**Title**: Integration of heterogeneous data sources for autoimmune diseases
**Start date**: From September 2018
**Contract duration**: 24 months
**Gross salary**: commensurate to experience
**Primary Contact**: Arthur Tenenhaus (arthur.tenenhaus@centralesupelec.fr) and Ivan Moszer (ivan.moszer@icm-institute.org)

**Description of the position.**

iMAP is a “Recherche Hospitalo-Universitaire” grant aimed at developing low dose IL2 as a therapy of autoimmune diseases, and likewise study the biology of IL2 in humans. This project is coordinated by the i3 laboratory (www.i3-immuno.fr) located on the Pitié-Salpêtrière hospital campus in Paris (13ème).

The iMAP database includes, for several groups of patients (different pathologies and healthy controls), a variety of information, acquired at multiple time points: (i) clinical data, (ii) standard biology, (iii) multiomics data (flow cytometry, immunoproteomics, TCR repertoire, transcriptomics and microbiome). We use the term "modality" for each type of acquisition. Data integration is the process of integrating multiple modalities to produce more consistent, accurate, and actionable information than that provided by any individual modality. Motivations for multimodal data integration are numerous: obtaining a global picture of the system at hand; identifying common or distinctive elements across modalities or time; improving decision making; etc.

We have proposed a general statistical framework for data integration called Regularized Generalized Canonical Correlation Analysis (RGCCA) [Garali, 2017, Tenenhaus et al. 2017, Tenenhaus et al 2015, Lofstedt et al 2015, Tenenhaus & Tenenhaus 2014a, Tenenhaus et al 2014b, Tenenhaus & Tenenhaus 2011]. Within this framework, the selected candidate will contribute to improving both theoretical and applicative components for multimodal data integration, in particular by developing novel strategies for handling missing data imputation and longitudinal study schemes. This methodology will be applied on the iMAP dataset to address various questions raised by scientists and clinicians of the iMAP consortium, especially for the identification of biomarkers to sustain the rational development of innovative biotherapies.

**Requirements (training/expertise) and profile.** PhD in applied mathematics/statistics/machine learning. Previous experience in multivariate data analysis applied to biological data will be strongly appreciated. Strong programming skills in at least one programming language (R, Matlab or Python) are mandatory. The selected candidate will work within the Bioinformatics/Biostatistics core facility (iCONICS) of the Brain and Spine Institute (ICM, Paris 13ème), and will interact with researchers of the Laboratoire des Signaux et Systèmes (L2S, CentraleSupelec, Gif-sur-Yvette) and investigators involved in the iMAP project.

To apply, submit a cover letter indicating past research experience, motivation for the position, expected availability date, a curriculum vitae, and at least 2 references, to arthur.tenenhaus@centralesupelec.fr and ivan.moszer@icm-institute.org.

**Bibliography**

Tenenhaus, M., Tenenhaus, A., Groenen, P. J. (2017). Regularized generalized canonical correlation analysis : a framework for sequential multiblock component methods. Psychometrika, 82(3), 737-777

Garali, I., Adanyeguh, I., Ichou, F., Perlbarg, V., Seyer, A., Colsch, B., Moszer, I., Guillemot, V., Durr, A., Mochel, F., Tenenhaus A. (2017). A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia. Briefings in Bioinformatics, bbx060

Tenenhaus, A., Philippe, C., Frouin, V. (2015). Kernel Generalized Canonical Correlation Analysis.Computational Statistics & Data Analysis, 90, 114-131.

Löfstedt T., Hadj-Selem F., Guillemot V., Philippe C., Duchesnay E., Frouin V., Tenenhaus A. (2015). Structured variable selection for generalized canonical correlation analysis. In The Multiple Facets of Partial Least Squares Methods, Proceedings in Mathematics and Statistics, Springer New York

Tenenhaus A., Philippe C., Guillemot V., Lê Cao K.-A., Grill, J., Frouin V., (2014). Variable Selection for Generalized Canonical Correlation Analysis, Biostatistics, 15 (3) : 569-583

Tenenhaus A., Tenenhaus M., (2014). Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. European Journal of Operational Research, 238(2), 391-403.

Tenenhaus A., Tenenhaus M., (2011). Regularized Generalized Canonical Correlation Analysis, Psychometrika, vol. 76, Issue 2, pp. 257-284