ML-Asset Management: Curation, Discovery, and Utilization

Mengying Wang
Case Western Reserve University
Cleveland, OH, US
mxw767@case.edu

Moming Duan
National University of Singapore
Singapore
moming@nus.edu.sg

Yicong Huang
University of California, Irvine
Irvine, CA, US
yicongh1@ics.uci.edu

Chen Li
University of California, Irvine
Irvine, CA, US
chenli@ics.uci.edu

Bingsheng He
National University of Singapore
Singapore
hebs@comp.nus.edu.sg

Yinghui Wu
Case Western Reserve University
Cleveland, OH, US
yxw1650@case.edu

ABSTRACT

Machine learning (ML) assets, such as models, datasets, and metadata—are central to modern ML workflows. Despite their explosive growth in practice, these assets are often underutilized due to fragmented documentation, siloed storage, inconsistent licensing, and lack of unified discovery mechanisms, making ML-asset management an urgent challenge. This tutorial offers a comprehensive overview of ML-asset management activities across its lifecycle, including curation, discovery, and utilization. We provide a categorization of ML assets, and major management issues, survey state-of-the-art techniques, and identify emerging opportunities at each stage. We further highlight system-level challenges related to scalability, lineage, and unified indexing. Through live demonstrations of systems, this tutorial equips both researchers and practitioners with actionable insights and practical tools for advancing ML-asset management in real-world and domain-specific settings.

PVLDB Reference Format:
Mengying Wang, Moming Duan, Yicong Huang, Chen Li, Bingsheng He, and Yinghui Wu. ML-Asset Management: Curation, Discovery, and Utilization. PVLDB, 18(12): 5493 - 5498, 2025.
doi:10.14778/3750601.3750701

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://ml-assets-management.github.io/.

1 INTRODUCTION

The realm of Machine Learning (ML) and Artificial Intelligence (AI) witnesses the creation of large amounts of ML models and relevant data resources. For example, data platforms such as HuggingFace [1] host over 1.5 million models, with 100, 000 new models added each month, occupying over 17 PB of storage [24]. These are valuable ML *assets*. Generally speaking, ML assets are high-value artifacts that may contribute to ML-driven data analysis workflows. Such ML assets include, but not limited to:

Proceedings of the VLDB Endowment, Vol. 18, No. 12 ISSN 2150-8097.
doi:10.14778/3750601.3750701

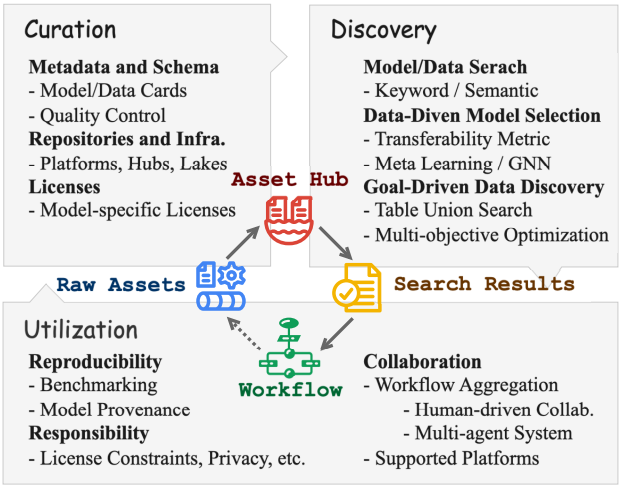

Figure 1: ML-Asset Management Lifecycle Overview

- *Datasets*: raw datasets, standardized documentation, annotated (training) data, validation data, test data, generated (benchmark) datasets, open data samples, feature vectors, etc;
- *Models*: pre-trained, fine-tuned, or foundation ML models; and relevant supporting code resources such as training pipelines, software libraries (e.g., AutoML components, LLM agents, statistical or physical models);
- *Metadata*: ontologies or data constraints/rules; (open source or proprietary) licenses, scripts and prompts (*e.g.,* for LLMs), provenance data, data sources (*e.g.,* contributors), hardware metadata, domain-specific in-lab/instrumental/experimental data, etc.

The lack of management of rich sets of ML assets leads to high maintenance costs, underutilized datasets and models, inefficiency in workflow development, and security and trustworthiness concerns. For example, over half of the models hosted in HuggingFace have no accompanying model card (documentation), and less than 8% are properly licensed [24]. Having this said, there still lack a standard characterization and through investigation of ML-asset management issues. A cornerstone step is to establish a systematic characterization of ML-asset management tasks and critical issues, and to provide a structured management infrastructure for modern ML-asset management. Data management community plays an essential role in contributing fundamental and advanced data management techniques to support such needs – by treating ML assets

Part 1: Motivation and Background (5 mins)
Part 2: ML-Asset Curation (22 mins)
- **Metadata and Schema**: Support Structured Understanding.
- **Repositories and Infrastructure**: Backbone of Discovery.
- **Licenses**: Enable Responsible Reuse.

