TO: Mr.X , Director, Data Science Department

FROM: Ms.Y , Data Scientist

DATE: November 18, 2018

Subject: Requesting to continue the funding for data curational activities

Data curation is an essential part of data science. It ensures that the analysis we undertake, the findings and insights we come up with as a result of the analysis are meaningful and accurate. I see the reason for your concern here is that the funding for data curational activities in the past has been significantly higher than the funding for actual data analysis. But I can certainly vouch for the fact that the data science department's achievements towards bringing the company great profit in the past can be attributed to successful data curational activities. I would like to present to you, below, few key curatorial activities without which our analysis can be valueless.

Workflow:
With enormous amount of data flowing in every single day, due the heterogeneous nature of the data, it is very important for us to perform continuous transformations like integration or format conversions to make all the data that we have consistent. This transformation will be disastrous if not for the presence of data curatorial activities as data will contain lot of errors and sustain data losses during the migration. This can be easily avoided with the help of another data curation activity called "Identity". Identity ensures that the data is transferred correctly and there has been no data loss during the process.

Provenance and Metadata:
The process of transformation of data is an on-going process and it is necessary to document every step in the transformation. Provenance helps us build more understanding of the data in its original form and also a sense of trust that the data is not corrupted. The attitude of the employees in our company before the practice of provenance used to be like -"The process is on the top of my head and I can talk about it when needed. I don't have to waste my time on documentation." But after a period of time the details would fade away from memory and would be very hard to trace back the root of a data or an algorithm. The whole procedure and format of data would then be based on memory and would be very unreliable. But post introduction of metadata, data can be trusted. This also enables provenance. The algorithms can also be reused for different applications with minor tweaks, without the need to build the entire thing.

Thus, I request you to kindly consider all the possible effects of data curatorial activities on our department/company and come forward with a decision whether to release adequate funds to these activities.

With Regards,
Ms.Y