**AAP Contrats doctoraux en Intelligence artificielle**

**Cofinancé par l'ANR**

# BigSF: Developing innovative machine learning methods to build an empirical model of the galactic star formation

## 1. DESCRIPTION OF THE PHD THESIS PROJECT

**(Up to 4 pages)**

### 1.1 OBJECTIVES OF THE PROJECT BASED ON THE CURRENT STATE OF THE ART

*(State how the project is relevant to the research axis.)*

The formation of high mass stars is one of the pending issues in modern astrophysics. These stars are rare and evolve rapidly, making their early study more difficult. They have a profound impact on their surrounding and are a key ingredient of galaxies' evolution. The fundamental properties of star formation in galaxies such as the rate and the efficiency at which they form, mainly controlled by high mass stars, are used to describe galaxies' evolution. However, the lack of spatial resolution of observations in distant galaxies make the derivation of these quantities highly uncertain. In particular, the way the efficiency and the rate of star formation change as a function of the environment is still strongly debated.

Our Galaxy has been observed on a large range of spectral domains at various spatial and spectral resolutions. This results in a panchromatic data goldmine in which the study of star formation plays a central role. The advent of the combination of Big Data (BD) and Machine Learning (ML) makes that we are in a position where the fundamental properties of star formation, on all spatial scales, can be unveiled in our Galaxy. However, the data complexity (multi wavelength, multi resolution and of different nature, i.e. images and spectra) makes this quest challenging. Specific BD handling and Innovative ML techniques have to be developed in order to reach this objective. This is the aim of the BigSF project.

BigSF is an interdisciplinary PhD project. Using the unique opportunity offered by the existence of BD on our Galaxy and the strongly rising importance of the ML in fundamental research, BigSF proposes to mix the two approaches and to develop new tools adapted for an innovative research in Galactic Star Formation. The ultimate aim of the BigSF project is to build and deliver an empirical model for star formation in our Galaxy.

Star formation remains one of the top unsolved questions in astrophysics. In particular, the formation of massive stars is key because these stars control the evolution of galaxies and played a central role in the re-ionization of the Universe. From a machine learning point of view, this requires a synergy between learning methods and physical models. This is challenging since the most powerful ML methods are quasi black-boxes, leading to the need of explainable or interpretable methods designed for this task.

In our Galaxy the wealth of existing data allows to envision a global approach of the star formation process. However, the data heterogeneity renders this approach very challenging. Indeed, data are obtained with different instruments, at different wavelength and different resolution. Moreover, their nature is different. The

data can be 2-dimension images or 3-dimension spectra of different molecules. On one hand, the richness of the available data offers a unique opportunity to tackle, for the first time, the fundamental process of star formation in our Galaxy. On the other hand, the complexity of the database makes this research project presently unreachable using the current available tools. The only way to envision this research program is to put together theoretical, observational and practical expertise in the involved key scientific domains: ML, BD and Star Formation. This is what this thesis proposes to do.

## 1.2 METHODOLOGY

*(Proposed methodology for the thesis project)*

The volume and heterogeneity of the available data creates both the richness and challenge of the BigSF project. Tackle the question of star formation properties in the Galaxy as a function of the environment with ML requires development of new ML algorithms, in particular by exploiting the new paradigm of deep learning to extract more information on both tabular and imaging data. Indeed, current ML algorithms cannot deal with BD and multi view. The scope of this part of the project is immense as these new developments will benefit a broad range of fields, well beyond academic research, mainly driven by the unprecedented impact of BD and ML in our society.

The proposed methodology for BigSF combines both approaches and expertise of all partners: the exploration and organization of the available big data sets on star formation in the Galaxy analysed with rising level of complexity ML algorithms to reveal new patterns of the star formation processes.