Part 3: ML-Asset Search and Discovery (23 mins)
- **Model and Dataset Search.**
- **Data-driven Model Selection.**
- **Model-driven Data Discovery.**

Part 4: ML-Asset Utilization (25 mins)
- **Collaboration**: Workflow Aggregation and Automation.
- **Reproducibility**: Benchmarking and Model Provenance.
- **Responsibility**: Licensing and Ethical Asset Governance.

Part 5: System Challenges and Opportunities (15 mins)
- **Storage and Scalability.**
- **Versioning and Lineage.**
- **Indexing and Searching.**

Part 6: Demonstration (Split and merged into Part 2-4)

Figure 2: Tutorial Outline (90 minutes)

with the same rigor as data objects, we can enable storage and modeling, queries and indexing, version control, lineage tracking, and provenance, among other critical capabilities. **This tutorial** aims to provide an overview of major ML-asset management tasks, synergetic research efforts that provide enabling techniques for effective ML-asset management tasks, and a vision for future opportunities.

Target Audience. Our targeted audience comprises academic researchers, industry practitioners and stakeholders in both (1) data management, data science, machine learning, AI, and (2) multidisciplinary areas where ML-driven analysis plays critical roles.

Difference with Existing Tutorials. We structure our tutorial following a "life-cycle" analysis of ML assets, from the modeling and curation, resource access and discovery, to their utilization in downstream tasks. To ensure effective learning experience, we will incorporate a series of live, interactive demo to reveal the asset management techniques and their connections.

Learning Outcomes. Participants will gain a practical and conceptual understanding of ML-asset management across, including
(1) How to characterize and curate ML assets effectively;
(2) How to store, index, and retrieve ML assets at scale;
(3) How to manage provenance, versioning, and licensing; and
(4) How can ML-asset management support reproducibility, responsible, and collaborative AI?

2 TUTORIAL OUTLINE

We propose a 1.5-hour tutorial to review the existing progress, challenges, and opportunities of ML-asset management (outlined in Fig. 2). We next describe the detailed agenda for each part.

2.1 ML-Asset Curation

ML-asset curation is essential to harnessing the growing abundance of datasets and models. Effective curation ensures that assets are not merely stored but also well-described, reusable, and regulated.

***Metadata and Schema**: Support Structured Understanding.* Metadata describes the essential properties of datasets and models, including their origin, modality, training configuration, evaluation metrics, known limitations, etc [2, 37]. It also captures interactions between assets, such as model performance across different datasets [32]. High-quality metadata helps clarify where an asset works well and where it might not, thereby minimizing misuse.

Community efforts have introduced concepts like Data Cards [57] and Model Cards [44] to add structures to meta-information of ML assets. One step forward, recent work such as CRUX [72] provide knowledge graphs of ML assets by linking various types of "Cards" (structured data objects). Efforts on linking ML assets with data dependencies, license compatibility, and functional calls among ML assets with graph models opens the door for formal modeling, schema, and normalization, declarative manipulation, provenance, compatibility checking, among other issues that are of both theoretical and practical interests.

***Repositories and Infrastructure:** Backbone of Discovery.* Open platforms such as Hugging Face [1], Kaggle [28], TensorFlow Hub [65], and OpenML [50] have become community standards for hosting ML models and datasets. They offer structured metadata templates, tagging, version control, and integration with popular ML frameworks for easy access and reuse. More recently, the emergence of Data Markets [16] and Model Lakes [51] reflects a shift toward infrastructure that supports scalable discovery and reuse of ML assets. These repositories move beyond siloed storage, enabling asset-centric querying, composition, and integration at scale.

***Licenses**: Enable Responsible Reuse.* A unique aspect of ML assets is the emergence of model-specific licenses, which are designed to govern the use and distribution of model-related components such as weights, checkpoints, optimizer states and architectures through legal terms and agreements. Model-specific licenses introduce three key differences compared to traditional software licenses [59]: governance over remote access (e.g., Model-as-a-Service [36]), restrictions on responsible AI use [9] and conditions pertaining to model distillation and generated content. For example, Gemma License [21] includes web access in its definition of "Distribution" and states that the "transfer of patterns of the weights" constitutes "Model Derivatives" governed by this license. Additionally, Article 3.1 of Gemma License prohibits licensees from using the Gemma model or its "Model Derivatives" in violation of its Prohibited Use Policy. These tailored terms in model-specific licenses significantly broaden the scope of governance and increase curated objects, thereby complicating legal compliance in ML systems.

