Ikhlas Attarwala
MSAI 339 – Data Science Seminar
10-16-18 – Background Reading

Summarizing "Data Curation at Scale: The Data Tamer System" by Stonebraker, Bruckner, …

Michael Stonebraker, Daniel Bruckner and a team of scientists introduced a new form of data curation at M.I.T. Brandeis & QCRI, called 'Data Tamer'. Data curation is the process of cleaning, transforming, deduplicating and integrating new and various sources of data into an end-to-end system. They stress that aggregating tens of thousands to more of curations manually is not feasible, and thus requires the need for machine learning. Data Tamer is a system that accepts a variety of sources of data that identifies attributes, tabulates them, rids the data of duplicate information and even creates a visual representation of the data to assist us in understanding the data source. The authors believe that their project, the Data Tamer, can solve issues present in other systems, such as through automated algorithms for large-scale projects, by cleaning through inaccuracy and imperfect data, easily teaching less-skilled programmers, and open for continual modification.

One of the strengths of this research proposal include a very thorough explanation of all the uses for Data Tamer, along with reasons for why their technique is far superior to the next. In Figure 2 on page 8 for example, they show the near perfect precision of Data Tamer's reporting (98.9%) and deduplication (100.0%) in 50 data sources, citing correctly over 180,000 duplicates. In addition, the authors include multiple example applications of different fieldwork to showcase what Data Tamer can be used in (section 2, page 2). This is especially helpful and coaxing for readers that may be unfamiliar with the program's details, but have familiarity in the area of use described. Finally, Stonebraker & co. express the need for future enhancements, which is great for advertising a growing system and expect investing eyes, but also for the sake of research in data integration.

I found a couple weaknesses in this essay. One, oddly notably in a few of Stonebraker's research proposals I've come across, there are excessive uses of jargon such as using the term "offshore 'wrapper foundry'" (page 2). I'm not sure what he meant by this nor would I expect the average reader to. Secondly, while the paper does well to explain what Data Tamer is capable of, it doesn't entirely explain how the user would do so. If it would be as easy as allowing non-programmers to handle it effectively, it could be expected to include a guide or application screenshots for ease of use. Lastly, while the authors began and ended with bragging of Data Tamer's "better results at lower cost than their current solution" (page 2), they made fewer than expected comparisons of performance to other systems.

All in all, this was a well-written introduction for the Data Tamer System. I would expect future directions to follow the limitations already discussed throughout the paper but explicitly again on page 9. The largest of these issues I see is in data cleaning, where the system could expect problems with common, semantic value in information gathered, that could be useful for its users.