.
- JAN RAMON was a member of the INRIA-Lille committee emploi-recherche (CER).
- AURÉLIEN BELLET is member of the Operational Committee for the assesment of Legal and Ethical risks (COERLE).
- PASCAL DENIS is administrator of Inria membership to Linguistic Data Consortium (LDC).

## 9.2 Teaching - Supervision - Juries

### 9.2.1 Teaching

- Licence SCE: FABIO VITALE, Apprentissage et émergence des comportements, 30h, L2, Université de Lille.
- Licence MIASHS: FABIO VITALE, Algorithmique et programmation, 42h, L3, Université de Lille.
- Licence SoQ: FABIO VITALE, Bases de données et SQL, 16h, L1, Université de Lille.

[9] https://www.societe-informatique-de-france.fr/recherche/prix-de-these-gilles-kahn/

[10] https://www.hceres.fr/en

- Licence MIASHS: Fabio Vitale, Introduction aux bases de données, 18h, L3, Université de Lille.
- Licence SHS: Marc Tommasi, Data Science, 24h, L2, Université de Lille.
- Licence MIASHS: Marc Tommasi, Python Programming, 48h, L2 Université de Lille.
- Licence MIASHS: Marc Tommasi, Data Science, 24h, L2 Université de Lille.
- Licence MIASHS: Mikaela Keller, Python II, 40h, L2, Université de Lille.
- Licence MIASHS: Mikaela Keller, Traitement de données, 42h, L2, Université de Lille.
- Licence SoQ (SHS): Mikaela Keller, Algorithmique de graphes, 24h, L3, Université de Lille.
- Master MIASHS: Marc Tommasi, Python Programming, 45h, M1 Université de Lille.
- Master MIASHS: Mikaela Keller, Algorithmes fondamentaux de la fouille de données, 60h, M1, Université de Lille.
- Master Computer Science: Marc Tommasi, Machine Learning, 48h, M1, Université de Lille.
- Master Data Science: Marc Tommasi Seminars 24h.
- Master Data Science: Marc Tommasi Reading groups 24h.
- Master Data Science: Fabio Vitale, Algorithms and their complexity, 30h, M1, Ecole Centrale de Lille.
- Master Data Science: Aurélien Bellet, Privacy Preserving Machine Learning, 24h, M2, Université de Lille and Ecole Centrale de Lille.
- Master Data Analysis & Decision Making: Aurélien Bellet, Machine Learning, 12h, Ecole Centrale de Lille.
- Master / Master Spécialisé Big Data: Aurélien Bellet, Advanced Machine Learning, 15h, Télécom ParisTech.
- Formation continue (Certificat d'Études Spécialisées Data Scientist): Aurélien Bellet, Supervised Learning and Support Vector Machines, 14h, Télécom ParisTech.
- Master Informatique: Pascal Denis, Foundations of Machine Learning, 46h, M1, Université de Lille.
- Master SCE: Michaël Perrot, Machine Learning for Cognitive Sciences, 30h, M2, Université de Lille.

### 9.2.2 Supervision

- Postdoc: Mohamed Maouche, supervised by Aurélien Bellet, Marc Tommasi, Privacy attacks on representation learning for speech processing, since November 2019.
- PhD in progress: Onkar Pandit, Graph-based Semi-supervised Linguistic Structure Prediction, since Dec. 2017, Pascal Denis, Marc Tommasi and Liva Ralaivola (Aix-Marseille Université).
- PhD in progress: Mariana Vargas Vieyra, Adaptive Graph Learning with Applications to Natural Language Processing, since Jan. 2018. Pascal Denis and Aurélien Bellet and Marc Tommasi.
- PhD in progress: Brij Mohan Lal Srivastava, Representation Learning for Privacy-Preserving Speech Recognition, since Oct 2018 Aurélien Bellet and Marc Tommasi and Emmanuel Vincent.
- PhD in progress: Mahsa Asadi, On Decentralized Machine Learning, since Oct 2018. Aurélien Bellet and Marc Tommasi.

