SORBONNE UNIVERSITÉ

PROGRAMME INTITUTS ET INITIATIVES

Appel à projet – campagne 2021

Proposition de projet de recherche doctoral (PRD)

SCAI - Sorbonne Center of Artificial Intelligence

Intitulé du projet de recherche doctoral (PRD): Machine Learning and the Cosmic Dawn

Directeur.rice de thèse porteur.euse du projet (titulaire d'une HDR) :

NOM : **Semelin** Prénom : **Benoit**
Titre : Professeur des Universités ou
e-mail : benoit.semelin@obspm.fr
Adresse professionnelle : LERMA, 77 av Denfert Rochereau, 75014 Paris
(site, adresse, bât., bureau)

Unité de Recherche :
Intitulé : LERMA
Code *(ex. UMR xxxx)* : UMR 8112

École Doctorale de rattachement de l'équipe (future école doctorale du.de la doctorant.e) : **ED127-AstronomieAstrophysiqueIdF**

Doctorant.e.s actuellement encadré.e.s par la.e directeur.rice de thèse (préciser le nombre de doctorant.e.s, leur année de 1e inscription et la quotité d'encadrement) : 1. doctorant, première inscription 2018, 100%

Co-encadrant.e :

NOM : Prénom :
Titre : Choisissez un élément : ou HDR ☐
e-mail :

Unité de Recherche :
Intitulé :
Code *(ex. UMR xxxx)* :

École Doctorale de rattachement : **Choisissez un élément :**
Ou si ED non Alliance SU :

Doctorant.e.s actuellement encadré.e.s par la.e co-directeur.rice de thèse (préciser le nombre de doctorant.e.s, leur année de 1e inscription et la quotité d'encadrement) :

Co-encadrant.e :

NOM : Prénom :
Titre : Choisissez un élément : ou HDR ☐
e-mail :

Unité de Recherche :
Intitulé :
Code *(ex. UMR xxxx)* :

École Doctorale de rattachement : **Choisissez un élément :**
Ou si ED non Alliance SU :

Doctorant.e.s actuellement encadré.e.s par la.e co-directeur.rice de thèse (préciser le nombre de doctorant.e.s, leur année de 1e inscription et la quotité d'encadrement) :

Cotutelle internationale : ☐ Non ☐ Oui, précisez Pays et Université :

Selon vous, ce projet est-il susceptible d'intéresser une autre Initiative ou un autre Institut ?
☐ Non ☐ Oui, précisez Choisissez l'institut ou l'initiative :

SORBONNE UNIVERSITÉ

Description du projet de recherche doctoral *(en français ou en anglais) :*

Ce texte sera diffusé en ligne : il ne doit pas excéder 3 pages et est écrit en interligne simple.

Détailler le contexte, l'objectif scientifique, la justification de l'approche scientifique ainsi que l'adéquation à l'initiative/l'Institut.

Le cas échéant, préciser le rôle de chaque encadrant ainsi que les compétences scientifiques apportées. Indiquer les publications/productions des encadrants en lien avec le projet.
Préciser le profil d'étudiant(e) recherché.

Context:

From 2027 on, the Square Kilometer Array radio-interferometer (SKA) will start observing the Cosmic Dawn (CD) and the Epoch of Reionization (EoR). Following the global recombination of Hydrogen in the cooling universe at redshift z~1000 (400 000 years after the Big Bang), the baryonic universe is cold and neutral. Only with the formation of the first stars at z~20-30 (i.e. during the Cosmic Dawn, 100-200 Myr after the Big Bang) are ionizing photons emitted again. As star clusters grow into primordial galaxies, bubbles of ionized gas form around them and expand into the intergalactic medium (IGM) until they overlap. Around z~6 (1 billion years after the Big Bang), the universe is completely reionized. During this first billion years, patches of cold and neutral hydrogen persist in the low-density regions of the IGM. They emit and absorb 21 cm photons through a hyperfine transition in the ground state of the atoms. Because these 21 cm photons redshift, as they travel, due to the expansion of the universe, and since the low density of the universe allows them to escape the core of the line by redshifting before being reabsorbed, they are received on Earth in a range of frequencies (50 to 200 MHz) that directly map to distances from the observer to the emission point. Thus, the tremendous sensitivity and collecting area of the SKA will allow us to perform a 3-dimensionnnal mapping of the IGM using this faint signal, unveiling a wealth of information about structure formation and the nature of the first sources of light. In the meantime, a first tentative and unexpected detection of the signal averaged over the whole sky (thus with fluctuations only along the line of sight) in a narrow redshift range around z ~17 has been performed with the EDGES instrument (Bowman et al., 2018). It needs however to be confirmed by observations with radio-interferometers able to measure angular fluctuations.

LOFAR, NenuFAR or MWA are such radio-interferometers, currently attempting a first detection of the power spectrum of the signal, a statistical quantity that benefits from a more favorable signal-to-noise ratio than a full 3D imaging. The upcoming HERA and SKA should measure this quantity with high accuracy, but only the SKA will have sufficient sensitivity to provide a full 3D imaging of the signal. Once the observed data are available, their interpretation in terms of relevant astrophysical information will be a challenge. Indeed, the local intensity of the signal depends on several properties of the hydrogen: the ionization state, the density, the velocity along the line of sight, and the kinetic temperature, but also on the local Lyman-alpha flux (see Furlanetto et al. 2006 for a review). Thus, the interpretation will rely on two building blocks. The first is an accurate modelling of the signal that will take as inputs a few parameters describing (yet) unconstrained astrophysical processes. The second building block will be inversion methods able to put constraints on those parameters using the observable quantities that encode the most information about the cosmological signal (powerspectrum, other summary statistics or the full 3D signal).

