# Thesis subject Proposal

## Representation learning for multimodal data

**Scientific director:**

**Nistor Grozavu**, Full Professor
ETIS UMR 8051, CY Cergy Paris University, ENSEA, CNRS
E-mail : nistor.grozavu@cyu.fr
Web : http://www.grozavu.fr
Commitment on this project (%): 50%

**Xiaoli LI,** Full Professor
Institute for Infocomm Research, A*STAR, Singapore.
Email: xlli@i2r.a-star.edu.sg
Web: https://personal.ntu.edu.sg/xlli/
Commitment on this project (%): 50%

Shared ARAP scheme funding CYU (France) – A*STAR (Singapore)

- **Salary**: 18 months CYU contract (legal doctoral salary, around 1500-1600€/month) + 18 months A*STAR contract (S$ 2700/month)
- Running cost: 1000 €/year
- Round trip to the other country: 600 - 1000 €
- Per Diem in the other country: 100€
- Intl conferences (at least 3): 2000€ including flight + registration + housing * 3

### Description

The performance of machine learning algorithms is largely determined by data representation, which we believe is due to the fact that different representations may entangle and hide distinct explanatory aspects of variation behind the data to varying degrees [1; 5].

In recent years, the in the Artificial Intelligence (AI) community, a lot of research and applicative works have appeared dealing with the learning of neural networks architectures, reflecting a resurgence of research activities around deep neural networks that have been eclipsed by kernel machines. This revival is largely explained by their ability to obtain excellent performance for classification and dimension reduction tasks, and by their ability to extract in their successive hidden layers increasingly high-level information about the system entry. The text, image, videos applications are impressive with the extraction of visually very relevant and appropriate primitives for the types of data used during the learning process.

Deep Neural Networks, are none other than multilayer networks, of classic architecture, but they include several hidden layers and it is the way of managing their learning that has given, since 2006, a resurgence of interest in their study.

Indeed, even if the approximation theorems of the end of the 80s assert that, in theory, a single hidden layer suffices for the approximation of any sufficiently regular function, nothing prevents a priori from implementing learning by back-propagation in networks comprising sev-

eral hidden layers. Therefore, some results highlight the interest of considering two or more hidden layers to obtain more parsimonious and efficient networks, by composing several levels of nonlinearities.

On the other hand, the advanced acquisition systems allow recording several types of data for a deeper understanding of complex scenarios in which the data are very often continuously gathered from different sources and available in different modalities, involving multimodal dynamical data. This results in heterogeneous multimodal data, including time series, text, images, etc. present methodological issues in the modelling and learning from such complex data. This challenge renders the supervision process almost impracticable and very uncertain given the unknown dynamical nature of the observed systems, for which the classical 'static' machine learning and statistical inference approaches are inappropriate and unfeasible [3; 4].

Indeed, despite this important increase in the accessibility to multimodal complex dynamic data, there is a lack of approaches to deal with, particularly for the representation learning. Most classical machine learning and statistical inference systems dedicated to multimodal and/or complex data, whether they are based on random models, empirical measures or prototype-based models, rely on a strong hypothesis, consisting in supposing at least that the structure of the data generating process for the observed scene is fixed, though it can be supposed unknown. In an unsupervised context, some existed works on Ensemble and Collaborative machine learning approaches were proposed but are limited to the same data distribution, i.e. in a multi-view context [2;3].

We want through this thesis to explore the unsupervised topological learning of multimodal data presenting a complex structure allowing to learn their representations. We are particularly interested in heterogeneous data whose representation may have been informed in different ways: expert representation which may be complex (for example: dynamic multi-graphs which may possibly have different topologies for each observation of the dataset and for every moment) or automatically learned representation (for example: embedding in a high-dimensional space with dense vectors and potential correlations between the components), images, signals.

**Références :**


1. Bengio Y., Courville A., Vincent P. (2013), Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence , Volume 35 Issue 8 Publisher: IEEE Computer Society
2. Rastin, P., Matei, B., Cabanes, G., Grozavu, N., Bennani, Y.: Impact of Learners' Quality and Diversity in Collaborative Clustering. Journal of Artificial Intelligence and Soft Computing Research, 9(2): 149–165 doi:10.2478/jaiscr-2018-0030 (2018).
3. W. Guo, J. Wang and S. Wang, "Deep Multimodal Representation Learning: A Survey," in *IEEE Access*, vol. 7, pp. 63373-63394, 2019, doi: 10.1109/ACCESS.2019.2916887.
4. Rastin, P., Matei, B., Cabanes, G., Bennani, Y., Marty, J.M.: A new sparse representation of complex data: application to dynamic clustering of web navigation. Pattern Recognition, 91: 291–307, doi:10.1016/j.patcog.2019.02.020 (2019).
5. Bengio Y. (2013), Deep learning of representations: looking forward. SLSP'13: Proceedings of the First international conference on Statistical Language and Speech Processing.