Will LLMs reshape, supercharge, or kill data science? (VLDB 2023 Panel)

Alon Halevy
Meta Reality Labs
ayh@meta.com

Yejin Choi
University of Washington and AI2
yejin@cs.washington.edu

Avrilia Floratou
Microsoft
avflor@microsoft.com

Michael J. Franklin
University of Chicago
mjfranklin@uchicago.edu

Natasha Noy
Google
noy@google.com

Haixun Wang
Instacart
haixun@gmail.com

ABSTRACT

Large language models (LLMs) have recently taken the world by storm, promising potentially game changing opportunities in multiple fields. Naturally, there is significant promise in applying LLMs to the management of structured data, or more generally, to the processes involved in data science. At the very least, LLMs have the potential to provide substantial advancements in long-standing challenges that our community has been tackling for decades. On the other hand, they may introduce completely new capabilities that we have only dreamed of thus far. This panel will bring together a few leading experts who have been thinking about these opportunities from various perspectives and fielding them in research prototypes and even in commercial applications.

BIOGRAPHIES

Moderator

Alon Halevy is a director at Meta's Reality Labs Research, where he works on Human Value Alignment, the combination of neural and symbolic techniques for data management and on responsible personal information management. Prior to Meta, Alon was the CEO of Megagon Labs (2015-2018) and led the Structured Data Group at Google Research (2005-2015), where the team developed WebTables and Google Fusion Tables. From 1998 to 2005 he was a professor at the University of Washington, where he founded the database group. Alon is a founder of two startups, Nimble Technology and Transformic Inc. (acquired by Google in 2005). Alon co-authored two books: The Infinite Emotions of Coffee and Principles of Data Integration. In 2021 he received the Edgar F. Codd SIGMOD Innovations Award. Alon is a Fellow of the ACM and a recipient of the PECASE award and Sloan Fellowship. Together with his co-authors, he received VLDB 10-year best paper awards

for the 2008 paper on WebTables and for the 1996 paper on the Information Manifold data integration system.

Panelists

Yejin Choi is Wissner-Slivka Professor and a MacArthur Fellow at the Paul G. Allen School of Computer Science & Engineering at the University of Washington. She is also a senior director at AI2 overseeing the project Mosaic and a Distinguished Research Fellow at the Institute for Ethics in AI at the University of Oxford. Her research investigates if (and how) AI systems can learn commonsense knowledge and reasoning, if machines can (and should) learn moral reasoning, and various other problems in NLP, AI, and Vision including neuro-symbolic integration, language grounding with vision and interactions, and AI for social good. She is a co-recipient of 2 Test of Time Awards (at ACL 2021 and ICCV 2021), 7 Best/Outstanding Paper Awards (at ACL 2023, NAACL 2022, ICML 2022, NeurIPS 2021, AAAI 2019, and ICCV 2013), the Borg Early Career Award (BECA) in 2018, the inaugural Alexa Prize Challenge in 2017, and IEEE AI's 10 to Watch in 2016.

Avrilia Floratou is a Principal Scientist Manager at Microsoft's Gray Systems Lab (GSL), an applied research group working at the intersection of databases/systems and machine learning. Her research broadly lies in the area of data management with a recent focus on leveraging Large Language Models to improve user productivity in domains, such as data integration, data exploration and code migration. Avrilia and her team are actively collaborating with various product teams in Azure Data to deliver LLM-based technologies to Microsoft customers. Her research interests also include semantics-aware data science, large-scale stream processing, and benchmarking Big Data platforms. Prior to her current role, she was a research scientist at IBM Almaden Research Center working on SQL-on-Hadoop engines and natural language interfaces for databases. Avrilia received her Ph.D. from University of Wisconsin-Madison and a B.S in computer science from University of Athens, Greece.

Michael J. Franklin is the Morton D. Hull Distinguished Service Professor of Computer Science and Sr. Advisor to the Provost for Computing and Data Science at the University of Chicago. At Chicago, he served as Chair of the Computer Science Department and co-founded the Data Science Institute. Previously, he was Thomas M. Siebel Professor of Computer Science at the University of California, Berkeley, and served a term as Chair of the Computer Science Division. He was Co-Director of the Algorithms, Machines

and People Laboratory (AMPLab) and is one of the original creators of Apache Spark, a leading open source platform for advanced data analytics and machine learning that was initially developed at the lab. He is a Member of the American Academy of Arts and Sciences and is a Fellow of the ACM and the American Association for the Advancement of Science. He received the 2022 ACM SIGMOD Systems Award with the team that developed Spark, and is a two-time recipient of the ACM SIGMOD "Test of Time" award. He holds a Ph.D. in Computer Sciences from the Univ. of Wisconsin (1993).

Natasha Noy is a scientist at Google Research where she works on making structured data accessible and useful. She leads the team building Dataset Search, a search engine for all the datasets on the Web. Prior to joining Google, she worked at Stanford Center for Biomedical Informatics Research where she made major contributions in the areas of ontology development and alignment, and collaborative ontology engineering. Dr. Noy is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI). She has served as the President of the Semantic Web Science Association from 2011 to 2017.

Haixun Wang is an IEEE fellow, editor in chief of IEEE Data Engineering Bulletin, and a VP of Engineering and Distinguished Scientist at Instacart. Before Instacart, he was a VP of Engineering and Distinguished Scientist at WeWork, a Director of Natural Language Processing at Amazon, and he led the NLP team working on Query and Document Understanding at Facebook. From 2013 to 2015, he worked on natural language processing with Google Research. From 2009 to 2013, he led research in semantic search, graph data processing systems, and distributed query processing at Microsoft Research Asia. He had been a research staff member at IBM T. J. Watson Research Center from 2000 to 2009. He received his Ph.D. in Computer Science from the University of California, Los Angeles, in 2000. He has published more than 150 research papers in referred international journals and conference proceedings. He served as PC Chair of conferences, such as SIGKDD'21, and is on the editorial board of journals, such as IEEE Transactions of Knowledge and Data Engineering (TKDE) and Journal of Computer Science and Technology (JCST). He won the best paper award in ICDE 2015, the 10-year best paper award in ICDM 2013, and the best paper award of ER 2009.