

Managing ML Pipelines: Feature Stores and the Coming Wave of Embedding Ecosystems

Laurel Orr, Atindriyo Sanyal, Xiao Ling, Karan Goel, and Megan Leszczynski

Stanford University, Uber AI, Apple

{lorr1,kgoel,mleszczy}@cs.stanford.edu,atin@uber.com,xiaoling@apple.com

Abstract

The industrial machine learning pipeline requires iterating on model features, training and deploying models, and monitoring deployed models at scale. Feature stores were developed to manage and standardize the engineer’s workflow in this end-to-end pipeline, focusing on traditional tabular feature data. In recent years, however, model development has shifted towards using self-supervised pretrained embeddings as model features. Managing these embeddings and the downstream systems that use them introduces new challenges with respect to managing embedding training data, measuring embedding quality, and monitoring downstream models that use embeddings. These challenges are largely unaddressed in standard feature stores. Our goal in this tutorial is to introduce the feature store system and discuss the challenges and current solutions to managing these new embedding-centric pipelines.

PVLDB Reference Format:

Laurel Orr, Atindriyo Sanyal, Xiao Ling, Karan Goel, and Megan Leszczynski. Managing ML Pipelines: Feature Stores and the Coming Wave of Embedding Ecosystems. PVLDB, 14(12): 3178-3181, 2021. doi:10.14778/3476311.3476402

1 Introduction

Building industrial machine learning pipelines is an iterative cycle of gathering and curating training data, training and deploying models, and monitoring the model once put into production. When errors or undesirable behavior are found, the cycle is repeated. Without tools to manage this process, production models become hard to maintain and difficult to reproduce. In 2017, a subset of the authors built the first industrial *feature store* [13]—a system designed to standardize and manage model features and workflows. The feature store both reduced engineer effort and improved model quality. Feature stores, however, have yet to adapt to a growing trend in model development: incorporating pretrained embeddings.

Pretrained embeddings are becoming standard inputs to modern machine learning pipelines. These embeddings are typically trained in a self-supervised fashion over massive data sets and encode knowledge about words, entities, graphs, and images. Once trained, they can provide lift in numerous downstream tasks like recommendation systems [21], information retrieval [15], and data integration [20]. A subset of the authors saw first-hand from their

experience building and deploying an industrial self-supervised entity disambiguation system that pretrained embeddings are shifting industrial pipelines towards “hands-free” models that require limited hand-curated data and model engineering [14, 19, 24]. There is therefore an increasing need for the next generation of feature store systems to help manage and monitor the embedding training data, pretrained embeddings, and downstream systems that consume the embeddings.

Goal The goal of this tutorial is to expose the interplay between data management and modern, self-supervised embedding ecosystems. We will first introduce feature store systems and the challenges they address. We will then introduce self-supervised pretrained embeddings and explain the new challenges associated with managing these embedding pipelines. We then explore how data management can help build, monitor, and maintain these self-supervised embedding ecosystems. Lastly, we will discuss future data management challenges and research directions.

Scope This tutorial focuses on modern feature stores and the challenges with supporting embeddings as first class citizens in feature stores. We highlight that managing these self-supervised embedding systems is a fundamental data management problem. This tutorial is intended for researchers with some familiarity with deep learning and pretrained embeddings who are interested in the interaction of data management and deep learning pipelines.

Outline This tutorial is split into three parts over 1.5 hours.¹

- (1) **Feature Stores.** We will give an overview of the modern machine learning pipeline and feature store systems. We will describe the core challenges these systems solve and give an overview of the technical contributions.
- (2) **Embedding Ecosystems.** We will introduce pretrained embeddings and discuss the new challenges faced by feature stores in treating embeddings as first class citizens. We then discuss recent solutions to some of these challenges.
- (3) **Future Directions and Challenges.** We will conclude with a discussion of the future directions and challenges.

2 ML Feature Stores

In 2017, a subset of the authors built the first industrial feature store [13]. Using this first-hand experience and the lessons learned, we introduce the modern ML pipeline and describe the challenges faced in maintaining and deploying models. We then introduce feature store systems and the technical innovations that help solve the aforementioned challenges. We focus the first part of this tutorial on traditional *tabular* feature data (i.e., not embeddings).

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

Proceedings of the VLDB Endowment, Vol. 14, No. 12 ISSN 2150-8097. doi:10.14778/3476311.3476402

¹This tutorial has not been presented in any venues prior and is most similar to recent SIGMOD and VLDB tutorials in data integration and cleaning with machine learning.