On the ML side, we propose to extend the most recent methods in both supervised and unsupervised learning for big datasets and test some multi view learning methods for managing the high heterogeneity of the data. We will focus on three challenges that must be tackled: 1) mixing data of different resolutions (form different wavelength) from several instruments for one given task, 2) building explainable, in an astrophysical sense, neural networks and combining them with expertise from astrophysicists and 3) dealing with missing data (highly dependent of instrument sensors). These challenges require new architectures and/or new training procedures that developments will be at the heart of this PhD project.

On the astrophysical side, we expect the first tested methods to give insights on how the environment influences the star formation, using homogeneous datasets. Then, with the new methods, we expect the hypothesis to be more and more precise using bigger datasets. At the end, all the results will be compared with different methods in order to avoid ill interpretation and extract the main common clues.

The first step of the thesis will be to organize the data access. In the meantime, machine learning techniques will be explored and developed to be trained on single-view datasets (one type of astrophysical object, image at one wavelength, and taken with a single instrument at one spatial resolution). We will then increase the complexity of the dataset, considering different type of astrophysical objects and introducing multi view data (images and spectra at different wavelength and different resolutions). The first year of the PhD will be dedicated to this part.

In parallel to this project's implementation during the first year, the second step of the PhD will be dedicated to the supervised and unsupervised application of machine learning techniques to the global dataset. One specific task will be on adding knowledge inside current ML methods while handling the biases. This will allow to reveal the fundamental properties of star formation in the Galactic Plane and how these properties are linked to the environment. This part about learning will last for 8 months.

The third step will be dedicated to the analysis of the results. The revealed patterns will be characterized. This last part will last for 10 months. We estimate the available data volume on the Galactic Plane to be around 5000 To.

The main results at each step will be published in referee journals with high impact in both disciplines. We foresee a minimum of 3 papers in the course of this PhD: one about the project presentation, one about the used and developed machine learning techniques and one about the results and deliverables. These papers are the PhD milestones. Moreover, the results from innovative ML developments obtained in this project will be useful for similar kind of data, like medical observations, where having a physical model can help.

We point out that this PhD is proposed in an international context with collaborators in Rome (part of the CIVIS European University) and Naples and that it will benefit from extra research funding from the Institut Universitaire de France.

## 1.3 WORK PLAN

*(Including a chronogram of the activities or Gantt chart)*

The first step of the project will be to organize the data access. In the meantime, machine learning techniques will be explored and developed to be trained on single-view datasets. The first year of the PhD will be dedicated to this part.

In parallel to this project's implementation during the first year, the second step of the PhD will be dedicated to the supervised and unsupervised application of machine learning techniques to the global dataset. This will allow to reveal the fundamental properties of star formation in the Galactic Plane and how these properties are linked to the environment. This part about learning will last for 8 months.

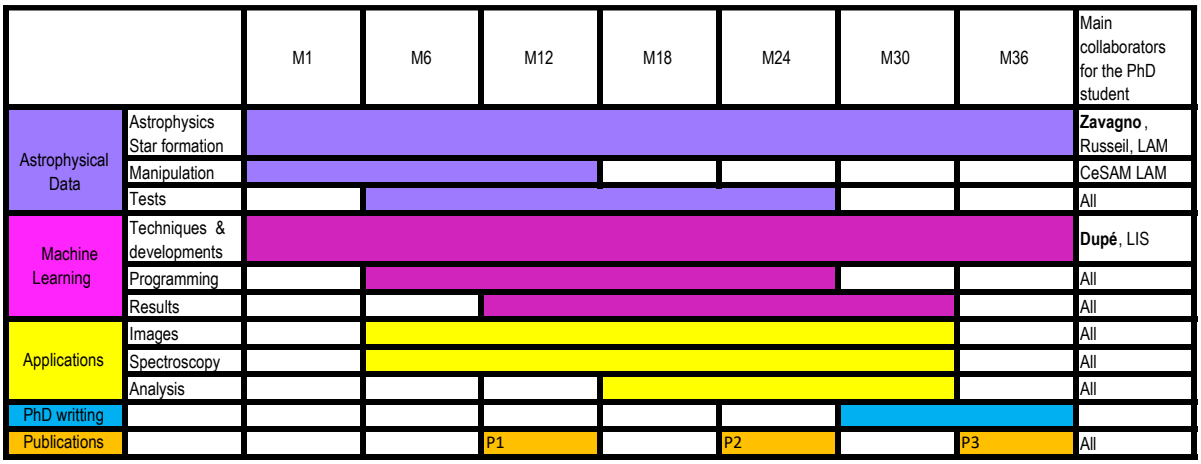
The third step will be dedicated to the analysis of the results. The revealed patterns will be characterized. This last part will last for 10 months. We estimate the available data volume on the Galactic Plane to be around 5000 To. With this project's outputs we aim at delivering an empirical model of the Galactic star formation, derived using Big Data and Machine Learning.

