Opportunities for Data Management Research in the Era of Horizontal AI/ML

Theodoros Rekatsinas
University of Wisconsin-Madison
rekatsinas@wisc.edu

Sudeepa Roy
Duke University
sudeepa@cs.duke.edu

Manasi Vartak
Verta.AI
manasi@verta.ai

Ce Zhang
ETH Zurich
ce.zhang@inf.ethz.ch

Neoklis Polyzotis
Google Research
npoly@google.com

ABSTRACT

AI/ML is becoming a horizontal technology: its application is expanding to more domains, and its integration touches more parts of the technology stack. Given the strong dependence of ML on data, this expansion creates a new space for applying data management techniques. At the same time, the deeper integration of ML in the technology stack provides more touch points where ML can be used in data management systems and vice versa.

In this panel, we invite researchers working in this domain to discuss this emerging world and its implications on data-management research. Among other topics, the discussion will touch on the opportunities for interesting research, how we can interact with other communities, what is the core expertise we bring to the table, and how we can conduct and evaluate this research effectively within our own community. The goal of the panel is to nudge the community to appreciate the opportunities in this new world of horizontal AI/ML and to spur a discussion on how we can shape an effective research agenda.

PVLDB Reference Format:
T. Rekatsinas, S. Roy, M. Vartak, C. Zhang, and Neoklis Polyzotis. Opportunities for data management research in the era of horizontal AI/ML. *PVLDB*, 12(12): 2323-2324, 2019.
DOI: https://doi.org/10.14778/3352063.3352149

1. THE ERA OF HORIZONTAL ML/AI

ML/AI is applied in an increasing range of domains, e.g., medical diagnosis, transportation, farming, chip design, and scientific research, to name a few. In each of these domains, the application of ML/AI is transformational, as the ongoing advances in ML (mostly on deep neural networks) are enabling big leaps in terms of domain-specific technologies and functionality.

Proceedings of the VLDB Endowment, Vol. 12, No. 12
ISSN 2150-8097.
DOI: https://doi.org/10.14778/3352063.3352149

In parallel to this expansion, ML is being integrated in larger parts of the technology stack. Machine-learned models have successfully replaced parts of the stack, from smaller-scope modules (e.g., replacing heuristic strategies for caching) to larger subsystems (e.g., using models for end-to-end translation). In addition, the touch points to ML have spurred new research directions in other communities, e.g., hardware accelerators for ML.

Overall, the emerging reality is that ML is becoming "horizontal": it applies to more domains, it works in increasingly diverse hardware and software environments, and it has acquired a larger user base with a diverse skill set. At a high level, this is good news for the data-management community. One obvious benefit is that it becomes easier to integrate ML in our own stack and thus enable improvements and simplification to internal components. Perhaps more importantly, our community gains an expanded impact surface due to the close relationship between ML and data management. For instance, it is well known that high-quality data is critical for training good models, and also that ML pipelines are themselves data flows that resemble and extend data-analysis flows that we know how to optimize. Techniques and methods for data management in these contexts have already proven valuable for ML, and now we have the opportunity to expand our research to even more contexts in which ML is applied.

What are some concrete directions for new and exciting data-management research in this new era of horizontal ML/AI? How do we engage with the ML community to popularize our research? How do we ensure that our community is able to conduct and evaluate this research in the first place, and what should we do to better prepare new researchers and practitioners in our domain? These are some of the questions that we will attempt to answer in this panel.

2. OPPORTUNITIES

In traditional software development when the *capacity* of programming models increases, the users resort to *software engineering* for guidance to deal with the ever-increasing complexity of software systems. The development of machine learning applications ("Software 2.0" as many people are calling it) is no different. The last decade has witnessed the dramatic increase of the *capacity* of ML systems — in terms of its speed and the level of automation. This makes it increasingly easy for the user to get *some* ML models out

of these systems, but does not necessarily make the whole development process beyond a single model easier. Future ML systems need to manage the whole ML development cycle, much like how software engineering guides the development process of traditional software ("Software Engineering 2.0" for the development of "Software 2.0"). This calls for research efforts all the way from feasibility study, model development, model maintenance, model/data validation and testing, sharing and versioning of datasets, building robust data pipelines for ML, model debugging and profiling (which often requires fast querying capabilities on non-relational data) to continuous delivery, monitoring, and integration.

The data management community is uniquely equipped to address these questions given the indelible link between data and the development of ML models. However, the context of ML introduces new technical wrinkles that require further investigation. Take for example the problem of data validation, which has been extensively studied in data management systems. For ML, it becomes necessary to go beyond existing theory and develop robust systems that support continuous operation. To this end, we need new models that extend and combine the relational model with statistical models and causal models. In turn, these extensions require connecting the relational model (and algebra) with probabilistic inference, the ability to reason about robustness in the presence of noise, and frameworks that support interpretability for ML models. All in all, this calls for a new theory that is closer to modern analytics and goes beyond relational algebra and graph processing.

Exploring these new directions must take into account a critical principle related to the application of ML and AI: responsible use of data. The topic of ethical and responsible data science is increasingly being discussed and studied in the ML/AI and the data management communities. Clearly, when data is overwhelmingly being used to help the process of decision making, both for better efficiency and better accuracy, the question of its societal, financial, or judicial aspects is unavoidable. Several notions of fairness have been studied in the ML/AI literature, and more recently in the data management literature, for common data science applications like predictions and recognition. The study of causality is also relevant – while significant effort in big data applications has gone into predictive analysis, we need to be careful that 'correlation does not imply causation', especially in this era of democratization of data where not all data enthusiasts are trained in computer science or statistics. Causality has been studied extensively in the Statistics, AI, and social science communities, where it has been established that even without randomized controlled experiments, and just from observational data as used in the database community, causality can be inferred under certain conditions and assumptions. Several major areas studied in the data-management community, such as data cleaning, recommendations, privacy, crowdsourcing, information extraction and retrieval, and data mining, should be revisited to study whether they are fair and transparent. Being able to answer 'why' an output is produced for applications in these areas, by incorporating research in fairness and causality, can take us a step further in the direction of interpretability and transparency.

3. CHALLENGES

The proposed agenda raises an inevitable question: how do we engage with the ML community so as to popularize this research? A clear-cut answer might be hard to find, but some challenges and tradeoffs are immediately obvious.

One challenge is that our community may not be adequately trained to evaluate this research, particularly if it involves deep dives into the ML domain. This pertains to both the technical content (e.g., evaluating the novelty and technical merit of ML-related techniques) and to the domain (e.g., understanding the importance or relevance of a problem statement in the context of ML). One particularly bad outcome would be if data-management researchers reinvented techniques already used in the ML community, or worked on irrelevant problems. Moreover, it is unclear whether we have adequate methodologies (e.g., metrics or benchmarks) to experimentally evaluate this type of work.

One way to address this challenge is to engage with the ML community and perhaps favor different publication venues. This may not be full proof, as ML reviewers may lack the background to evaluate the novelty and merit of data-related techniques. One interesting middle ground might be newly emerging conferences like SysML which lie in the intersection of the two communities, but this can end up being either the best or the worst of both worlds.

Finally, how do we help new researchers and practitioners come up to speed so that they can contribute towards this new agenda? "More ML courses!" is an obvious answer, but it might also miss some important nuances. ML/AI is a large, fast-evolving field. What type of training can provide adequate coverage?