Ikhlas Attarwala
MSAI 339 – Data Science Seminar
10-02-18 – Background Reading

Summarizing the "What goes around, comes around" by Stonebraker, Hellerstein

Michael Stonebraker and Joseph Hellerstein collaborated on a paper titled 'What Goes Around Comes Around' to provide a thorough analysis of the various styles of data modeling, and to highlight that there may be a pattern in how we've proposed models over the course of 35 years from the late 1960s. It's worth noting that as it is currently 2018, this article seems to have been written around 15 years ago. Both authors begin their analysis by grouping all models into 9 distinct categories (into the 9 "eras" they were proposed in) and provide a breadth of knowledge over the history of that era, what variables were present in the proposed schema, how information was hierarchically depicted, they show examples to illustrate the scope of the program, discuss the history of its application, and more. As they reach the end segment of each "era", the authors show how data models from older eras (such as the "IMS era" from 1968) were considered far too complex to handle for their time, and couldn't bring a lot of traction or interest in then. The early IMS era proposed instances, values and keys in a record type arranged in a tree to log information. In the search for simplicity, data scientists moved on to a new era where they introduced a collection of information in a network as opposed to a tree. As time progressed, multivariate relational models were introduced, later we reached an object-oriented era, and then worked around semi-structured data (whether a schema should be proposed first for data to fall in and fit a model, or not be proposed first, in which case the data would be entered by the user themselves). Stonebraker and Hellerstein stress how important it was for us to simplify models then. Today, we not only have more data to use, but we have a need to again simplify from our current complex ones.

A few of the authors' strengths include how in-depth and informative this research article is, along with its structure and comprehensibility. As mentioned above, they cover every major development and descriptive detail in hierarchical models and data representations per era, but do so in a manner that does not disrupt the flow of the passage. As this passage is very well-structured, a reader does not need to have a background in data science to understand how models developed and improved over the 35 year span. Finally, the passage does not introduce jargon. Anything that needs to be properly defined or discussed further is explored further.

The authors don't make the best attempts to show reason for this case, however. For example when summarizing, they write that "To avoid repeating history, it is always wise to stand on the shoulders of those who went before, rather than on their feet." The authors neglected to use this space to actually discuss what not to do to avoid falling into the "cycle" of logical data independence and complexity. In addition, if they didn't want to see a repeat of the 1970s, they didn't make great comparisons between the IMS era to now. Yes, there was a large amount of complexity then and it was hard to understand what to do, but it would help to show more similarity between now and then to further their aim. The authors also fail to discuss further limitations, even though simplification is not the only factor that's changed since the IMS era.

While the authors mean well, this article didn't serve to scare people; it should, but it seemed more like a method to vent frustration. It would help their case if they directed this with the intent of showing more similarity to how we are repeating history and why it's for certain that we're acting recursively despite all else we've gained in the past 3 and a half decades.