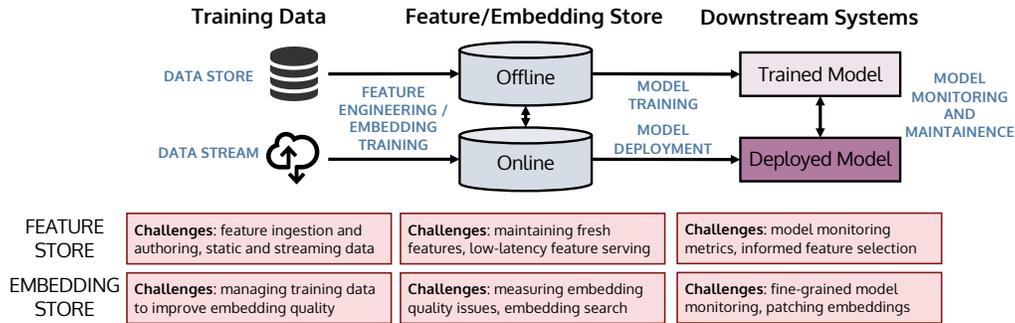


Figure 1: Modern machine learning pipeline with the challenges associated with managing a feature store (top row) and the additional challenges with managing an embedding ecosystem (bottom row).

2.1 Machine Learning Pipelines

In industrial machine learning (ML) pipelines, as shown in Figure 1, engineers need to quickly ingest training data for feature curation, train and deploy models, and monitor and maintain the model once deployed. We describe each step below.

- (1) **Training Data.** Data is scraped, mined, or retrieved from a variety of different sources. The data needs to be cleaned, checked, and featured for downstream models.
Challenges: Engineers author custom features that, if not shared and managed, can result in repeated work and lack of definitional consistency. Feature definitions can become stale if not kept up-to-date as data changes over time.
- (2) **Model Training and Deployment.** Using a set of features, engineers need to train and deploy models.
Challenges: As data changes over time and updates occur at different intervals, models can become stale if not given the most up-to-date features. Further, model reproducibility becomes a challenge as engineers try to keep up with changing data and model parameters.
- (3) **Model Maintenance and Monitoring.** Once deployed, models need to be monitored and maintained.
Challenges: Models can struggle in the face of distribution shift and out-of-domain inputs [26]. Further, once model errors are detected, engineers are often lacking guidance as to what features need to be corrected.

2.2 Feature Stores

Feature stores (FSs) arose to address these challenges by providing a centralized repository of reusable features across the ML pipeline and automating the management of this pipeline [1–3, 13]. Below, we dive into how feature stores address the three challenges above.

2.2.1 Training Data

Structured data can be in the form of raw tables as well as streams that users access when curating features. To facilitate sharing of features across an organization and maintaining features if they get updated, feature stores allow for *feature authoring and publishing* [4]. Users provide simple definitional metadata, e.g., the feature update cadence and a definition SQL query, and upload the definition to the FS. When the underlying data changes, the FS orchestrates the updates to the features based on the user-defined cadence.

For streaming features, users provide aggregation functions that are applied on the raw streaming features. The aggregated features are persisted to the online store and logged to the offline store.

2.2.2 Model Training and Deployment

Feature Storage Once features are curated, users need the ability to construct feature sets on the most recent data to train and deploy models. FSs support this workflow by partitioning features on date and providing APIs to allow for time based joins. Further, FSs must support feature quality metrics to support the detection and mitigation of feature errors. For example, FSs measure feature freshness, null counts, and mutual information across features.

Model Storage Once a model is trained, relevant parameters and artifacts need to be stored for provenance and reproducibility. Although model storage is not traditionally part of a FS, some FSs [3, 13] do support model management by integrating a separate model store [8, 28].

Online Feature Serving Once a model is deployed, features need to be continuously provided to deployed models even as the feature data is updated over time. To provide low latency feature serving, FSs are typically a dual datastore: one for offline training (e.g., SQL warehouse) and for online serving (e.g., in-memory DBMS).

2.2.3 Model Monitoring and Maintenance

FSs must additionally support model quality metrics (as well as feature quality metrics described above) [28]. For example, FSs support critical model metrics such as training-deployment data skew and near real-time outlier and input drift detection. These metrics allow users to be informed of potential ‘gremlins’ in the system. Once an error is discovered, engineers can use the FS metrics to detect the offending set of features and select a more optimal feature set for serving (or retraining).

