Data Sci. Eng. (2017) 2:197–198
https://doi.org/10.1007/s41019-017-0049-x

Data-Intensive Modelling and Simulation in Life Sciences and Socio-economical and Physical Sciences

Contributions from the Coordinated Work in *cHiPSet*, the COST Action on *High-Performance Modelling and Simulation for Big Data Applications*

Andrea Bracciali[1] · **Elisabeth Larsson**[2]

This special issue of the journal Data Science and Engineering is a result of the work fostered by the ICT COST Action IC1406 *High-Performance Modelling and Simulation for Big Data Applications (cHiPSet)*, an EU-funded research network across 30+ European countries and overseas partners (http://chipset-cost.eu/).

The Big Data era poses critically difficult challenges and striking development opportunities in high-performance computing (HPC), which appears to be, in several domains, an enabling technology for the ongoing efforts to efficiently turn massively large data into valuable information and meaningful knowledge. On the other hand, modelling and simulation (MS) offer suitable abstractions to manage the complexity of analysing Big Data in the various domains.

Unfortunately, modelling and simulation of big data problems does not always naturally lend itself to efficient HPC solutions. MS communities often lack the detailed expertise required to exploit the full potential of HPC solutions, and HPC architects may not be fully aware of specific MS requirements.

cHiPSet is an opportunity to coordinate European research, with the support of overseas colleagues, and facilitate interactions among data-intensive MS and HPC experts, both from research and industry. This Action aims to support the development of the field, which is strategic and of long-standing interest. *cHiPSet* is organised around four working groups: WG1 and WG2 on HPC infrastructures and programming models for MS, while WG3 and WG4 are thematic umbrellas for data-intensive MS in Life Sciences and for Socio-economical and Physical Sciences.

This *DSE* special issue collects contributions originating from the coordinated work of the "modelling" working groups WG3 and WG4 across the first two years of the Action. Andrea Bracciali and Elisabeth Larsson, guest editors of this volume and chairs of WG3 and WG4, respectively, have compiled papers presenting general examples of modelling and simulation for big data problems in the context of the two working groups. Within the Action, these and other approaches are of strong interest as paradigmatic case studies for the coordinated development of HPC solutions and for fostering collaboration across the various scientific communities in the cHiPSet network. Analogously, readers interested in state-of-the-art examples of efficient big data modelling will find the contributions of this volume of interest.

This special issue is based upon work from COST Action IC1406 cHiPSet, supported by COST (European Cooperation in Science and Technology).

✉ Andrea Bracciali
andrea.bracciali@stir.ac.uk

Elisabeth Larsson
elisabeth.larsson@it.uu.se

[1] Computing Science and Mathematics, Stirling University, Stirling FK9 4LA, UK

[2] Department of Information Technology, Uppsala University, Box 337, 751 05 Uppsala, Sweden

The special issue contains five papers, and below we give a brief overview of what type of data is central for each paper and what kind of outcomes are expected from the methods presented.

Paper 1: *Trust-based Modelling of Multi-criteria Crowd-sourced Data*, by Fátima Leal, Benedita Malheiro, Horacio González-Vélez, and Juan Carlos Burguillo, considers the problem of providing high quality personalised recommendations for travellers, based on crowd-sourced data available from sites such as Expedia and TripAdvisor. When an individual rates, e.g. a hotel, there are different aspects that can be rated such as cleanliness or comfort (multi-criteria). Furthermore, the ratings from different individuals may be more or less relevant for a specific recommendation and/or more or less trustworthy considering the overall rating behaviour of that individual. These characteristics are taken into account in the novel approaches presented here.

Paper 2: *Tracking time evolving data streams for short-term traffic forecasting*, by Amr Abdullatif, Francesco Masulli, and Stefano Rovetta, addresses the problem of processing large volumes of traffic flow data collected in a real-time setting. A challenging feature of the data is that it is non-stationary. The traffic behaviour can change both abruptly and with a gradual drift over time. In order to perform short-term traffic forecasts, the forecaster needs to continuously learn from the data stream, while adapting to the current situation as well as detecting anomalous data points.

Paper 3: *Using GUHA data mining method in analysing road traffic accidents occurred in the years 2004–2008 in Finland*, by Esko Turunen, presents another application in the traffic domain. Investigated in this paper is how to use traffic accident data collected over time to learn about which (risk) factors are strongly correlated with, for example, single vehicle accidents or accidents leading to severe injuries. Knowledge about these relations can then be used to inform preventive work at the societal level.

Paper 4: *Robust cross-platform workflows: How technical and scientific communities collaborate to develop, test and share best practices for data analysis*, by Steffen Möller, Stuart W. Prescott, Lars Wirzenius, Petter Reinholdtsen, Brad Chapman, Pjotr Prins, Stian Soiland-Reyes, Fabian Klötzl, Andrea Bagnacani, Matúš, Kalaš, Andreas Tille, and Michael R. Crusoe lies in the context of bioinformatics, a rapidly evolving area where the data volumes are becoming very large (whole genomes of multiple individuals), and where evolving data collection methods, improved data quality, and novel data analysis methods imply that software must adapt and provide new functionality at quite short-time scales. Bioinformaticians need to spend significant amounts of time to adjust workflows to the current state of data and software. This paper discusses how to integrate and coordinate open source packages, e.g. through the emerging Common Workflow Language, to realise such workflows, efficiently.

Paper 5: *A review of scalable bioinformatics pipelines* by Bjørn Fjukstad and Lars Ailo Bongo, is a review paper focusing on scalability, an important open question for bioinformatics workflows. Many types of analyses are provided as web services. This means that portals need to be scalable with respect to the number of users accessing them simultaneously, and the service needs to be scalable with respect to increasing data volumes. The backend system providing the service can be a cloud system or a high-performance computing (HPC) cluster, leading to different issues. An advantage of the cloud is that resource can be provided elastically, depending on the load situation. The HPC system cannot do that in general, and instead it becomes important that the service is scalable over the hardware with respect to nodes and cores.

As guest editors of this special issue we would like to thank the journal editors and the editorial staff for their support during the publication process. We would also like to thank the anonymous reviewers that have supported the review process for this special issue. Lastly, we want to acknowledge the support of COST (www.cost.eu) that, through the *cHiPSet* Action, has provided a collaborative environment fostering interaction amongst European, and overseas, researchers, across a particularly large number of countries.