Letter from the 2017 IEEE TCDE Early Career Award Winner

Rethinking Data Analytics with Humans-in-the-loop

From large-scale physical simulations, to high-throughput genomic sequencing, and from conversational agent interactions, to sensor data from the Internet of Things, the need for data analytics—extracting insights from large datasets—has never been greater. At the same time, current data analytics tools are powerless in harnessing the hidden potential within these datasets. The bottleneck is not one of “scale”—we already know how to process large volumes of data quickly—but instead stems from the *humans-in-the-loop*. As dataset sizes have grown, the time for human analysis, the cognitive load taken on by humans, and the human skills to extract value from data, have either stayed constant, or haven’t grown at a commensurate rate. Thus, at present, *there is a severe lack of powerful tools that incorporate humans as a “first-class citizen” in data analytics, helping them interactively manage, analyze, and make sense of their large datasets.*

My research has centered on the design of efficient and usable *Human-in-the-Loop Data Analytics* (HILDA) tools, spanning the spectrum from *manipulate* → *visualize* → *collaborate* → *understand*: (a) For users not currently able to even examine or *manipulate* their large datasets, I am developing DATASPREAD, a *spreadsheet-database hybrid* (dataspread.github.io). (b) Then, once users can examine their large datasets, the next step is to *visualize* it: I am developing ZENVISAGE, a *visualization search and recommendation system*, to allow users to rapidly search for and identify visual patterns of interest, without effort (zenvisage.github.io). (c) Then, to *collaborate* on and share the discovered insights with others, I am developing ORPHEUS, a collaborative analytics system that can *efficiently manage and maintain dataset versions* (orpheus-db.github.io). (d) Finally, to *understand* data at a finer granularity by using humans to annotate data for training machine learning algorithms, I am developing POPULACE, an *optimized crowdsourcing system* (populace-org.github.io).

Developing these HILDA tools requires techniques not just in database systems, but also in data mining and in Human-Computer Interaction (HCI)—we’ve had to evaluate our systems not just in terms of scalability and latency, but also accuracy and utility (from data mining), and interactivity and usability (from HCI). In developing these tools, we’ve also had to go outside of our comfort zone in talking to *real users*: biologists, battery scientists, ad analysts, neuroscientists, and astrophysicists, in identifying usage scenarios, pain-points, and challenges, thereby ensuring that our tools meet real user needs. Indeed, many of these individuals and teams have access to large datasets, and a pressing need to extract insights and value from them, but are not able to do so. This is due to the lack of powerful tools that can reduce the amount of human effort, labor, time, and tedium, and at the same time, minimize the need for sophisticated programming and analysis skills.

While our tools represent a promising start, we are barely scratching the surface of this nascent research field. Future research on HILDA will hopefully enable us to make steps towards meeting the grand challenge of empowering scientists, business users, consultants, finance analysts, and lay users with a new class of tools that equips them with what they need to manage, make sense of, and unlock value from data. We envision that data-driven discovery of insights in the future will no longer be bottlenecked on the “humans-in-the-loop”, and will instead depend on fluid interactions facilitated by powerful, scalable, usable, and intelligent HILDA tools.

Aditya Parameswaran
University of Illinois UC