3 Data Management and Embedding Ecosystems

Drawing on our first-hand experiences developing entity embeddings across numerous downstream products at a large technology company, we now introduce self-supervised pretrained embeddings and the ecosystems around them. We then discuss new challenges and solutions with managing these ecosystems. We believe treating embeddings as first class citizens is the next evolution of feature stores.

3.1 Embedding Ecosystem

We define the embedding ecosystem as the embedding training data, embeddings, and downstream systems that consume them. As shown in Figure 1, the embedding ecosystem pipeline is similar to that of the feature store. However, FSs are unable to support end-to-end embedding management. With embeddings, standard metrics and tools for managing tabular features are no longer adequate as embeddings are derived data. For example, embeddings are often compared by dot product similarity, and existing FS metrics such as null value count do not capture drifts or changes in embeddings with respect to this metric.

We now highlight the additional challenges associated with each step of the ML pipeline when incorporating embeddings (the same challenges from FSs still apply to embedding ecosystems).

- (1) **Training Data:** In an embedding ecosystem, the raw training data is used to pre-train the embeddings. As this data is self-supervised, it will not be hand-labeled or curated.

Challenges: Embeddings will encode any inherent biases that exist in the self-supervised training data, e.g., the embeddings do not well represent rare things [22, 27].

- (2) **Model Training and Deployment:** Once embeddings are trained, they need to be stored and served to downstream systems that use embeddings for training and deployment.

Challenges: As embeddings get retrained and updated, just like features, the downstream models can become stale and out-of-date. Users need to understand and monitor the embedding changes and search over possible embeddings and select the best ones for their task. Unlike features, standard tabular metrics are inadequate for embeddings.

- (3) **Model Maintenance and Monitoring:** Like with FSs, deployed models need to be monitored and maintained, especially with respect to the embedding inputs.

Challenges: Any inherent embedding quality issue will impact all downstream models using those embeddings. Users need to be able to understand and isolate downstream quality issues in the underlying embeddings. Once found, users need methods for correcting errors in downstream products.

3.1.1 Self-Supervised Training Data

Unlike feature curation data, which is tabular and pre-labeled, self-supervised training data is often unstructured and is not hand-curated. This lack of manual curation results in data that may not accurately represent the data seen upon deployment [5, 11, 16] and is often biased toward popular things [22]. Embeddings trained on this data can inherit these biases.

To improve embedding quality of rare things through training data management, recent work from [22] explored incorporating structured data into entity embedding pretraining through named entity disambiguation, the task of mapping from strings to things in a knowledge base. They showed that by adding structured data of the type of an entity and its knowledge graph relations, they could boost performance over rare entities by 40 F1 points. We believe merging structured and unstructured data is a promising management technique for improving quality in training data.

3.1.2 Embedding Management

Both traditional features and pretrained embeddings are served to

downstream models. The uniqueness of an embedding ecosystem is that users need to be able to understand the difference in embedding quality as embeddings are updated over time and need guidance over what embeddings to use for their task.

To measure quality, Wendlandt et al. [29] and Hellrich and Hahn [12] discuss analyzing word embeddings with respect to an embedding's nearest neighbors. The work of Leszczynski et al. [17] is uniquely looking at the quality of an embedding with respect to a downstream task. The authors define the metric of downstream instability, the number of predictions that change with different embeddings, to measure downstream embedding instability.

There is little available work on finding the right embedding to use, especially given compute or memory constraints. The work of May et al. [18] takes a first step by a variant of the eigenspace overlap score as a way of predicting downstream performance. However, their work is focused on measuring the performance of non-contextualized word embeddings.

3.1.3 Fine-Grained Monitoring and Patching

Downstream models in traditional FSs and embedding ecosystems need to be monitored and maintained. In an embedding ecosystem, however, the challenge is in giving users the tools to find meaningful subpopulations of errors and connecting the downstream errors to embedding quality issues. These errors then need to be corrected *through the underlying embedding*.

In terms of monitoring downstream models, recent works provide toolkits for measuring language model performance at a semantic, fine-grained level [10, 25]. Goel et al. [10] in particular focuses on allowing users to define custom sub-population functions to explore performance across different models.