- PhD in progress: NICOLAS CROSETTI, Privacy Risks of Aggregates in Data Centric-Workflows, since Oct 2018. FLORENT CAPELLI and SOPHIE TISON and JOACHIM NIEHREN and JAN RAMON.
- PhD ROBIN VOGEL, Learning to rank by similarity and performance optimization in biometric identification, 2017-2020 (CIFRE thesis with IDEMIA and Télécom ParisTech). AURÉLIEN BELLET, STÉPHAN CLÉMENÇON and ANNE SABOURIN.
- PhD in progress: MOITREE BASU, Integrated privacy-preserving AI, since 2019. JAN RAMON.
- PhD in progress: CÉSAR SABATER, Privacy Preserving Machine Learning, since 2019. JAN RAMON.
- PhD in progress: PAUL MANGOLD, supervised by AURÉLIEN BELLET and MARC TOMMASI and JOSEPH SALMON, since October 2020.
- Engineer YANNICK BOUILLARD, supervised by AURÉLIEN BELLET since Nov. 2020.
- Engineer SOPHIE VILLEROT, supervised by JAN RAMON since Nov. 2020.

### 9.2.3 Juries

- AURÉLIEN BELLET was member of the PhD jury of Kevin Elgui (Télécom Paris)
- JAN RAMON was member of the PhD jury of Robin Vandaele (Ghent, BE)
- MARC TOMMASI was member of the PhD jury of Antoine Chatalic (Rennes Univ.), Maziar Moradi Fard (Grenoble Univ.), Léo Gautheron (Saint-Etienne Univ.)
- MIKAELA KELLER was member of the PhD jury of Timothée Lacroix (Paris-Est Univ.)

# 10 Scientific production

## 10.1 Major publications

[1] A. Bellet, R. Guerraoui and H. Hendrikx. 'Who started this rumor? Quantifying the natural differential privacy guarantees of gossip protocols'. In: *DISC 2020 - 34th International Symposium on Distributed Computing*. Freiburg / Virtual, Germany, Oct. 2020. URL: https://hal.inria.fr/hal-02166432.

[2] A. Bellet, R. Guerraoui, M. Taziki and M. Tommasi. 'Personalized and Private Peer-to-Peer Machine Learning'. In: *AISTATS 2018 - 21st International Conference on Artificial Intelligence and Statistics*. Lanzarote, Spain, Apr. 2018, pp. 1–20. URL: https://hal.inria.fr/hal-01745796.

[3] M. Dehouck and P. Denis. 'Delexicalized Word Embeddings for Cross-lingual Dependency Parsing'. In: *EACL*. Vol. 1. EACL 2017. Valencia, Spain, Apr. 2017, pp. 241–250. DOI: 10.18653/v1/E17-1023. URL: https://hal.inria.fr/hal-01590639.

[4] M. Dehouck and P. Denis. 'Phylogenetic Multi-Lingual Dependency Parsing'. In: *NAACL 2019 - Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, United States, June 2019. URL: https://hal.archives-ouvertes.fr/hal-02143747.

[5] O. Kuželka, Y. Wang and J. Ramon. 'Bounds for Learning from Evolutionary-Related Data in the Realizable Case'. In: *International Joint Conference on Artificial Intelligence (IJCAI)*. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) 2016. New York, United States, July 2016. URL: https://hal.archives-ouvertes.fr/hal-01422033.

[6] E. Lassalle and P. Denis. 'Joint Anaphoricity Detection and Coreference Resolution with Constrained Latent Structures'. In: *AAAI Conference on Artificial Intelligence (AAAI 2015)*. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015). Austin, Texas, United States, Jan. 2015. URL: https://hal.inria.fr/hal-01205189.

[7] C. Pelekis, J. Ramon and Y. Wang. 'H\"older-type inequalities and their applications to concentration and correlation bounds'. In: *Indagationes Mathematicae* 28.1 (2017), pp. 170–182. DOI: 10.1016/j.indag.2016.11.017. URL: https://hal.archives-ouvertes.fr/hal-01421953.

[8] T. Ricatte, R. Gilleron and M. Tommasi. 'Skill Rating for Multiplayer Games Introducing Hypernode Graphs and their Spectral Theory'. In: *Journal of Machine Learning Research* 21 (2020), pp. 1–18. URL: https://hal.inria.fr/hal-02566930.

[9] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi and E. Vincent. 'Evaluating Voice Conversion-based Privacy Protection against Informed Attackers'. In: *ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing*. IEEE Signal Processing Society. Barcelona, Spain, May 2020, pp. 2802–2806. URL: https://hal.inria.fr/hal-02355115.

[10] P. Vanhaesebrouck, A. Bellet and M. Tommasi. 'Decentralized Collaborative Learning of Personalized Models over Networks'. In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. Fort Lauderdale, Florida., United States, Apr. 2017. URL: https://hal.inria.fr/hal-01533182.

