Databases and AI: The Twain Just Met

… after no progress for two decades, a challenge by

Michael L. Brodie, Mark Greaves, James A. Hendler

Since its inception the database community has focused predominantly on engineering aspects of data management, with particular interest in data engineering at scale, but with little focus on the meaning of the data. Since Ted Codd introduced the relational data model (RDM) in 1969, the research and industrial database communities focused on relational technology and within the reasoning and expressive power of the RDM and its underlying First Order Logic (FOL). Conversely, the Artificial Intelligence (AI) community focused predominantly on more powerful and expressive reasoning with little focus on engineering aspects especially for problems at (database-) scale, and did not adopt standards for the representation of AI data at scale.

Based on the premise that database applications required more expressive power than that offered by relational data and the RDM there were attempts in the late 1980s and early 1990s to investigate the semantics of data in databases through semantic data models and conceptual modelling. Despite the amount of work on database semantics there was no impact in research or industry.

In the 1980s and 1990s there were attempts to understand or broach the data engineering-semantics divide between leaders in the Database and AI communities also with little if any impact. The database community continued making significant improvements in engineering relational databases at vastly increasing scale. Yet the database community made almost no progress on dealing better with database semantics. In 2011 relational technology dominates the research and industrial data management communities with essentially no use of even modest RDM extensions such as the Entity-Relational data model. Practical semantic integration of data is largely done manually, by groups of application-specific experts sitting around tables and laboriously iterating solutions. Conversely the AI community continued to make advances in knowledge representation and automated reasoning with essentially no practical contributions to reasoning at scale.

The lack of interaction between the Database and AI communities was driven by more than just the engineering-semantics divide between the disciplines. In the 1980s and 1990s there was little overlap in the basic problem domains addressed by the two communities. Database applications involved simple, relational data at scale while AI applications involved more complex, e.g., modal, reasoning over data that was complex or uncertain but at modest scale. AI and Databases lacked a shared use case.

The 1991 launch of the Web changed everything. Most dramatically, the machine learning community within AI began to be flooded with textual data, and in response started to pay serious attention to data engineering issues to develop their scalable, learning-based text processing algorithms. The symbolic AI community also began to address data issues more directly when they started to apply less expressive description logic technologies to Web markup in what a decade later (2001) was called The Semantic Web. The evolving Semantic Web has grown to include use cases that require accessing and reasoning over billions of heterogeneous data elements. Yet 20 years on (2011), web-scale data engineering challenges of the Semantic Web, including the popular Linked Open Data (LOD) approach, remain open issues. Realizing the Semantic Web requires addressing data engineering, management, query, and integration aspects of Web-scale distributed data over vast numbers of heterogeneous data sources, as well as addressing traditional AI challenges such as search and reasoning – in the same computational space. Mostly, though, AI researchers continue to hope that the database community will simply provide optimal out-of-the-box solutions to their data problems. Continuing to ignore issues of data engineering at scale may relegate AI to another AI Winter since the Semantic Web cannot be realized without addressing issues of scale.

The database community faces data engineering and data integration challenges similar to those of the Semantic Web. Large enterprises have vast numbers of semantically heterogeneous databases, few of which are purely relational having been extended with 100s of attributes. High-end database applications now participate in information ecosystems that require data to be meaningfully integrated from 100s or 1,000s of databases. At the low end, the proliferation of lightweight web-driven mashups driven by data sources like Amazon and Google Maps are transforming the larger data environment and

challenging traditional database doctrines like transactional integrity and single version of truth. Integrating semantically heterogeneous data sources pose challenges of the meaning of data that faster, more efficient relational database engines do not address. Continuing to ignore semantics may relegate the database community to the status of plumbers.

Under the *Patient Protection and Affordable Care Act* and the *Health Information Technology for Economic and Clinical Health* (HITECH) *Act*, the future US Healthcare Systems will be driven by the quality of healthcare delivered. The PCAST report, *Report to the President: Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward,* envisions the US Healthcare System being driven, i.e., automated, by a set of Web-scale data integration experts, called data element access services (DEAS), that access and integrate data from millions of heterogeneous data sources distributed across US Sates and territories world-wide. The PCAST data-centric solution involves all healthcare data, e.g., patient records, lab tests, diagnoses, drug treatments, and clinical research data, possibly residing at their places of origin and tagged with meta-data, such as provenance, to facilitate discovery and exchange using a universal exchange language for healthcare information and an infrastructure for locating relevant data. This vast information ecosystem is one of many emerging or in operation today; others include advertising, e-commerce, and social network data. Collectively these information ecosystems are the missing shared use case that demands data management solutions at massive scale from the database community and powerful search and reasoning solutions from the AI community over the same data space.

Our Digital Universe, increasingly composed of information ecosystems, is on the brink of the data engineering-semantics divide. Few realize that there is a divide or that there is a problem at all. Most users naïvely assume that information in the Digital Universe is correct and instantly accessible. Most database researchers do not understand the challenges of integrating semantically heterogeneous data or the conditions under which the problems arise. The industrial users of data may be even more naïve. The fastest growing segment in computing is Big Data. The fastest growing segment in Big Data is Business Intelligence and Analytics that are built on data warehouses that are built on databases. Most data warehouses are loaded automatically from a large number of semantically heterogeneous data sources. Automatic integration of semantically heterogeneous data into a data warehouse cannot resolve deeper issues of semantic heterogeneity, hence the resulting data is unlikely to be semantically correct, e.g., a single version of truth under FOL, which is the underlying assumption of relational technology.

The data engineering-semantics divide can be clearly seen right in the heart of data use cases in our data-driven Digital Universe. The database community is continuing to shut their eyes tightly and concentrate on the traditional use cases at which databases excel. Conversely, the AI community may have deep insight into challenges of reasoning across heterogeneous spaces and are exploring solutions for the Semantic Web, but most AI researchers still do not understand data engineering at scale. Currently there is essentially no industrial Semantic Web community. Without a data engineering solution for reasoning over Web-scale data, there never will be.

It's time for the database and AI communities to come together after two decades, and work together on issues of common concern to bridge the data engineering-semantics divide.