Once an error is discovered, the challenge is in how to correct that error in the underlying embedding. By correcting the error in the embedding, all downstream systems using those embeddings will be patched, which maintains product consistency. The work in Orr et al. [22] gives a proof-of-concept that using data management techniques such as augmentation [7], weak supervision [23], and slice-based learning [6] can correct underperforming sub-populations of data [22].

4 Future Directions

We end with a discussion of future directions.

Embedding Enhanced Feature Stores We believe the next evolution of a feature store is one with native support for embeddings. While we discussed some challenges and potential solutions, this is just the beginning. Users need tools for searching and querying these embeddings as well as support for versioning, provenance, and downstream quality metrics. For example, if an embedding gets updated but a model that uses it does not, the dot product of the embedding with model parameters can lose meaning which leads to incorrect model predictions. Further, performing these operations at industrial scale will be non-trivial as the size of embeddings and their associated models are continuing to increase.

End-to-End Model Patching Through Data An open area of research is in automatically correcting the errors discovered in the downstream model error analysis through the underlying embedding. While prior work showed you can patch errors through methods

like data augmentation and slice finding [9, 22], there are remaining challenges in how to automate and manage this process. How can you predict if an augmentation strategy will have the desired result? If an embedding gets patched, what is the optimal way to propagate that patch downstream?

5 Biographical Sketches

Laurel Orr is a PostDoc in the Computer Science Department at Stanford University advised by Christopher Ré. She graduated from the University of Washington in the Database group and is a lead on the Bootleg project, a self-supervised system for entity disambiguation. Bootleg is in production at Apple and used by academic research groups. She was awarded the NSF GRFP as a graduate student and is a current IC Postdoc Fellow.

Atindriyo Sanyal is a technical lead on the Michelangelo team at Uber AI. He leads various feature engineering efforts across Uber. Prior to that, he worked at LinkedIn and Apple where he was a Senior Software Engineer on Siri and was the founding engineer behind SiriKit (Siri API). He did his Masters at UCLA, where he worked at the Networks Research Lab building routing algorithms for pedestrians with skin conditions on open source navigation systems. He's one of the winners of Microsoft's Imagine Cup, won many hackathons at university, an IEEE presidential Award nominee, and a Math Olympiad winner.

Xiao Ling is a Machine Learning engineer at Apple where his work spans from information extraction for knowledge base construction to open-domain question answering. He earned his PhD in Computer Science and Engineering from the University of Washington in 2015. He was an early engineer at Lattice Data Inc., which was acquired by Apple in 2017.

Megan Leszczynski is a PhD student in the Computer Science Department at Stanford University advised by Christopher Ré. She is one of the original developers of Bootleg, a self-supervised system for named entity disambiguation, which has since been deployed in industry. She has also given an invited lecture to the Stanford CS224N (NLP with Deep Learning) course led by Christopher Manning. Her research has been recognized with a NSF GRFP.

Karan Goel is a 3rd year CS PhD student at the Stanford AI Lab. He leads the Robustness Gym (RG) project, whose goal is to facilitate fine-grained evaluation and maintenance of ML models. RG is actively deployed at Salesforce with users in academia and industry. He wrote one of the first papers on "model patching", and his work has been recognized with a Siebel Scholarship (2018) and a Salesforce Research Grant (2020).

Acknowledgements We acknowledge the support of the IC Postdoctoral Research Fellowship Program and NSF Graduate Research Fellowship under No. DGE-1656518.

References

- [1] 2021. Feast: An Open Source Feature Store for Machine Learning. <https://feast.dev/>.
- [2] 2021. Feature Store for ML. <https://www.featurestore.org/>.
- [3] 2021. Hopsworks: The Enterprise Feature Store. <https://www.hopsworks.ai/>.
- [4] Wail Alkawaileet, Sattam Alsubaiee, Michael J Carey, Chen Li, Heri Ramampiaro, Phanwadee Sinthong, and Xikui Wang. 2018. End-to-end machine learning with Apache AsterixDB. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*. 1–10.
- [5] Michael S Bernstein, Jaime Teevan, Susan Dumais, Daniel Liebling, and Eric Horvitz. 2012. Direct answers for search queries in the long tail. In *SIGCHI*.
- [6] Vincent S Chen, Sen Wu, Zhenzhen Weng, Alexander Ratner, and Christopher Ré. 2019. Slice-based learning: A programming model for residual learning in

