

Bastien Dussap, phd

Data Scientist

* 17 June 1998

☎ 07 77 99 21 57

✉ bastiendussapapb@gmail.com

🌐 bastiendussap.github.io

🐦 [BastienDussap](#)

🔄 [BastienDussap](#)

🔗 [bastiendussapapb](#)

Education

- 2021–2024 **PhD thesis in Machine Learning**, *Université Paris Saclay*
PhD thesis in Machine Learning apply to the comparison of Cytometric data.
- 2019–2021 **Master Mathématiques et Applications**, *Université Paris Saclay*
Master's degree in Mathematics apply to Artificial Intelligence.
- 2016–2019 **Licence de Mathématiques**, *Université Paris Saclay*
Bachelor's degree of Mathematics. The first two years at Evry and the last one at Orsay.

PhD thesis

Title *A unified framework for label shift quantification*

Supervisors Gilles Blanchard and Marc Glisse

Abstract In supervised classification, it is not uncommon that the information sought is not local, meaning the label associated to each data point, but global: obtaining the proportions of the different labels within the sample directly. This problem, which we have chosen to refer to as label shift quantification but which is also known by many other names in the literature, has seen a proliferation of publications since the mid-2000s. However, these works often proceed in parallel, coming from communities with limited dialogue, resulting in a scattered bibliography. In this manuscript, we first provide an overview of these diverse works with a twofold aim: first, to bridge the gap between these communities by presenting results from different research areas, and on the other hand, contextualise the subsequent work, particularly focusing on efforts to unify methods. Second, we propose a framework that unifies several classical methods from the literature based on mean vectorisations. We examine the theoretical guarantees of these methods and demonstrate their robustness when the central assumption of label shift is violated. We also extend this work by focusing on kernel-based vectorisations using covariance information rather than just the mean. Finally, we explore the use of a specific vectorisation based on Random Fourier Features in applications related to flow cytometry.

Experience

2024– **Data Scientist**, *Metafora Biosystems*
Data Scientist at Metafora Biosystems.

Teaching

2022–2023 **Mathematics for Management**, *IUT Sceaux*, L1 B.U.T GEA
Taught by Patrick Pamphile.

Seminar

2022–2024 **Seminar**, *Université Paris-Saclay*, Master 2
Co-organizer of the seminar for master students in Statistics and Machine Learning at Université Paris-Saclay.

Publication

- 2023 **Label Shift Quantification with Robust Guarantees via Distribution Feature Matching**, *ECML/PKDD 2023*, With Gilles Blanchard and Badr-Eddine Chérif-Abdellatif

Quantification learning deals with the task of estimating the target label distribution under label shift. There exist two main classes of quantifiers in the literature: classification-based methods vs statistical mixture modeling approaches. In this paper, we propose an efficient and scalable quantifier that belongs to the second class, and we present a unifying framework based on feature distribution matching that recovers estimators from both quantification families. In particular, we derive a general consistency theorem under label shift which improves upon the bounds that can be found in the literature, investigate the misspecified setting where the exact label shift hypothesis is challenged, and provide a detailed numerical study on simulated and real-world datasets.






Award








- 2023 **Best Student Paper - Research Track**, *ECML/PKDD 2023*, Label Shift Quantification with Robust Guarantees via Distribution Feature Matching

Languages

French	Native	
English	B2	CEFR rating

Computer skills

	basic knowledge		extensive project experience
	intermediate knowledge with some project experience		deepened expert knowledge
			expert / specialist

	Level	Skill	Years	Comment
Language:		Python	4	Used Python and standard Machine Learning packages such as numpy, matplotlib, scikit-learn or pytorch. Creation of custom packages.
		SQL		Online course on SQLite
		R	1	Used R for Machine Learning project during my education.
		Zotero	3	Used Zotero during the PhD to manage my bibliography.
		LaTeX	5	
OS:		Linux	4	I only use Ubuntu for work.
		Windows	10+	Use Windows at home.

References

- [1] Bastien Dussap, Gilles Blanchard, and Badr-Eddine Chérif-Abdellatif. Label shift quantification with robustness guarantees via distribution feature matching.

