Laurent Gatto
Department of Biochemistry, University of Cambridge
Tennis Court Road, Cambridge, CB2 1GA, UK
Email: lg390@cam.ac.uk
Tel: +44 (0) 1223 760253

After a PhD in Evolutionary Genetics at the Free University of Brussels, and complementary studies in Computer Science at the University of Namur, Belgium, I worked for 3 years in industry with a focus on computational pipelines R&D. In 2010, I moved to the University of Cambridge, where I concentrated on data and software infrastructure for proteomics, first as a Post-doctoral Research Associate and later as a Senior Research Associate. In 2014, I was awarded an Sustainability Institute fellowship and recently have become a Software Carpentry instructor.

Since the completion of my PhD, the emphasis of my work has focused on delivering software solutions to tackle specific data-driven scientific questions, that have culminated in several state-of-the-art software, data and educational packages. The extend to which I can deliver research software is contstraint by the limited opportunities for research software engineering in academia. The fellowship will give me a unique opportunity to pursue my inter-disciplinary career focusing on research software engineering.

Project summary Modern science relies on ever increasing quantities of data (some have coined it a 'data deluge') and on trustworthy and sustainable software to analyse and interpret them. Each of these pillars of modern scientific activities are generally developed and maintained independently. On one hand, a typical data pipeline starts with the submission of annotated data to an official repository that, after some basic curation, serves them to the wider community. This flow is unidirectional and, once released, becomes inert (updates and further annotations are very rare) and, most often, dormant. On the other hand, the exposure of research software is often limited by the application (or applicability) to a reduced number of data sets. Surprisingly, while each data/software information and development streams can't be conceived in isolation, and their true potential emerges from their interoperation, little efforts exist to facilitate streamlined interactions and discoverability and enable new data/software meta-analyses.

A recent project involved developing, implementing and applying a transfer learning algorithm that learns from different data sources to classify features. In our specific question, the classification is aimed at inferring the spatial sub-cellular localisation of proteins using experimental data and third-party gene ontology data. Our algorithm was initially inspired from a leaves image classification problem and was improved and adapted for our use cases. We would like to apply it to another domain, genomics data being an obvious candidate; however, manual exploration and reliance on personal contacts are not effective and substantially slow cross-fertilisation between domains. The transition from the first publication of the algorithm in the International Conference on Machine Learning to a functional implementation and application in spatial proteomics initially relied on a fortuitous discovery. What further transitions and improvements could be anticipated? If these are prohibitively slow or difficult, the dissemination, improvement and generalisation of the software is dramatically limited. The project's goal is to offer systematic ways to enable such transitions.

Concretely, the first major milestone of the project is a set of interconnected databases of experimental data, software and publications, all spanning different disciplines, created from mining relevant repositories. The relations between records within and between these databases will be browseable and searchable through an online portal and programmatically accessible through specific APIs. Given specific properties of existing records, or new, user-defined properties describing software, data or publications, new matches can then be identified. Properties related to quality and sustainability of software and data (metadata, documentation, testing, ...) will be used to favour the promotion of more sustainable code. Known cross-disciplinary data/software outputs will be used to optimise and assess the relevance of the infrastructure. The second major milestone aims at facilitating the application of the relations identified above by enabling transparent integration of different data source (local or remote, public or private), data sets (transparent to file formats) through data APIs and software architectures (local or remote) through software APIs/pipelines. Milestone one enables systematic exploration of RSE entities and milestone two supports their automated and standardised meta-analysis. Each of these deliverables will be complemented by dedicated open communication channels to openly publicise and disseminate the new software/data relations and use cases.