Research Fellow Position
Data Quality, Uncertainty, and Lineage

Stéphane Bressan & Pierre Senellart
National University of Singapore & ENS, PSL University

We describe a two-year Research Follow (post-doctoral) position available at CNRS@CREATE, Singapore, in the framework of a collaboration between National University of Singapore and the Valda team of École normale supérieure (PSL University), CNRS & Inria (France).

Background and Research Description

Real-world data is not always readily usable: it may be marred with *uncertainty*, of *low quality*, or *too costly* to be obtained (because of its price, access limitations, or for privacy reasons). There is actually often a *trade-off* between the cost and quality of data: acquiring more data or higher-quality data is often feasible, but at the cost of paying for it, spending communication resources downloading it, or spending computation resources running costly data enrichment tools.

An approach to keep track of the lineage and uncertainty of information is through the use of provenance-aware, probabilistic, database management systems [1]. Such a state-of-the-art system, ProvSQL [2] (`https://github.com/PierreSenellart/provsql/`) is in development in the Valda group, relying on the semiring provenance model of [3].

The goal of this Research Fellow position is to develop new models, algorithms, implementations for the management of the provenance and probability of uncertain, low-quality, costly-to-access data, by extending the ProvSQL system and rigorously evaluating them on real-world data (including data from the DesCartes project, see further). In particular, the following aspects can be investigated:

- How to add robust support for continuous probability distributions into probabilistic database query evaluation?
- How to handle arbitrary aggregate queries over provenance-aware and probabilistic databases?
- How to properly deal with database operations beyond regular query evaluation: updates, user-defined functions, out-of-database operations?

Environment

The Research Fellow will be employed for this two-year position by CNRS@CREATE, the subsidiary of CNRS in Singapore. He will be based at the CREATE Tower, within National University of Singapore. He will work under the co-supervision of Stéphane Bressan, Associate Professor at NUS and Pierre Senellart, Professor at ENS. This position is funded in the context of the DesCartes project[1] of CNRS@CREATE, a French–Singaporean collaborative project in Singapore, on the general topic of *Intelligent Modelling for Decision-Making in Critical Urban Systems*. The PhD is part of WorkPackage 2 of DesCartes on *Learning from Smart and Complex Data for Hybrid AI*; it may involve use cases relevant to the project such as *structure monitoring*, *control of drones*, *digital energy*, though it remains a basic research project in computer science.

Pierre Senellart will spend part of his time in Singapore over the period of the PhD. Regular meetings will also be held by video-conference. Finally, there will also be opportunities for the Research Fellow to spend time in Paris, for extended research trips working in the Valda group.

Conditions

Starting date Flexible. The starting date should be in 2022 or in the first half of 2023. The earlier, the better.

Deadline for application Applications will be continually considered as they are received.

Prerequisites Doctoral degree in computer science (or similar field) and research experience in the area of data management, machine learning, or data mining. An interest for and willingness to develop both theoretical and systems aspects of data management research is desired.

How to apply

Contact Stéphane Bressan (steph@nus.edu.sg) and Pierre Senellart (pierre@senellart.com) with all relevant background information, including a CV.

References

[1] Pierre Senellart. Provenance and Probabilities in Relational Databases: From Theory to Practice. *SIGMOD Record*, 46(4), December 2017.

[2] Pierre Senellart, Louis Jachiet, Silviu Maniu, and Yann Ramusat. ProvSQL: Provenance and Probability Management in PostgreSQL. In *Proc. VLDB*, pages 2034–2037, Rio de Janeiro, Brazil, August 2018. Demonstration.

[3] Todd J. Green, Gregory Karvounarakis, and Val Tannen. Provenance semirings. In Leonid Libkin, editor, *Proceedings of the Twenty-Sixth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 11-13, 2007, Beijing, China*, pages 31–40. ACM, 2007.

[1] https://www.cnrsatcreate.cnrs.fr/descartes/