PhD Position
Data Quality, Uncertainty, and Lineage

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

We describe a PhD position available at the National University of Singapore, in collaboration with the Valda team of École normale supérieure (PSL University), CNRS & Inria.

Background and Topic 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.

We intend to address these issues in this PhD research work by considering some or all of the following problems:

- How to assess the quality of a given dataset or collection of datasets? This is the problem of deriving a formal (e.g., probabilistic) model of the uncertainty contained in a dataset.
- How to efficiently manage (model, store, query, keep track of) large quantities of uncertain data, possibly with a cost model on data access?
- How to best address the trade-off between cost of data and quality of resulting applications?
- In the presence of too costly data, how to obtain synthetic datasets that can be used to realistically complement a dataset, with as high quality as possible?

These questions are fairly abstract, so consider a specific use case, that of extracting structured theorems and proofs from the scientific literature, a current topic of interest in the ENS research group [1, 2, 3]:

- acquiring PDF articles from the scientific literature is costly (some are freely available but to be crawled from various sources, some are fully inaccessible because of intellectual property reasons, some can be obtained for a cost);
- current information extraction techniques are imperfect, so produce results of varying quality;
- there are often alternatives for such information extraction tools, each coming with a specific cost and quality trade-off;

- human experts can be relied on, at a (monetary or availability) cost to hand-label articles;
- it is critical to keep track of the lineage of every piece of data item (which article it comes from, which techniques have been used to produce it, etc.) as well as the remaining uncertainty about this data.

This is just one example, other relevant examples can be derived from the use cases of the Descartes project (see further).

In order to address the varying issues relevant to data quality, cost, lineage, it will be possible to rely on the following approaches and tools:

Truth finding. Truth finding, or truth discovery, is the problem of estimating the truth in the presence of contradicting sources of information. This can be solved using a variety of approaches, in particular based on optimization and machine learning. It has been a subject of interest in both the NUS and Valda groups. [4, 5, 6]

Provenance management and probabilistic databases. An approach to keep track of the lineage and uncertainty of information is through provenance-aware, probabilistic, database management systems [7]. Such a state-of-the-art system, ProvSQL [8] is in development in the Valda group, relying on the semiring provenance model of [9].

Privacy optimization. One particular cost to take into account while dealing with a dataset is the *privacy budget* involved in this particular dataset. A cost-based approach to privacy optimization has been proposed in the NUS group. [10]

Reinforcement learning. Reinforcement learning is learning towards an objective in a dynamic situation where each action may lead to a reward (or may have a cost). It is a potentially useful framework to model the optimization process of the cost–quality trade-off. Joint work on this topic has been conducted between the NUS and Valda groups. [11, 12, 13, 14]

Environment

The 4-year PhD thesis will be carried out in the School of Computing of the National University of Singapore (NUS), under the co-supervision of Stéphane Bressan, Associate Professor at NUS and Pierre Senellart, Professor at ENS. The PhD thesis is carried out and 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 PhD candidate to spend time in Paris during his or her PhD, for extended research trips working in the Valda group.

The PhD student will need to satisfy the requirements of a PhD in Computer Science at NUS, see `https://www.comp.nus.edu.sg/programmes/pg/phdcs/` for details. The deadline for August 2022 intake is **January 5th, 2022**, see `https://nusgs.nus.edu.sg/programme/`

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

phd-department-of-computer-science/ for applications. But please contact Stéphane Bressan and Pierre Senellart by email before making a formal application for this PhD programme, preferably **by December 15, 2021.**

Conditions

Starting date August 2022 (or January 2023).

Deadline for application December 15, 2021 to contact PhD advisors; January 5, 2022 for August 2022 intake; June 15, 2022 for January 2023 intake.

Prerequisites Bachelor's degree in computer science (or equivalent diploma) and research experience in the area of data management, machine learning, or data mining. Holders of a Master's degree (or planning to obtain a Master's degree in 2022) are also welcome.

References

[1] Shrey Mishra, Lucas Pluvinage, and Pierre Senellart. Towards Extraction of Theorems and Proofs in Scholarly Articles. In *Proc. DocEng*, Limerick, Ireland, August 2021.

[2] Lucas Pluvinage. Extracting scientific results from research articles. Master's thesis, ENS, PSL University, 2020.

[3] Lucas Pluvinage. A knowledge base of mathematical results. Master's thesis, ENS, PSL University, 2020.

[4] Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. Corroborating Information from Disagreeing Views. In *Proc. WSDM*, pages 131–140, New York, USA, February 2010.

[5] Mouhamadou Lamine Ba, Roxana Horincar, Pierre Senellart, and Huayu Wu. Truth finding with attribute partitioning. In Julia Stoyanovich and Fabian M. Suchanek, editors, *Proceedings of the 18th International Workshop on Web and Databases, Melbourne, VIC, Australia, May 31, 2015*, pages 27–33. ACM, 2015.

[6] Ngurah Agus Sanjaya, Mouhamadou Lamine Ba, Talel Abdessalem, and Stéphane Bressan. Harnessing truth discovery algorithms on the topic labelling problem. In Maria Indrawan-Santiago, Eric Pardede, Ivan Luiz Salvadori, Matthias Steinbauer, Ismail Khalil, and Gabriele Anderst-Kotsis, editors, *Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services, iiWAS 2018, Yogyakarta, Indonesia, November 19-21, 2018*, pages 8–14. ACM, 2018.

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

[8] 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.

[9] 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.

[10] Ashish Dandekar, Debabrota Basu, and Stéphane Bressan. Differential privacy at risk: Bridging randomness and privacy budget. *Proc. Priv. Enhancing Technol.*, 2021(1):64–84, 2021.

[11] Miyoung Han, Pierre Senellart, Stéphane Bressan, and Huayu Wu. Routing an Autonomous Taxi with Reinforcement Learning. In *Proc. CIKM*, Indianapolis, USA, October 2016. Industry track, short paper.

[12] Miyoung Han, Pierre-Henri Wuillemin, and Pierre Senellart. Focused Crawling through Reinforcement Learning. In *Proc. ICWE*, pages 261–278, Cáceres, Spain, June 2018.

[13] Debabrota Basu, Qian Lin, Weidong Chen, Hoang Tam Vo, Zihong Yuan, Pierre Senellart, and Stéphane Bressan. Regularized Cost-Model Oblivious Database Tuning with Reinforcement Learning. *Transactions on Large-Scale Data and Knowledge-Centered Systems*, 28:96–132, 2016.

[14] Debabrota Basu, Pierre Senellart, and Stéphane Bressan. BelMan: Bayesian Bandits on the Belief–Reward Manifold. In *Proc. ECML/PKDD*, pages 167–183, Würzburg, Germany, September 2019.