Shrashti Singhal
Santa Fe, New Mexico
SS55@illinois.edu

24th November 2019

Mr. Manjunath Bhide
L&T Technology Services
India

Dear Mr. Manjunath Bhide,

I take great pleasure in welcoming you to the Data Science Department and wish you success during your upcoming tenure here.

I see this as an appropriate opportunity to bring several matters to your attention regarding the field of Data Curation. I want to advocate for the continuation of data curation activities in our company. Our company is recognized for providing state-of-the-art engineering services in research and innovative problems. Our history has shown that such innovation has been possible due to the appropriate preservation, policies, and standards. These have maintained the provenance of our technology so far. It is essential that we maintain good data curation practices that allow us to build on top of our continuous achievements. This will also save us time and money, and also enable us to keep pushing forward in the innovation frontier.

As the quantity of data increases, so does the need for the managing & organizing this data. Data curation seeks to provide integrity, accessibility, and usability of data to enable research and development within the realm of data science. It is essential to dedicate time and resources to define our processes for handling and storing data so we can ensure that data is not lost, corrupted or misinterpreted. Information needs to be validated, and constraints can be applied to prevent inaccuracies, inconsistencies, and errors that would otherwise have to be resolved and cleaned upon use. Keeping our data organized, documented and groomed means less time spent curating and cleaning data during analysis of data science projects and research. It also means we can use our data with confidence, which is reliable and generated after sound curating processes. Data curation ensures that data can be reliably retrieved and maintained for present and future use.

Provenance is crucial for data-driven processes. As we use the results from our data to make decisions for the company and our clients, it's critical that we have the means to show our process of generating this data which guides these decisions. Ensuring provenance provides us with this transparency of data creating methods. Provenance supports the identification of inputs, calculations, and actions responsible for our data values. Provenance ensures that we have each step documented from input to output which ensures our processes are repeatable and auditable. It makes our procedures rigorous, which builds confidence in our reports and creates trust from transparency.

Preservation is another significant Data curation activity. It ensures that data will be understandable and usable in the future. For example, there are cases where a customer asks for a solution, and our company spends time, money and other resources to solve it. Very likely, the same customer comes again wanting for a similar solution. In such cases, it is essential to have a system that supports long-term preservation, so that we don't need to spend resources again. We can as well use the previous solution or work on modifying it rather than allocating resources from scratch. Preservation includes not just bit sequence preservation and syntax documentation, but also the documentation of semantics for data elements and the generation. Preservation of all metadata is needed to ensure that the data is usable and understandable, and can be authenticated and audited for provenance. This preservation system may consist of a framework to document contextual and procedural information. This data should be captured in metadata by following a set of standards.

We often miss the importance of data and how it can be used to empower us. At the same time, ignored and improper handled data can cause us trouble. Recent revelations showed that digital consultants of the Trump campaign misused the data of millions of Facebook users, which set off a furor on both sides of the Atlantic. Trump Campaign supporters with the help of Cambridge Analytica illegally obtained data from Facebook to build voter profiles. The news put Cambridge under investigation and thrust Facebook into its biggest crisis ever. Poor data management with regards to compliance and security led to a breach of data privacy which affected millions of users on Facebook and tarnished the reputation of such a pioneering company forever. Ninety-seven percent of Fortune 500 companies have been hacked. A very public and embarrassing example of this was Sony Pictures. This was due to lax security around the data and poor policies in place, including but not limited to retention schedules and proper disposition. Our department, definitely, cannot afford to neglect data curation and its overarching goals otherwise we risk a repetition of past mistakes.

In conclusion, it is evident that data is increasing exponentially, which means that any problem regarding data curation that will be an expensive problem to solve in the future. It is best that the goals of data curation are observed continuously by an active and separate Data Curation committee within this department to ensure that we 'curate' our data effectively to meet the demand and government regulations. Data curation is essential for reliable and efficient analysis. Moreover, most of the cost and workforce associated with using data is in curation, not in analysis. As a data manager from this industry I can tell you, it is curatorial work where industries make the most substantial investment of money, staff, time, and effort.

Without successful data curation, successful data analysis is not possible, it would be prohibitively expensive and dangerously unreliable, which can have an enormous societal impact due to the nature of work which our company undertakes. Given to all this, I believe our organization should prioritize funding for this vital area. Data curation is a primary asset to the continued success of our company, and stopping it could severely impact our performance. I implore you to take a moment to think about the consequences of improperly managed, erroneous or lost data. Data curation is an investment in the health and longevity of data, and it plays a fundamental part in guaranteeing the success to our organization's directives. Curating data is as important as collecting or analyzing it.

Sincerely,

Shrashti Singhal