Automation of Data Prep, ML, and Data Science: New Cure or Snake Oil?

Arun Kumar
University of California, San Diego
United States of America
arunkk@eng.ucsd.edu

ABSTRACT

As machine learning (ML), artificial intelligence (AI), and Data Science grow in practical importance, a large part of the ML/AI software industry claims to have built tools and platforms to automate the *entire workflow* of ML. That includes vexing problems of *data preparation (prep)*, studied intensively by the database (DB) community for decades, with basically no resolution so far. Such claims by the ML/AI industry face a stunning lack of scientific scrutiny from the DB and ML research worlds, largely due to the lack of meaningful, large, and objective *benchmarks*. As such tools rapidly gain adoption among enterprises and other customers, this panel will debate whether the new ML/AI industry is basically selling "snake oil" to such users, how to evolve away from the status quo by instituting meaningful new benchmarks, creating new partnerships between industry and academia for this, and other pressing questions in this important arena. We aim to spur vigorous conversations that will hopefully lead to genuine new cures for an age-old affliction in Data Science.

ACM Reference Format:
Arun Kumar. 2021. Automation of Data Prep, ML, and Data Science: New Cure or Snake Oil?. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21), June 20–25, 2021, Virtual Event, China.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3448016.3457537

1 THE NEW DATA PREP BATTLEFIELD

For almost 30 years, the DB / data management community has intensively studied the vexing pains of data integration, cleaning, and transformation. This research has largely been in the contexts of RDBMSs, SQL-oriented business intelligence (BI), and knowledge base construction. But as the emerging interdisciplinary field of Data Science gains prominence, the massive pain of such data "grunt work" in the context of machine learning (ML) and artificial intelligence (AI) applications has taken center stage [9].

Surveys show data scientists spend large amounts of time (e.g., 45% [3], 60% [4], or worse!) on data grunt work, often loosely dubbed *data preparation* (prep). Sadly, the DB community's tools and techniques, primarily logic-based, have failed to improve the lives of most data scientists. To paraphrase Ihab Ilyas, a leading expert in data cleaning research: "decades of research, tons of papers, but very little success in practical adoption" summarizes the state of affairs [10].

Naturally, new lines of research use different vantage points: a human-in-the-loop self-service approach (e.g., Tableau Prep and Trifacta), automation using ML/deep learning models, automation using program synthesis techniques, and various hybrid approaches. But out in the ML/AI marketplace a new breed of "end-to-end" Automated ML (AutoML) platforms are promising the moon with almost no scrutiny or participation from the DB world: the *entire workflow* from raw data to ML model will be automated, including data prep! Examples include Salesforce Einstein, Alteryx, DataRobot, H2O Driverless AI, and emerging tools from the cloud "whales."

Such platforms are rapidly growing in adoption among small-and-medium enterprises [9]. For instance, Salesforce Einstein alone is apparently used on "hundreds of thousands of datasets" by enterprises [1]. Such customers either cannot afford data scientists (e.g., small city governments) or they use such tools for first-cut proofs of concept. Alas, there is no objective data on how "good" these tools are, especially on data prep. The ML and data mining worlds have studied to death the automation of the ML algorithmics part of the pipeline via so-called "AutoML heuristics" for feature engineering/extraction, hyperparameter tuning, and algorithm/architecture selection [8]. But the implications of how *data prep automation* and its failures affect end-to-end AutoML pipelines are shockingly ill-understood.

The above status quo has led to serious questions being raised many researchers and practitioners on whether enterprise and other customers are being sold effectively "snake oil" by much of this newly ascendant ML/AI industry [2, 11, 20, 22]. How to evolve away from this abysmal status quo? As David Patterson famously put it: "benchmarks shape a field . . . good ones accelerate progress . . . bad ones help sales." [13]. Both the DB and ML communities have long studied and valued benchmarks: TPC and ImageNet revolutionized the RDBMS and ML worlds, respectively. Yet, curiously no such benchmarks of renown exist for automated data prep, certainly not in this fast-growing Data Science arena, although some exist for the ML algorithmics part of the pipeline [12, 19]. All this poses many urgent questions for the DB research community:

(1) Is the ML/AI industry basically selling "snake oil," especially on the automation of data prep in so-called AutoML platforms? What kinds of users in terms of technical and/or domain expertise can *reliably* benefit from such tools?

(2) Why is the DB research community mostly slumbering in this fast-growing Data Science arena in contrast to its continued obsession with traditional SQL analytics and BI applications? Which parts of this space can learn from the long

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGMOD '21, June 20–25, 2021, Virtual Event, China
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8343-1/21/06.
https://doi.org/10.1145/3448016.3457537

history of prior work, including failed ideas, in the DB world for various kinds of users?

