Data Management Opportunities for Foundation Models

Laurel Orr, Karan Goel, Christopher Ré
Stanford University
{ljorr1, kgoel, chrismre}@cs.stanford.edu

There is a paradigm shift in industrial machine learning pipelines where data is becoming one of the most important factors when building performant systems [5, 18, 17, 13, 9, 3]. Previously, ML pipelines followed a more "model-centric" paradigm, where engineers customized model architectures and hand curated features for training. These pipelines are being replaced by "foundation model" [4] ecosystems that follow a "data-centric" [14] viewpoint—commoditized architectures (e.g., Transformers [20] or MLPs) are trained *without manual labels* (i.e., with self-supervision) on massive corpora and adapted to hundreds of downstream tasks. In this new paradigm, the differentiating factor between models is the data they are "fed", not the architecture. Managing these foundation models is essentially the problem of managing their data lifecycle.

As foundation models are guided by their training data, one of their most important industrial benefits is that they often require dramatically less engineer effort to maintain (and can provide an improvement in quality compared to their hand-tuned predecessors). Before foundation model pipelines, teams of engineers spent hundreds of man-hours manually labeling data or curating features for each desired locale or domain. Now an engineer simply has to gather and manage unlabeled data for each locale. This reduction in engineer effort has contributed to their rapid rise in importance. Foundation models are already in production (e.g., Google Search [19] and Bootleg [11]), and a research institution recently formed a large interdisciplinary community to study them [1].

This paper seeks to highlight key data management challenges and opportunities with the foundation model data lifecycle. Our perspective is shaped by our experiences building and deploying a first-of-its-kind self-supervised model for entity disambiguation with Apple and a toolkit for end-to-end model monitoring with Salesforce. We believe the data management community is uniquely situated to address these challenges as many of them are twists on traditional data management problems spanning data integration, data exploration, and data programming.

12th Annual Conference on Innovative Data Systems Research (CIDR '22) January 9-12, 2022, Chaminade, USA.

Challenges and Opportunities.

We now discuss three challenges and opportunities for data management.

1. **Data Integration**: Recent works [11, 12] show how enhancing unstructured text data with structured metadata allows foundation models to have better factual recall and improved performance over rare items that do not occur frequently in training data but are common in deployed settings. At its core, this is a data integration problem between unstructured and structured data. Foundation models bring a new set of challenges to the data integration problem as they are trained on complex multi-modalities including code [2], images, videos, text, and robotic simulations [4]. Integration solutions must adapt to the heterogeneous nature of the data and must scale to handle petabytes of business and personal data used for training [10].

2. **Model Monitoring**: As foundation models are trained on typically unlabeled datasets with less engineer manipulation, there is an increased importance on monitoring model performance and detecting undesirable behavior. Recent works highlight how engineers benefit from the ability to easily filter relevant subpopulations and construct test examples for targeted performance metrics [8, 15, 16]. This is the data management challenge of data exploration over model predictions. Foundation model ecosystems make the data exploration challenge more complex as there are model artifacts for both the pretrained models (i.e., embeddings) and downstream models (i.e., model predictions). Further, model monitoring solutions need to tie performance results back to the underlying training data that caused the behavior (i.e., provenance for model predictions).

3. **Model Patching and Maintenance**: Foundation models are continuously evolving and need constant maintenance. When a model demonstrates undesirable behavior, an engineer needs to quickly update the model and correct for the error. Recent work by a subset of authors describe how models can be corrected all through the underlying training data [11, 7]. The authors used common data engineering methods, like weak supervision and augmentation, to "correct" a pretrained entity disambiguation model. The ability to engineer data for models is a fundamental data management challenge. As foundation models are adapted

to hundreds of downstream tasks, model patching solutions need to correct the underlying pretrained model rather than each downstream model individually. This presents an efficiency challenge as pretraining can take on the order of days to weeks [6]. Further, users will need automated suggestions for how to engineer the data given an undesirable behavior.

Call to Action.

We believe the data management community can have an outsized impact on deciding how these models are managed and maintained. We hope the community is excited by these ideas and can build off them.

Acknowledgments: We gratefully acknowledge the support of NIH under No. U54EB020405 (Mobilize), NSF under Nos. CCF1763315 (Beyond Sparsity), CCF1563078 (Volume to Velocity), and 1937301 (RTML); ONR under No. N000141712266 (Unifying Weak Supervision); ONR N00014-20-1-2480: Understanding and Applying Non-Euclidean Geometry in Machine Learning; N000142012275 (NEPTUNE); the Moore Foundation, NXP, Xilinx, LETI-CEA, Intel, IBM, Microsoft, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, the Okawa Foundation, American Family Insurance, Google Cloud, Salesforce, Total, the HAI Cloud Credits for Research program, the Stanford Data Science Initiative (SDSI), and members of the Stanford DAWN project: Facebook, Google, and VMWare. Laurel Orr is supported by the IC Postdoc Fellowship Program. The Mobilize Center is a Biomedical Technology Resource Center, funded by the NIH National Institute of Biomedical Imaging and Bioengineering through Grant P41EB027060. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of NIH, ONR, or the U.S. Government.

REFERENCES

[1] Introducing the center for research on foundation models (crfm). https://hai.stanford.edu/news/introducing-center-research-foundation-models-crfm, 2021.

[2] Openai codex. https://openai.com/blog/openai-codex/, 2021.

[3] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 291–300. IEEE, 2019.

[4] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[5] E. Breck, N. Polyzotis, S. Roy, S. Whang, and M. Zinkevich. Data validation for machine learning. In *MLSys*, 2019.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019.

[7] K. Goel, L. Orr, N. F. Rajani, J. Vig, and C. Ré. Goodwill hunting: Analyzing and repurposing off-the-shelf named entity linking systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 205–213, 2021.

[8] K. Goel, N. Rajani, J. Vig, S. Tan, J. Wu, S. Zheng, C. Xiong, M. Bansal, and C. Ré. Robustness gym: Unifying the nlp evaluation landscape. 2021.

[9] F. Hohman, K. Wongsuphasawat, M. B. Kery, and K. Patel. Understanding and visualizing data iteration in machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[10] B. Marr. Really big data at walmart: Real-time insights from their 40+ petabyte data cloud. https://www.forbes.com/sites/bernardmarr/2017/01/23/really-big-data-at-walmart-real-time-insights-from-their-40-petabyte-data-2017.

[11] L. Orr, M. Leszczynski, S. Arora, S. Wu, N. Guha, X. Ling, and C. Re. Bootleg: Chasing the tail with self-supervised named entity disambiguation. 2020.

[12] M. E. Peters, M. Neumann, R. L. Logan IV, R. Schwartz, V. Joshi, S. Singh, and N. A. Smith. Knowledge enhanced contextual word representations. 2019.

[13] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich. Data management challenges in production machine learning. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1723–1726, 2017.

[14] C. Ré. The road to software 2.0 or data-centric ai. https://hazyresearch.stanford.edu/data-centric-ai, 2021.

[15] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Association for Computational Linguistics (ACL)*, 2020.

[16] S. Sagadeeva and M. Boehm. Sliceline: Fast, linear-algebra-based slice finding for ml model debugging. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2290–2299, 2021.

[17] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[18] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger. Automating large-scale data quality verification. *Proceedings of the VLDB Endowment*, 11(12):1781–1794, 2018.

[19] B. Schwartz. Google: Bert now used on almost every english query. https://searchengineland.com/google-bert-used-on-almost-every-english-query-342193, 2020.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.