The main results at each step will be published in referee journals with high impact in both disciplines. We foresee a minimum of 3 papers during this PhD: one about the project presentation, one about the used and developed machine learning techniques and one about the results and deliverables. These papers are the PhD milestones.

At each step, the PhD student will be guided by the two supervisors and their main collaborators.

The proposed project is part of the research program for the two PhD advisors and their teams. This ensures a rapid progression of the project where the PhD student will play the central role. The Gantt chart summarizes the expected PhD progression.

The PhD advisors and their team members all have experience in managing research projects. In addition to the student's supervision, monthly common progress meetings will be organized (at LAM or LIS) to follow the general progression of the project.




## 1.4 SUPERVISOR AND RESEARCH GROUP DESCRIPTION

*(Are there other funds for the research? Is the project part of a larger research programme?)*

The project is included in larger research programmes ongoing in each team at LIS and LAM. Funds have been allocated already (ANR programmes, European programme) and have been requested. For example, we (A. Zavagno and F.-X. Dupé) had a funding from the CNRS PEPS Astroinformatics and bought a dedicated machine for this project. The selection of the project by this CNRS Interdisciplinary Action points out his high scientific potential.

The Qarma team of LIS is a research team with strong interest in machine learning and its application. They are 10 permanents (with 1 full professors) and 5 PhD students. This thesis will be on behalf of two ANR projects: Lives (ANR-15-CE23-0026) and Deep In France (ANR-16-CE23-0006). The first project is about the theoretical framework and possibility of multi-view learning, i.e. learning highly heterogeneous data (e.g. text, image, sounds, …). The second project aims to develop new deep learning methods and to better understand how these methods perform.

LAM has been part of the VIALACTEA European FP7 SPACE project (PI: S. Molinari, INAF-Rome). This project has been allocated 2.5 M€ (2013-2016) to exploit the scientific output of the Herschel Hi-GAL survey consisting in the complete imaging survey of the Galactic Plane in five infrared bands using the infrared space observatory Herschel. In this large consortium, Prof. Brescia, Drs. Longo and Pasian (INAF - Osservatorio Astronomico di Capodimonte, Naples) are experts in machine learning techniques adapted to astrophysics and are associated with this project. As active and central members of the VIALACTEA consortium, we benefit from all the developments and results achieved in this project.

# 2. RECENT PUBLICATIONS

*(Insert your team's most recent publications connected to the thesis project; the idea is to give the candidate reading material for a better understanding of your team's work)*

Zavagno, A., André, P., Schuller, F. et al. 2020, A&A in press, The role of Galactic HII regions in the formation of filaments. High resolution submillimeter imaging of RCW120 with ArTéMiS (arXiv 2004.05604)
Zhang, S., Zavagno, A., Yuan, J. et al. 2020, A&A in press, HII regions and high-mass starless clump candidates I: Catalogs and properties (arXiv 2003.11433 )
Xu, J-L, Zavagno, A., Yu, N. et al. 2019, A&A, 627, A27 "The effect of ionization feedback on star formation: a case study of the M16 HII region"