(3) What is the analogue of ImageNet and TPC for data prep in Data Science? There has been a lot of talk in the DB world on *benchmark datasets* for data prep, e.g., on blogs [21] and at NSF meetings [17] but nothing concrete yet. Why? What are the roadblocks involved and how to tackle them? Or is the DB community just full of "hot air" on benchmarks?

(4) How to incentivize DB researchers, ML researchers, and industry to work together on instituting such rigorous Data Science benchmarks and not just keep publishing narrow, "algorithmically novel," and often glib "delta" papers? Likewise how to incentivize the ML/AI industry to take such benchmarks seriously?

(5) Are human-in-the-loop approaches overrated in this space given the massive scale and low-expertise target user base of AutoML platforms? Does this approach not violate the promise of "end-to-end" automation? Or was this promise a pipe dream all along? Is it even possible to "benchmark" human-in-the-loop approaches in scientifically valid and reproducible manners?

(6) Many AutoML platforms use ad hoc and brittle rule-based heuristics for automating data prep. How promising are ML/deep learning and/or program synthesis techniques to supplant such heuristics? What are the challenges involved and how to tackle them? How to assess their real-world effectiveness and robustness, including the impact of their failures on downstream AutoML heuristics and enforcement of emerging criteria such as fairness of ML predictions?

2 PANELISTS

Felix Naumann. Hasso Plattner Institute and University of Potsdam.
Bio: He works on prep for file ingestion and has recently surveyed commercial data prep tools [6].
Rationale: Perspectives from his research and community organization; as PC co-chair for VLDB'21, along with Luna, he introduced the Benchmarks dimension to VLDB's scope and started the Scalable Data Science Research category.

Ihab Ilyas. Inductiv (based on HoloClean [14] and acquired by Apple), Tamr, and University of Waterloo.
Bio: He works on logic-based methods and ML methods for data prep and has product experience in enterprise data software.
Rationale: Perspectives from his research and his companies' customers.

Joseph Hellerstein. Trifacta and University of California, Berkeley.
Bio: He works on human-in-the-loop and program synthesis methods for data prep, has worked on platforms for enterprise ML (Apache MADlib [7]), and has product experience in enterprise data software.
Rationale: Perspectives from his research and his company's customers.

Sarah Catanzaro. Amplify Partners.
Bio: She invests in and advises high-potential startups in machine intelligence, data management, and distributed systems. She has also defined data strategy and led data science teams at startups and in the defense/intelligence sector.
Rationale: Perspectives bridging the worlds of research and industry, including through investments in data/AI software startups and interactions with their customers.

Xin Luna Dong. Amazon.
Bio: She works on ML/deep learning methods for data prep and has product experience in knowledge extraction, integration, cleaning, and mining across both Google and Amazon.
Rationale: Perspectives from her research and from major Web companies; as PC co-chair for VLDB'21, along with Felix, she introduced the Benchmarks dimension to VLDB's scope and started the Scalable Data Science Research category.

3 PANEL CHAIR

Bio snippet relevant for this panel: Arun's main research interests are in data management and systems for ML/AI analytics, focusing on issues of usability, scalability, and resource efficiency. Along with his PhD advisee, Vraj Shah, and other students, his recent work on the "ML Data Prep Zoo" is creating new labeled datasets and benchmarks for data prep automation [15, 16]. Google is exploring adoption of models from that work for the TensorFlow Extended platform. He has given invited talks on this effort at a joint seminar organized by the University of Wisconsin-Madison and Microsoft [18] and to Google Cloud (BigQuery and AutoML teams).

Panel chair experience: He moderated an acclaimed, provocative, and educational panel discussion at the SIGMOD Workshop on Data Management for End-to-End Machine Learning (DEEM) in 2018, also on the intersection of the DB and ML areas [10].

Community organization: He is an Associate Editor for VLDB's Scalable Data Science Research category [5], both in its inaugural year of 2021 and in 2022. He helped shape its rationale and criteria. He co-organized the SIGMOD DEEM Workshop in 2018.

4 ACKNOWLEDGMENTS

Thank you to all the panelists, SIGMOD 2021 Panels Chairs (Sam Madden and Tiziana Catarci), and Vraj Shah for their feedback on an earlier version of this document. Thank you to Michael Stonebraker for initiating a discussion on the CIDR 2021 Slack workspace on this topic and to Felix Naumann for helping channel those conversations into this panel proposal. The author's work on this topic is funded in part by an NSF OIA C-Accel grant under award number 2040727 and gifts from Amazon, Google, Oracle, and VMware. The content is solely the responsibility of the author and does not necessarily represent the views of any of these people or organizations.