There are several *challenges and open opportunities* in ML-asset curation. First, **incomplete or low-quality metadata**, much of the published metadata is manually entered without validation or quality control [47], which harms metadata-based discovery. One promising approach is automating metadata generation [5, 61] and utilizing LLMs for enhancing semantics. Second, **schema inconsistency**: different platforms adopt different metadata schemas, making integration and federation difficult. A community-wide standard schema or ontology is necessary to facilitate widespread adoption [23]. Third, **license ambiguity and conflicts**, for example, license restrictions can propagate through model fine-tuning chains, and some may be mutually exclusive (e.g., GPL-3.0 and

Llama3.1 Community License [17]). As model dependencies grow deeper and involve more components, manual legal compliance analysis becomes increasingly impractical [13, 14]. A formal license curation framework that enables automated dependency reasoning and compliance checking across assets is an essential next step.

2.2 ML-Asset Search and Discovery

Effective search and discovery are critical for the reuse of ML assets and the acceleration of workflow construction [54].

Model and Dataset Search. Beginning with **keyword and tag-based filtering**, which is widely used on platforms such as Hugging Face [1] and Kaggle [28], these systems support faceted search over structured metadata, enabling users to refine results based on modality, task, or license through exact matches [7]. Recent progress has brought about **semantic and vector-based retrieval**, embedding models or datasets within a unified space for similarity-based searches [31], with vector databases used to index model and document embeddings for rapid similarity queries [69]. These methods offer entry points for ML-asset discovery, but they tackle various asset types independently and neglect valuable interactions.

Data-driven Model Selection. Given high-value datasets and tasks at hand, a crucial question is: *Which model should we use?* Brute-force evaluation is often infeasible due to the scale of modern model hubs [34, 63]. To address this, several works propose **transferability metrics**, which rank pre-trained models by estimating the label evidence on a target dataset, based on features extracted from the models [48, 79]. Another direction leverages **meta-learning** for model recommendation, where a recommender is trained to predict model performance using the metadata of ML-assets. This approach enables more precise and context-aware ranking [33]. **Graph learning-based** recommenders may further improve the quality of suggested models, by exploiting enriched metadata/features, better-informed suggestions and annotations, and more efficient cold-start strategies for new datasets [39, 70].

Model-driven Data Discovery. Recent research also advocates that data discovery for ML models could be "model-driven", with a goal to identify data over which a given ML model has high expected performance and small training/testing overhead. This requires finer-grained data manipulation that may integrate feature engineering and data integration [38]. Methods utilizing **table union search** to enhance data completeness and schema compatibility may generate tables by semantically merging multiple contextualized columnar data sources [15]. **Goal-oriented data discovery** tailors data selection to specific downstream tasks, guided by a target utility function [18]. **Multi-objective data discovery** incorporates multiple user-defined model performance evaluation criteria, to generate datasets that may optimize model performance across various performances [71]. Recent research also develop model-aware data augmentation for LLM pretraining and fine-tuning [29, 74, 78].

Effective ML-asset discovery benefits from guarantees on *robust search*, *high-quality metadata*, and *context-awareness*. Yet, several challenges remain alongside opportunities. First, **cold-start issue**: current methods depend heavily on underlying metadata, which might be limiting for many tasks. Beyond enriching metadata, embedding techniques to derive standardized representations directly from raw assets content offers a promising solution. Second, **discovery at scale**, searching through huge number of assets with complex queries becomes computationally expensive. Scalable infrastructure (distributed retrieval, caching of embeddings, vector databases, etc.) is needed. Third, **semantic understanding**: interactive discovery requires systems to understand both sides (model and data) at a semantic level, making a unified representation space that encapsulates the characteristics of various asset types crucial.

2.3 ML-Asset Utilization

The ultimate payoff lies in utilizing these well-organized ML assets and supporting systems to improve applications and practices in *reproducible*, *ethical*, and *collaborative* data science [3].

***Collaboration**: Workflow Aggregation and Automation.* Workflow aggregation involves creating modular ML workflows by selecting compatible assets from repositories. This modularity enables collaboration [40], allowing teams from different disciplines and backgrounds to collaboratively integrate and reuse components from various sources [10]. Platforms such as Davos [62] and Texera [75] have emerged to support these efforts. Modularized workflows can be modeled as directed acyclic graphs (DAGs), providing a structure foundation for aggregation approaches [64]. Beyond human-driven collaboration, there is a trend toward automating workflow construction using agents. Unlike traditional AutoML, which typically requires extensive testing [45], multi-agent approaches frame workflow assembly as a goal-conditioned planning problem. Leveraging language agents, such a framework may reason over asset metadata, infer task requirements, and iteratively assemble workflows [80].