[11] F. Vitale, N. Parotsidis and C. Gentile. 'Online Reciprocal Recommendation with Theoretical Performance Guarantees'. In: *NIPS 2018 - 32nd Conference on Neural Information Processing Systems*. Montreal, Canada, Dec. 2018. URL: https://hal.inria.fr/hal-01916979.

## 10.2 Publications of the year

### International journals

[12] W. De Vazelhes, C. Carey, Y. Tang, N. Vauquier and A. Bellet. 'metric-learn: Metric Learning Algorithms in Python'. In: *Journal of Machine Learning Research* 21 (2020), pp. 1–6. URL: https://hal.inria.fr/hal-03100076.

[13] P. Mangold, A. Filiot, M. Moussa, V. Sobanski, G. Ficheur, P. Andrey and A. Lamer. 'A Decentralized Framework for Biostatistics and Privacy Concerns'. In: *Studies in Health Technology and Informatics* (23rd Nov. 2020). DOI: 10.3233/shti200710. URL: https://hal.inria.fr/hal-03110739.

[14] T. Ricatte, R. Gilleron and M. Tommasi. 'Skill Rating for Multiplayer Games Introducing Hypernode Graphs and their Spectral Theory'. In: *Journal of Machine Learning Research* 21 (2020), pp. 1–18. URL: https://hal.inria.fr/hal-02566930.

### International peer-reviewed conferences

[15] J. Bell, A. Bellet, A. Gascón and T. Kulkarni. 'Private Protocols for U-Statistics in the Local Model and Beyond'. In: AISTATS 2020 - 23rd International Conference on Artificial Intelligence and Statistics. Palermo, Italy, 26th Aug. 2020. URL: https://hal.inria.fr/hal-02310236.

[16] A. Bellet, R. Guerraoui and H. Hendrikx. 'Who started this rumor? Quantifying the natural differential privacy guarantees of gossip protocols'. In: DISC 2020 - 34th International Symposium on Distributed Computing. Freiburg / Virtual, Germany, 2020. URL: https://hal.inria.fr/hal-02166432.

[17] M. Maouche, B. M. L. Srivastava, N. Vauquier, A. Bellet, M. Tommasi and E. Vincent. 'A comparative study of speech anonymization metrics'. In: INTERSPEECH 2020. Shanghai, China, 25th Oct. 2020. URL: https://hal.inria.fr/hal-02907918.

[18] B. Paige, J. Bell, A. Bellet, A. Gascón and D. Ezer. 'Reconstructing Genotypes in Private Genomic Databases from Genetic Risk Scores'. In: 24th International Conference On Research In Computational Molecular Biology (RECOMB 2020). Virtual, Italy, 2020. DOI: 10.1101/2020.01.15.907808. URL: https://hal.inria.fr/hal-03100032.

[19] B. M. L. Srivastava, N. Tomashenko, X. Wang, E. Vincent, J. Yamagishi, M. Maouche, A. Bellet and M. Tommasi. 'Design Choices for X-vector Based Speaker Anonymization'. In: INTERSPEECH 2020. Shanghai, China, 25th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02610447.

[20] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi and E. Vincent. 'Evaluating Voice Conversion-based Privacy Protection against Informed Attackers'. In: ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing. Barcelona, Spain, 4th May 2020, pp. 2802–2806. URL: https://hal.inria.fr/hal-02355115.

[21] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé and M. Todisco. 'Introducing the VoicePrivacy initiative'. In: INTERSPEECH 2020. Shanghai, China, 25th Oct. 2020. URL: https://hal.inria.fr/hal-02562199.

[22] M. Vargas-Vieyra, A. Bellet and P. Denis. 'Joint Learning of the Graph and the Data Representation for Graph-Based Semi-Supervised Learning'. In: 14th Workshop on Graph-Based Natural Language Processing (TextGraphs 2020). Virtual, Spain, 2020. URL: https://hal.inria.fr/hal-03100039.

[23] V. Zantedeschi, A. Bellet and M. Tommasi. 'Fully Decentralized Joint Learning of Personalized Models and Collaboration Graphs'. In: AISTATS 2020 - The 23rd International Conference on Artificial Intelligence and Statistics. Palerme / Virtual, Italy, 26th Aug. 2020. URL: https://hal.inria.fr/hal-03100057.

