What is Our Agenda for Data Science?

AnHai Doan
University of Wisconsin-Madison
anhai@cs.wisc.edu

Data science (DS) has emerged as a major interdisciplinary field, and has attracted much attention. Our field however has been slow in reacting to this development. In this abstract I argue that DS is here to stay and will become even more important, and that we have a lot to contribute, but if we do not ramp up our efforts, we risk becoming increasingly irrelevant.

In particular, I believe the time has come for us to perhaps develop a data science agenda that builds on our strengths, attracts a broad participation from our community, and helps us shape this emerging field. The set of topics that this agenda can discuss includes (but is not limited to) the following:

Scope of the Field. I propose to view DS as a field that develops principles, algorithms, tools, and best practices to manage data, focusing on three topics: (a) analyzing raw data to infer insights, (b) building data-intensive artifacts (e.g., recommender systems, knowledge bases), and (c) designing data-intensive experiments to answer questions (e.g., A/B testing). This "working definition" should be sufficient to get us started. Viewed as above, DS is clearly here to stay (even though the name may change), for the simple reason that everything is now data driven, and will only become even more so in the future.

Education. There is a great demand for learning DS at both undergraduate and graduate levels. Unfortunately currently we only have isolated course offerings, of varying contents. If we can (more or less) agree on a standard RDBMS curriculum (at least for undergraduates), we ought to be able to do the same for DS. In particular, current DS courses do not emphasize enough data wrangling (the stage where raw data is acquired, extracted, cleaned, transformed, and integrated), even though this step takes up to 80% of analysts' time. This is where our community (especially researchers in data cleaning/transformation/integration) can really contribute.

In general, we have done a good job training our students/workforce for RDBMSs. Can we do the same for DS? A huge number of non-CS students have also been flocking to CS departments wanting to learn DS. What and how should we teach them? Is it even our job? Many academic departments are also looking into setting up their own DS curriculum. What can we do to help them, and how?

WOODSTOCK '97 El Paso, Texas USA

DOI: 10.475/123_4

Research. I believe we need more effort in data wrangling, data integration, data analysis, scalable techniques, and building data-intensive artifacts such as knowledge bases (KBs), especially methods to build domain-specific KBs.

We probably should rethink how we do research in DS. To really make impacts here, I believe we need to devote far more effort to building systems. DS by nature is engineering. We need to build systems to help evaluate and drive research, to integrate disparate R&D efforts, and to make practical impacts. Thus, DS research must be tied more closely to system/tool building.

System/Tool Building. This is where we may need a "sea change". I believe that building stand-alone DS systems/tools as we have done so far will become increasingly difficult and will not maximize our impacts.

Instead, there are already very popular ecosystems of open-source DS tools out there, e.g., in R and Python. It may be worthwhile to explore these ecosystems, study them, then build tools in data wrangling/cleaning/integration/etc directly into these ecosystems. This way we can build tools faster, contribute to a growing ecosystem of practical DS tools, and at the same time help this ecosystem solve difficult challenges in developing data intensive tools.

Relationship with Other Communities. Many academic disciplines have growing DS needs, and have been looking to us for help. How can we help them with their research? With education? Should we explore providing DS services to other academic departments (e.g., consulting, DS tools, cloud-based DS services)? Should we play a leadership role in setting up Data Science Institutes at universities, and if so, what should these institutes do?

My experience in the past few years suggests that while these academic disciplines are still looking to us for support, they have also been actively doing DS on their own: educating themselves in DS, setting up DS training programs, building domain-specific DS tools, etc. Likewise, while we can still contribute much to the various ecosystems of DS tools (e.g., in R/Python), they have also been growing rapidly without us. As a result, if we do not ramp up our efforts, and think strategically about what we want to do, we risk becoming increasingly irrelevant in this fast growing interdisciplinary field.