Data Science with Human-in-the-loop

PI: Michel Dumontier, Postdoc: Amrapali Zaveri

Enormous amounts of Life Science data, with valuable information, have been and are being produced at an unprecedented rate by researchers all over the world. Re-using this data requires complete, accurate, consistent i.e. good quality metadata. Currently, quality of the data as well as the metadata is extremely poor. Employing domain experts is expensive and time consuming to perform the curation and machines cannot always detect the problems that require human judgment. Consequently, crowdsourcing has emerged as a means of involving non-experts to carry out human intelligent tasks. Crowdsourcing is when you take a big task, break it down into smaller micro pieces and then send it out to a large group of people (non-experts) to do it in parallel at a minimal cost. This ensures that the tasks are time-efficient and inexpensive to execute. This thesis will involve developing algorithms and tools focused on combining machine-learning as well as crowdsourcing i.e. non-experts, in an inexpensive, time-efficient and scalable manner. The results of such collaborations will accelerate scientific discovery, which will shorten time to market for new healthcare products and services. Additionally, the tool will ensure that curated good quality datasets are made available using the FAIR principles, which will enable healthcare policy makers and decision makers make more informed and data-driven decisions regarding the investment of their resources.

Publications:

1. Crowdsourcing linked data quality assessment. M Acosta, **A Zaveri** , E Simperl, D Kontokostas, S Auer, J Lehmann. International Semantic Web Conference, 260-276. **Cited by 81.**
2. User-driven quality evaluation of DBpedia. **Amrapali Zaveri** , Dimitris Kontokostas, Mohamed A Sherif, Lorenz Bühmann, Mohamed Morsey, Sören Auer, Jens Lehmann. Proceedings of the 9th International Conference on Semantic Systems 2013. **Cited by 86** .
3. Quality assessment for linked data: A survey. **A Zaveri** , A Rula, A Maurino, R Pietrobon, J Lehmann, S Auer. Semantic Web 7 (1), 63-93. **Cited by 214** .
4. Triplecheckmate: A tool for crowdsourcing the quality assessment of linked data. Dimitris Kontokostas, **Amrapali Zaveri,** Sören Auer, Jens Lehmann. International Conference on Knowledge Engineering and the Semantic Web, 2013. **Cited by 37** .
5. Ranking adverse drug reactions with crowdsourcing. A Gottlieb, R Hoehndorf, M Dumontier, RB Altman/ Journal of medical Internet research 17 (3) . **Cited by 12** .