Palmerim, P., Zavagno, A., Elia, D. et al. 2017 A&A, 605, A35 “Spatial distribution of star formation related to ionized regions throughout the inner Galactic plane”
Brescia, M., Cavuoti, S., Longo, G. et al. 2014, PASP, 126, 783 “DAMEWARE: A Web Cyberinfrastructure for Astrophysical Data Mining”
Kadri, H., Ayache, S., Capponi, C., Koço, S., Dupé, F. X., & Morvant, E. (2013). The multi-task learning view of multimodal data. In Asian Conference on Machine Learning.
Dupé, F.- X., & Anthoine, S. (2018). Generalized greedy alternatives. Applied and Computational Harmonic Analysis.
Sellami, A., Dupé, F. X., Cagna, B., Kadri, H., Ayache, S., Artières, T., & Takerkart, S. (2020). Mapping individual differences in cortical architecture using multi-view representation learning. IEEE IJCNN 2020.

## 3. EXPECTED PROFILE OF THE CANDIDATE

*(Insert expected profile: interests, academic background, skills. Approximately half a page.)*

The selected candidate should possess an academic background and skills related to the central knowledge needed for this PhD thesis: statistics, signal and/or image processing, Machine Learning, Big Data and Astrophysics. Note that, because the central topic of this PhD project is the development of innovative Machine Learning algorithms for Astrophysics, knowledge in Astrophysics is not mandatory but will be considered as an added value for selecting the PhD candidate. The project requires a significant part of computation and software development in Python. Skills and strong interest in these fields are required for this PhD.

Because the proposed project involves many partners including a non-academic one, the candidate will have to demonstrate that he/she can work in groups and international collaborations and is open-minded.

Minorities are encouraged to apply.

## 4. SUPERVISORS’ PROFILE

- *(Insert a short professional profile of the SUPERVISOR)*
- *(Please mention the number of theses currently being supervised, and their starting date.)*

Professional profiles
Annie Zavagno
Full professor AMU
Senior Member of the Institut Universitaire de France since October 2017 (for 5 years)
Habilitation à Diriger des Recherches (HDR) obtained in May 2002
5 already supervised PhD Thesis, 1 started in October 2017 (Defense in September 2020)

| Name of the PhD student | Supervision | Year of PhD defense | PhD duration (months) | Peer-reviewed publications | Current status |
|---|---|---|---|---|---|
| M. Pomarès | Full | 2009 | 38 | 3 | Physics Teacher |
| J. Tigé | Co-supervision | 2014 | 40 | 2 | Programmer |
| H.-L. Liu | Co-supervision | 2016 | 36 | 4 | Post-doc |

| | | | | | |
|---|---|---|---|---|---|
| M. Figueira | Full | 2017 | 36 | 4 | Post-doc |
| A. Bernard | Co-supervision | 2017 | 36 | 3 | Post-doc |
| S. Zhang | Full | 2020 | | | |

Publications: 168 articles

A-ranked (referee publications): 121, B-ranked: 47
h-index: 47
Number of current PhD: 1, starting date October 2017 – Siju Zhang
A. Zavagno has also trained 7 post-doctoral students (2010-2017) working on the scientific results of the *Herschel* space mission and 5 research engineers.
Main research interests: star formation, impact of massive stars on star formation, multi scale and multi wavelength analysis

François-Xavier Dupé
Lecturer AMU
Main research interests: sparse constrained optimization, multi-view machine learning
Articles: A-ranked (referee publications): 6, B-ranked: 5
h-index: 10
F-X Dupé will defend his HDR in the next three years. The supervision of this PhD is key to meet this goal.

## VISA DU RESPONSABLE DE L'INSTITUT ET DU DIRECTEUR DE LABORATOIRE CONCERNÉS

**Visa du responsable de l'institut, NOM Prénom**

Emmanuel GODARD

**Fait à Marseille, le** 15/05/2020

**Signature**

**Visa du directeur du laboratoire, NOM Prénom**

**Fait à Marseille, le**

**Signature**