

**Memo (Part 2)**

ATTN: My new director

Dear Mr. Director,

As you are aware, data science is a very important field of our company's activities. It produces a solid basis for our operations, provides justification for many of our important decisions, and is of paramount importance for all data-related practices.

I would like to draw your attention to a very interesting fact. Data science can actually be defined as a combination of two activities: data curation and data analytics. The role of data curation is ensuring that data can be efficiently and reliably found and used, while data analytics is concerned with employing specific techniques to extract knowledge from data. Therefore, data curation is the first phase of data science, and any valid and justified data analysis is impossible without it.

I should also add that data curation enables the active and continuous management of data through its life cycle. Curation activities enable data discovery and retrieval, support quality, add value, ensure data re-use over time. All of this makes this "first stage" of data science a very important step on the path of getting the correct and valid results of data analysis. And, as you of course know, the competitiveness of our company is directly related to the quality of the analysis that we do. I am sure that without properly organized and managed data curation, our company will not implement its full potential and may get behind its competitors in a number of areas.

This is due to a simple fact that in the world of interdisciplinary challenges, curatorial activities provide us with means to manage large volumes of rapidly changing data which is often heterogeneous in nature and developed by different sources, and which "must be retrieved, authenticated, reformatted, integrated with other data and managed for effective use, and demonstrably reliable even after processing and preparation." It is needless to say that without successful data curation it is impossible to achieve successful data analysis. Otherwise, the latter becomes extremely expensive and very unreliable

Another interesting, but underestimated argument in favor of data curation is that most of the costs related to data use is associated with curation, and not analysis. Also, the much greater part of personnel focuses on curation, not analysis. I am sure that any data manager in our industry can confirm this to you right away.

I am aware of the fact that you are not sure if funding of curatorial activities should continue. Please consider these arguments when making your decision. I would like to assure you that data curation is the foundation of success in data science.

The key areas where data curation is very important for our company include, but are not limited to:

- Preservation – in order to ensure that data is understandable and usable for our future operations, we must maintain a strict, documented preservation strategy including not only preservation at the level of characters and syntax, but also the meaning of data and the generation and preservation of all metadata ensuring that the data is usable and understandable, can be validated and audited for provenance. We need to implement this strategy with discipline, documenting every action taken.
- Moreover, provenance is of paramount importance for proper identification of inputs, calculations, and actions responsible for data values in situations when one data set gets derived from another.

I hope that I have convinced you that data curation is too important to be disregarded and is extremely important for our mutual success as a company.

Respectfully,  
Andrew Nedilko