Signal dimensionality reduction using machine learning:

As the full 3D signal will suffer from a high level of instrumental noise even with the SKA, summary statistics are a good starting point for the inference of the best-fit model parameters. The powerspectrum is the most widely used statistic for this purpose. However, it ignores the non-gaussian information that we know is present in the signal from numerical simulations. Our team has an ongoing collaboration with a team a LPENS (F. Boulanger, E. Allys, F. Levrier) to explore the representation of the signal with scattering transforms (see Mallat, S., Commun. Pure Appl. Math., 2012). Yet another possibility is to employ Machine Learning (ML) to perform a dimensionality reduction on the full 3D signal and use the compressed information as a starting point for parameter inference. Usual dimensionality reduction methods, such as autoencoders, would be the obvious choice for this task. One difficulty is that our goal is not to encode the information necessary to rebuilt the exact same signal, but rather to rebuild any signal than could be emitted in a universe with a different realization of the gaussian random field describing the initial density field for matter, but with the same subsequent astrophysical processes (the parameters in our model). Identiying what part of the latent space stores the information of the gaussian random field and what part stores the information about the astrophysical processes will be necessary. This difficulty may also possibly be circumvented by using generative models (e.g. GANs). In this part of the PhD project, several ML tools to perform dimensionality reduction will be tested, and their results evaluated against other summary statistics (powerspectrum, scattering transforms…). This comparison will for example be performed by using these quantities as inputs for a code such as 21CMMC (Greig et al. 2014) to perform Bayesian MCMC inference.

Constraining model parameters using machine learning:

The Bayesian approach combined with a Markov Chain Monte Carlo (MCMC) exploration of the parameter space has been the most widely used method in the last decade to constrain model parameters using observations in astrophysics (for example cosmological parameter constraints using Cosmic Microwave Background observations). The main advantage of the method is to provide confidence contours without making assumptions about the shape of the posterior distribution of the parameters (compared to the Fisher matrix approach for example). The limiting factor is that it typically requires a large number of instances of forward modelling, to be able to explore this posterior distribution. Such an approach is not feasible if the modelling is performed with a full numerical simulation as is our case for the 21-cm signal from the Epoch of Reionisation. This is where machine learning can help. Based on a limited set of simulations (learning sample), algorithms can be trained to either emulate the forward modelling (from parameters to observable signal) or perform inverse modelling.

Our team has started exploring the potential of training various supervised learning algorithms to perform inverse modelling and predict maximum likelihood parameters values (Shimabukuro & Semelin 2017, Eames et al. 2019, Doussot et al. 2019). Even though we now reach an error level in the reconstructed parameters that is smaller than the typical amplitude of the uncertainty resulting from a SKA-like thermal noise level (as estimated with the Bayesian approach), improvement must still be achieved to make it negligible. However, the main direction in which progress is needed is to provide confidence levels, along with maximum likelihood values, to improve the interpretability of the reconstructed parameters. Bayesian Neural Networks appear a likely tool to reach that goal. Indeed, due advances in AI research, methods have been developed to quantify separately the aleatoric uncertainty, corresponding to the thermal noise in our case, from the epistemic uncertainty that comes from the imperfect training of the network and imperfect modelling among other factors.

In the course of the PhD project, the candidate will have to master this tool, apply it to a 21-cm signal learning sample, produced either with full numerical simulations or semi-numerical fast models. We will then compare the performances of the machine learning approach to the usual Bayesian approach in predicting confidence levels. Other methods to derive confidence levels using

machine learning may also be explored.

Environment of the PhD project and adequation to SCAI:

The PhD student will benefit from a lively scientific environment for this project. The collaboration with LPENS, together with Strasbourg Observatory is materialized by a ANR funding application currently under review (the KODAMA project), where 21-cm signal representation is a central axis of research. Moreover, the advisor is also the PI of the MINERVA project, funded by Observatoire de Paris, which goals is to apply Machine Learning in radio-astronomy. Three post-docs are funded by this project, all of them with some level of experience in Machine Learning. The advisor himself as a few years of experience in using machine learning in the context of the 21-cm signal from the Epoch of reionization (see publications below). Generally speaking, data science is a central subject in astrophysics where in situ experimentations are seldom possible, and thus acquiring as much data as possible trough observations is necessary. Radio-interferometers in general and the SKA in particular produce huge amounts of data (1 Po per day for SKA) and machine learning will be a necessary tool for their scientific exploitation. For these reasons, the current project should be a good fit for the goal of SCAI.

A few publications from the team on the subject:

Doussot, A., Eames, E., & Semelin, B. , MNRAS, 490, 371, 2019

Semelin, B. , MNRAS, 455, 962, 2016

Semelin, B., Eames, E., Bolgar, F., & Caillat, M., MNRAS, 472, 4508, 2017

Shimabukuro, H. & Semelin, B., MNRAS, 468, 3869, 2017

Profile of the student:

The candidate should idealy have a good background in astrophysics and cosmology and some basic knowledge of data science and machine learning.

Merci d'enregistrer votre fichier au format PDF et de le nommer :
«ACRONYME de l'initiative/institut – AAP 2021 – NOM Porteur.euse Projet »

Fichier envoyer simultanément par e-mail à l'ED de rattachement et au programme :
cd_instituts_et_initiatives@listes.upmc.fr avant le 20 février.