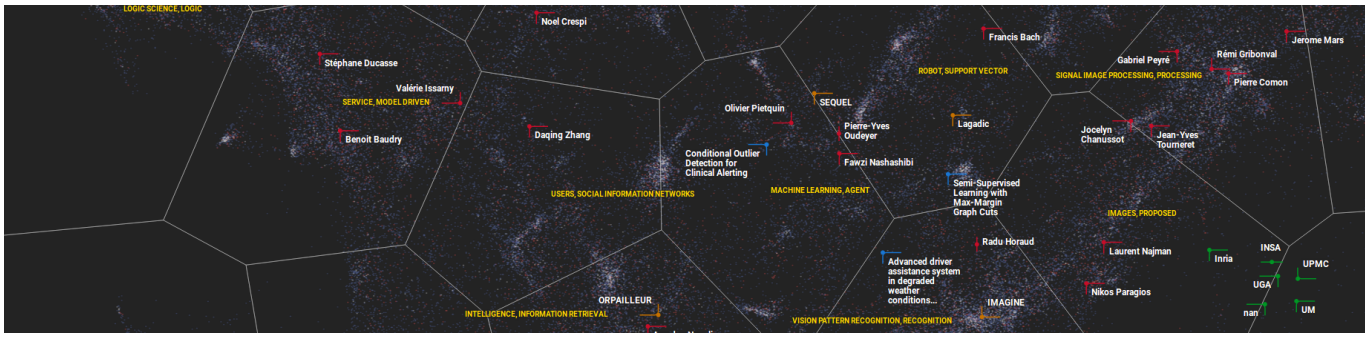
Engineer or M2 Internship offer

Hyperparameter exploration and optimization for big data visualisation.

General presentation of the topic :

Cartolabe lets explore a scientific landscape built from unsupervised machine learning algorithms. The user can search for authors, articles, research teams & labs or even scientific terms and see where they fit on a navigable map. You can look for the nearest entities of each kind to understand the underlying connections that shape a specific domain.



Objectives of the internship :

The intern will have to set a test protocol to evaluate the results of the scientific cartography provided by Cartolabe web application (cartolabe.fr). In a first step, quality indicators and a validation protocol will be defined. Applying Cartolabe to different datasets like Wikipedia for instance and the comparison with the results returned by other search engine for instance should allow to validate the chosen indicators and the test protocol. In a second step, optimizing the cartography hyperparameters will improve the reached quality.

Job description :

Cartolabe is a LRI – CNRS - Inria common project aiming at visualizing many publications, authors, labs and teams on a unique map (up to 10^6 points).
Cartolabe application builds a distance between these entities linked to publication by mean of articles text content. A data handling pipeline scraps the data from HAL open archive (https://hal.archives-ouvertes.fr/: 750 000 articles and authors as of today) and works them out using machine learning techniques. A single json (or feather) format file is produced as pipeline output. Then, a second part of the application (a web application) is in charge of visualizing the point cloud in a zoomable annotated heatmap. Full exploration possibilities are offered on the web client.

1 Version du 13 Novembre 2019

As an example, a natural intrinsic quality indicator could be to count, per author, the part of his articles which are reasonnably 'near' his own author point localisation.
Extrinsic quality indicators could be conceived by sending similar requests to independant applications like Google Scholar or LookInLabs (https://lookinlabs4halinria.cominlabs.u-bretagneloire.fr/) and comparing the outputs.
Manual quality indicators are also possible by asking scientific referees and experts in a recorded formal querying session to check the validity of the distances proposed by Cartolabe.
Some of the quality indicators can be confrontated with other available information articles citations.

Once the quality indicators will be defined, the second part of the internship will focus on optimizing the pipeline hyperparameters in order to establish their correlation with the indicators and to improve the output results on the map. Hyperparamters are either the choice of an algorithm among several possible others : LDA or LSA, chosen neighborood or projection method the choice of their parameters, like latent dimensions number for similarity computation.

Expected abilities of the candidate:

- Python programmation and tools : Anaconda, scikit-learn, pandas...;
- software environment tools : software forges, git ;
- Appreciated knowledge in one of the following: large corpus data visualization, machine learning, Natural Langage Processing, information retrieval : recall vs precision.
- Scientific english level required ;
- methodology, curiosity and team work ability are also required for this internship.

Organisational background:

TAU (TAckling the Underspecified) is an Inria team belonging to the LRI lab, the Reseacrh Laboratory in Computer Science. The LRI is attached to both the computing science department of Paris Saclay University and to the INS2I Institute of the CNRS. It is also tightly linked in a partnership with Inria and CentraleSupelec neighbor institutions. LRI hosts more than 250 people, 115 permanent people and 90 PhD on Plateau de Saclay.
Cartolabe project is a high potential project held by a team of scientifics and engineers from LRI and Inria. Both parts, data pipeline and visualization module, have an open architecture to be adapted to various application fields. The hyperparameters tuning part of the internship is essential for the project, because it will validate at least one Cartolabe instance, the one for the scientific publications model.

Recruiting level : M2 internship or engineering school 3rd or 4th year.

Internship fees and contract: Paris Saclay or INRIA Internship convention. About 550€ monthly fees, according to number of working days in the month.

Location : INRIA TAU team, at LRI, Laboratoire de Recherche en Informatique – Paris Saclay university – Building 660 – Shannon (Gif-sur-Yvette)

Duration : 4 to 6 months from march 2020.

Send CV and application letter to

Philippe Caillou : caillou@lri.fr and Anne-Catherine Letournel : acl@lri.fr