- critical data slices. *Advances in neural information processing systems* 32 (2019), 9392.
- [7] Nadiia Chepurko, Ryan Marcus, Emanuel Zraggen, Raul Castro Fernandez, Tim Kraska, and David Karger. 2020. ARDA: automatic relational data augmentation for machine learning. *arXiv preprint arXiv:2003.09758* (2020).
- [8] Gharib Gharibi, Vijay Walunj, Sirisha Rella, and Yuyung Lee. 2019. ModelKB: towards automated management of the modeling lifecycle in deep learning. In *2019 IEEE/ACM 7th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE)*. IEEE, 28–34.
- [9] Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. 2020. Model Patching: Closing the Subgroup Performance Gap with Data Augmentation. *arXiv preprint arXiv:2008.06775* (2020).
- [10] Karan Goel, Nazneen Rajani, Jesse Vig, Samson Tan, Jason Wu, Stephan Zheng, Caoming Xiong, Mohit Bansal, and Christopher Ré. 2021. Robustness Gym: Unifying the NLP Evaluation Landscape. *arXiv preprint arXiv:2101.04840* (2021).
- [11] Ben Gomes. 2017. Our latest quality improvements for Search. <https://blog.google/products/search/our-latest-quality-improvements-search/>.
- [12] Johannes Hellrich and Udo Hahn. 2016. Bad company—neighborhoods in neural embedding spaces considered harmful. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*. 2785–2796.
- [13] Jeremy Hermann and Mike Del Blaso. 2021. Meet Michelangelo: Uber's Machine Learning Platform. <https://eng.uber.com/michelangelo-machine-learning-platform/>.
- [14] Andrej Karpathy. 2020. Software 2.0. <https://medium.com/@karpathy/software-2-0-a64152b37c35>.
- [15] Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 39–48.
- [16] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiko Yasunaga, Richard Lanus Phillips, Sara Beery, et al. 2020. WILDS: A Benchmark of in-the-Wild Distribution Shifts. *arXiv preprint arXiv:2012.07421* (2020).
- [17] Megan Leszczynski, Avner May, Jian Zhang, Sen Wu, Christopher R Aberger, and Christopher Ré. 2020. Understanding the downstream instability of word embeddings. *arXiv preprint arXiv:2003.04983* (2020).
- [18] Avner May, Jian Zhang, Tri Dao, and Christopher Ré. 2019. On the downstream performance of compressed word embeddings. *Advances in neural information processing systems* 32 (2019), 11782.
- [19] Piero Molino, Yaroslav Dudin, and Sai Sumanth Miryala. 2019. Ludwig: a type-based declarative deep learning toolbox. *arXiv preprint arXiv:1909.07930* (2019).
- [20] Sidharth Mudgal, Han Li, Theodoros Rekatinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra. 2018. Deep learning for entity matching: A design space exploration. In *Proceedings of the 2018 International Conference on Management of Data*. 19–34.
- [21] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G Azzolini, et al. 2019. Deep learning recommendation model for personalization and recommendation systems. *arXiv preprint arXiv:1906.00091* (2019).
- [22] Laurel Orr, Megan Leszczynski, Simran Arora, Sen Wu, Neel Guha, Xiao Ling, and Christopher Re. 2021. Bootleg: Chasing the Tail with Self-Supervised Named Entity Disambiguation. *CIDR 2021, 11th Conference on Innovative Data Systems Research* (2021).
- [23] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, Vol. 11. NIH Public Access, 269.
- [24] Christopher Ré, Feng Niu, Pallavi Gudipati, and Charles Srisuwananukorn. 2020. Overton: A Data System for Monitoring and Improving Machine-Learned Products. In *CIDR 2020, 10th Conference on Innovative Data Systems Research, Amsterdam, The Netherlands, January 12-15, 2020, Online Proceedings*.
- [25] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118* (2020).
- [26] Sebastian Schelter, Felix Biessmann, Tim Januschowski, David Salinas, Stephan Seufert, and Gyuri Szarvas. 2018. On challenges in machine learning model management. (2018).
- [27] Timo Schick and Hinrich Schütze. 2020. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8766–8774.
- [28] Manasi Vartak, Harihar Subramanyam, Wei-En Lee, Srinidhi Viswanathan, Saadiyah Husnoo, Samuel Madden, and Matei Zaharia. 2016. ModelDB: a system for machine learning model management. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. 1–3.
- [29] Laura Wendlandt, Jonathan K Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. *arXiv preprint arXiv:1804.09692* (2018).