# Causal Statistical Matching with Optimal Transport

Master/Ecole d'Ingénieur internship

Expected starting: February/March 2021 (6 months)

## Keywords

Causal Machine Learning, Optimal Transport

**Context** In artificial intelligence, in many fields of application, statistical learning methods have demonstrated their high level of performance. One of the tasks often performed by this type of method consists in studying the dependence / statistical associations between variables in order to understand the relationship between these explanatory variables and a variable of interest, and to predict this variable from the observation of new individuals. The aim of this research project is to evaluate the performance of machine learning methods through the angle of causality. Indeed, if the inference approach has a formal and theoretical framework tested for many contexts, statistically significant associations are not necessarily linked to causal structures between the variables. These questions arise in various fields of application, notably when analyzing clinical research data (to evaluate the real impact of a treatment) or remote sensing data (to highlight explanatory causes of an observed phenomenon) for example.

**Objectives** Motivated by the methodological problems which arise for the analysis of data in this causal context [1], the objectives of this internship is to link recent methodological developments in machine learning (and more specifically in the context of optimal transport) to this causal inference problem. Optimal Transport is an old mathematical question that revealed to be extremely sound in machine learning frameworks. The recent theoretical and applied developments of optimal transport, and in particular regarding computational aspects, have opened the way to multiple applications in machine learning [2]. Our team has a strong background on this topic, and we propose in this internship to rely on a recent method published at NeurIPS 2020: Co-Optimal Transport [3]. Specifically, we will use this method in a context of statistical matching (find association in different databases) or imputation (infer missing variables in records) with optimal transport, in the spirit of [4, 5]. In a second time, we will extend this method by integrating causality notions.

**Outcomes** Potential outcomes of the internship will lead to publication in the machine or statistical fields, depending on the nature of the contributions. Let us finally note that this internship will be part of the AI chair OTTOPIA (Earth Observation with Optimal Transport for Artificial Intelligence) funded by ANR (starting beginning of 2021), for which potential funding is available for the candidate to **enter a PhD program** after the internship.

## Work program

In order to address the aforementioned objectives, a tentative work program is given below.

- Bibliographical study on causal machine learning, statistical matching, optimal transport and data imputation
- Evaluation of an existing approach for coupled sample/features matching
- Development in Python of operational machine learning tools
- Evaluation and benchmarking on real remote sensing and/or medical datasets

## Required background and skills

- Student in Master 2 or Ecole d'Ingénieur or equivalent with excellent academic track;
- Background in machine/statistical learning and/or applied mathematics;
- Excellent programming in Python (familiar with one of deep learning packages, such as PyTorch, is a plus.)

## Supervision

The expected intern will join the OBELIX research group (`www.irisa.fr/obelix`) from IRISA (UMR 6074) is located in the UBS (Université Bretagne Sud) campus in Vannes 56000, France.

The internship will be jointly supervised by **Dr. Chloé Friguet**[1] (Maîtresse de Conférences at UBS), **Dr. Valérie Garès**[2] (Maîtresse de Conférences at INSA Rennes) and **Prof. Nicolas Courty**[3] (Professor at UBS).

## Application

Send your CV + Motivation letter + Master transcripts to `chloe.friguet@irisa.fr`, `valerie.gares@insa-rennes.fr` and `nicolas.courty@irisa.fr` **(before 20 November 2020)**. Potential candidates will be contacted for interview. Feel free to contact us for any question.

## References


[1] B. Scholkopf (2019), Causality for Machine Learning, arXiv:1911.10500

[2] G. Peyré and M. Cuturi (2019). Computational optimal transport. Foundations and Trends in Machine Learning, 11(5-6), 355-607.

[3] I. Redko, T. Vayer, R. Flamary and N. Courty (2020). CO-Optimal Transport. NeurIPS'20, arXiv preprint arXiv:2002.03731.

[4] B. Muzellec, J. Josse, C. Boyer and M. Cuturi (2020) Missing Data Imputation using Optimal Transport, ICML'20, arXiv preprint arXiv:2002.03860.

[5] V. Garès and J. Omer (2020) Regularized Optimal Transport of Covariates and Outcomes in Data Recoding, Journal of the American Statistical Association. 1-14.


[1] `http://people.irisa.fr/Chloe.Friguet/`
[2] `http://vgares.perso.math.cnrs.fr/`
[3] `http://people.irisa.fr/Nicolas.Courty/`