### Conferences without proceedings

[24] O. Pandit, P. Denis and L. Ralaivola. 'Integrating knowledge graph embeddings to improve mention representation for bridging anaphora resolution'. In: CRAC 2020 - Third Workshop on Computational Models of Reference, Anaphora and Coreference. Virtual, France, 12th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03001157.

[25] J. Ramon and M. Basu. 'Interpretable privacy with optimizable utility'. In: ECML/PKDD 2020 - Workshop on eXplainable Knowledge Discovery in Data mining. Ghent / Virtual, Belgium, 14th Sept. 2020. URL: https://hal.inria.fr/hal-02950994.

### Scientific book chapters

[26] O. Bournez, G. Dowek, R. Gilleron, S. Grigorieff, J.-Y. Marion, S. Perdrix and S. Tison. 'Theoretical Computer Science: Computability, Decidability and Logic'. In: *A Guided Tour of Artificial Intelligence Research - Volume III: Interfaces and Applications of Artificial Intelligence (10.1007/978-3-030-06170-8)*. Springer International Publishing, 2020, pp. 1–50. DOI: 10.1007/978-3-030-06170-8_1. URL: https://hal.archives-ouvertes.fr/hal-03173193.

[27] O. Bournez, G. Dowek, R. Gilleron, S. Grigorieff, J.-Y. Marion, S. Perdrix and S. Tison. 'Theoretical Computer Science: Computational Complexity'. In: *A Guided Tour of Artificial Intelligence Research - Volume III: Interfaces and Applications of Artificial Intelligence (10.1007/978-3-030-06170-8)*. Springer International Publishing, 2020. URL: https://hal.archives-ouvertes.fr/hal-02995771.

### Reports & preprints

[28] A. Bellet, P. Denis, R. Gilleron, M. Keller and N. Vauquier. *For more transparency in the automatic analysis of public consultations: lessons learned from the French "Grand Débat National"*. 23rd Oct. 2020. URL: https://hal.inria.fr/hal-02860659.

[29] E. Cyffers and A. Bellet. *Privacy Amplification by Decentralization*. 2020. URL: https://hal.inria.fr/hal-03100005.

[30] B. M. L. Srivastava, M. Maouche, M. Sahidullah, E. Vincent, A. Bellet, M. Tommasi, N. Tomashenko, X. Wang and J. Yamagishi. *Privacy and utility of x-vector based speaker anonymization*. 15th Apr. 2021. URL: https://hal.inria.fr/hal-03197376.

[31] R. Vogel, A. Bellet and S. Clémençon. *Learning Fair Scoring Functions: Bipartite Ranking under ROC-based Fairness Constraints*. 2020. URL: https://hal.inria.fr/hal-03100014.

**Other scientific publications**

[32] C. Sabater, A. Bellet and J. Ramon. *Distributed Differentially Private Averaging with Improved Utility and Robustness to Malicious Parties*. Vancouver (Virtual Workshop), Canada, 11th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03117816.

[33] C. Sabater, A. Bellet and J. Ramon. *Échange de bruit corrélé pour le calcul distribué de moyenne avec garanties de confidentialité différentielle*. Vannes (Virtual), France, 23rd June 2020. URL: https://hal.archives-ouvertes.fr/hal-03117907.

## 10.3 Cited publications

[34] A. E. Elo. *The Rating of Chess Players, Past and Present*. Arco Publishing, 1978.

[35] R. Herbrich, T. Minka and T. Graepel. 'TrueSkill$^{TM}$: A Bayesian Skill Rating System'. In: *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*. Ed. by B. Schölkopf, J. C. Platt and T. Hofmann. MIT Press, 2006, pp. 569–576. URL: https://proceedings.neurips.cc/paper/2006/hash/f44ee263952e65b3610b8ba51229d1f9-Abstract.html.

[36] K. Markert, Y. Hou and M. Strube. 'Collective Classification for Fine-grained Information Status'. In: *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8–14, 2012, Jeju Island, Korea – Volume 1: Long Papers*. The Association for Computer Linguistics, 2012, pp. 795–804. URL: https://www.aclweb.org/anthology/P12-1084/.

[37] I. Rösiger. 'BASHI: A Corpus of Wall Street Journal Articles Annotated with Bridging Links'. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7–12, 2018*. Ed. by N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis and T. Tokunaga. European Language Resources Association (ELRA), 2018. URL: http://www.lrec-conf.org/proceedings/lrec2018/summaries/84.html.