REFERENCES

[1] 2019. Personal conversation with ML engineers at Salesforce Einstein.

[2] Ahmed Abbasi, Brent Kitchens, and Faizan Ahmad. 2019. The Risks of AutoML and How to Avoid Them. https://hbr.org/2019/10/the-risks-of-automl-and-how-to-avoid-them

[3] Datanami.com Alex Woodie. 2020. Data Prep Still Dominates Data Scientists' Time, Survey Finds. https://www.datanami.com/2020/07/06/data-prep-still-dominates-data-scientists-time-survey-finds/

[4] Figure Eight. 2016. CrowdFlower Data Science Report. https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

[5] Alon Halevy, Arun Kumar, and Nesime Tatbul. 2020. Scalable Data Science: A New Research Track Category at PVLDB Vol 14 / VLDB 2021. https://wp.sigmod.org/?p=3033

[6] Mazhar Hameed and Felix Naumann. 2020. Data Preparation: A Survey of Commercial Tools. *SIGMOD Rec.* 49, 3 (Dec. 2020), 18–29.

[7] Joseph M. Hellerstein, Christoper Ré, Florian Schoppmann, Daisy Zhe Wang, Eugene Fratkin, Aleksander Gorajek, Kee Siong Ng, Caleb Welton, Xixuan Feng, Kun Li, and Arun Kumar. 2012. The MADlib Analytics Library: Or MAD Skills, the SQL. *Proc. VLDB Endow.* 5, 12 (Aug. 2012), 1700–1711.

[8] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren (Eds.). 2018. *Automated Machine Learning: Methods, Systems, Challenges.* Springer. In press, available at http://automl.org/book.

[9] Kaggle. 2020. State of Data Science and Machine Learning. https://www.kaggle.com/kaggle-survey-2020

[10] Arun Kumar. 2018. ML/AI Systems and Applications: Is the SIGMOD/VLDB Community Losing Relevance? https://wp.sigmod.org/?p=2454

[11] Hilary Mason. 2021. Twitter. https://twitter.com/hmason/status/1363924362659782657?s=20

[12] OpenML. 2021. Website. https://www.openml.org/

[13] David Patterson. 2001. How to Have a Bad Career How to Have a Bad Career in Research/Academia. https://people.eecs.berkeley.edu/~pattrsn/talks/BadCareer.pdf

[14] Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, and Christopher Ré. 2017. HoloClean: Holistic Data Repairs with Probabilistic Inference. *Proc. VLDB Endow.* 10, 11 (Aug. 2017), 1190–1201.

[15] Vraj Shah and Arun Kumar. 2019. The ML Data Prep Zoo: Towards Semi-Automatic Data Preparation for ML. In *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning (DEEM'19).* Association for Computing Machinery, Article 11, 4 pages.

[16] Vraj Shah, Jonathan Lacanlale, Premanand Kumar, Kevin Yang, and Arun Kumar. 2021. Towards Benchmarking Feature Type Inference for AutoML Platforms. In *ACM SIGMOD.* https://adalabucsd.github.io/sortinghat.html.

[17] Lisa Singh, Amol Deshpande, Wenchao Zhou, Arindam Banerjee, Alex Bowers, Sorelle Friedler, H.V. Jagadish, George Karypis, Zoran Obradovic, Anil Vullikanti, and Wangda Zuo. 2019. NSF BIGDATA PI Meeting - Domain-Specific Research Directions and Data Sets. *SIGMOD Rec.* 47, 3 (Feb. 2019), 32–35.

[18] UW-Madison and Microsoft. 2020. Machine Learning Optimized Systems. https://remziarpacidusseau.wixsite.com/mlos

[19] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. 2013. OpenML: Networked Science in Machine Learning. *SIGKDD Explorations* 15, 2 (2013), 49–60.

[20] Denis Vorotyntsev. 2019. Towards Data Science: AutoML is Overhyped. https://towardsdatascience.com/automl-is-overhyped-1b5511ded65f

[21] Gerhard Weikum. 2013. Where's the Data in the Big Data Wave? http://wp.sigmod.org/?p=786

[22] Doris Xin, Eva Yiwei Wu, Doris Jung-Lin Lee, Niloufar Salehi, and Aditya Parameswaran. 2021. Whither AutoML? Understanding the Role of Automation in Machine Learning Workflows. In *CHI.* https://arxiv.org/abs/2101.04834.