***Reproducibility**: Benchmarking and Model Provenance.* ML-asset management streamlines benchmarking, allowing researchers to evaluate algorithms against standard datasets and baselines stored in asset repositories [41]. Employing established procedures by leveraging version-controlled and validated resources significantly improves the reproducibility of experiments [67]. A second critical aspect is tracking model provenance [46, 56], which aims to track a model's lineage data (training data, preprocessing, hyperparameters, source code, evaluation metrics and results). Model provenance allows others to replicate experiments, verify reported results, and gain insights in their reusability [60]. Data provenance such as "Why-provenance" can be adapted to generate post-hoc explanations for tracking ML model outputs, as observed in [8, 58].

***Responsibility**: Licensing and Ethical Asset Governance.* Responsible ML-asset utilization starts with clear licensing and agreements that govern how an asset may be used, adapted, or distributed. Licensing information often suffers from inconsistency during the reuse and republishing of licensed materials [77]. As an open standard for *AI Bills of Materials* (BOM), *Software Package Data Exchange 3.0* (SPDX 3.0) [4] enables the structured recording of ML assets and their associated licensing information throughout the development lifecycle, potentially supporting automated license-related analyses, such as detecting compatibility issues [30], inconsistencies [76, 77], and license proliferation [20]. Existing license compliant analysis tools such as FOSSology [27], Carneades [22], ModelGo Analyzer [14] and Black Duck [26] may extend to ML projects if AI BOM is avaliable. Unfortunately, SPDX 3.0 is not yet integrated into

mainstream ML tools, and model development disclosure remains unstandardized. Moreover, commonly used model file formats (e.g., Safetensors, GGUF, and OpenVINO IR) do not embed license metadata, leading to inconsistency of licensing information. ML-asset utilization remains subject to uncertain legal compliance risks.

Beyond licensing, responsible reuse also involves privacy and transparency issues which are expected to be captured through metadata documentation (e.g., model/data cards and provenance data). However, this remains challenging due to the variation in AI-related regulations across jurisdictions. Meanwhile, the use policies of the model vendors (enforceable under contract law) must also be complied with. Material breaches of either applicable laws or licenses/agreements may result in legal consequences.

2.4 System Challenges & Opportunities

Treating ML assets as first-class citizens indicates new types of ML-asset management systems. As ML assets grow rapidly in size, complexity, and volume, there is a need to revisit data management systems on how to best explore them in ML-asset management.

Storage and Scalability. ML assets, particularly large models and datasets, pose substantial storage challenges due to their rapidly increasing size and complexity. Techniques have emerged as scalable solutions, such as **compressed binary formats** like Safetensors provide safe, zero-copy tensor storage, enabling faster and secure model loading during deployment [6]. Earlier systems like ModelDB adopted a lightweight design by **storing only essential metadata** while keeping large binaries in external object storage[68]. Other efforts, such as Model Lake, leverage **distributed storage infrastructure**[19], etc. Ensuring efficiency, security, and consistency at scale remains an open research challenge.

Versioning and Lineage. Versioning and lineage tracking are crucial for reproducibility and auditability in data science. **Delta-based version control systems**, inspired by Git, allow efficient management of evolving datasets and models by capturing fine-grained changes along with detailed metadata [43]. **Provenance systems** like ProvDB represent ML workflows as graphs, enabling rich queries over asset lineage and dependencies [42]. Nonetheless, scalability remains a considerable challenge, and the potential for lineage reuse is an area that warrants further investigation [55].

Indexing and Searching. To efficiently search through a large volume of assets, powerful indexing mechanisms are essential. Supporting heterogeneous information, such as structured metadata, graph-based lineage, and semantic embeddings, poses challenges for hybrid query execution. Early attempts often relied on interfaces that surfaced all data types separately (e.g., via tabbed views) [19], but hybrid indexes offer a more unified and performant solution [53]. Additionally, it is important to maintain index freshness with minimal cost, while also considering privacy and security concerns. There are several potential directions: (1) Explore **vector database** systems [52] and optimization techniques in vector processing for large-scale ML-asset search. (2) Text-rich domain languages, datahubs, and application scenarios of ML assets continue to enrich their metadata and features, hence in turn providing opportunities of recent Large Langage Models (**LLMs**) and Retrieval Augmented Generation (**RAG**) methods in ML-asset recommendation.

3 DEMONSTRATION

We will walk through several ML-asset management tools and showcase how they may benefit ML-asset management tasks.

***CRUX**: ML-Asset Curation and Discovery.* CRUX [72] is a crowd-sourced platform for curating ML assets analysis for materials data science. It captures rich metadata and model–dataset interactions using domain-specific ontologies co-designed with materials scientists. CRUX supports model recommendation and data discovery [73]. We will demonstrate asset ingestion, metadata visualization, and search over asset knowledge graphs.

***ModelGo**: ML-Asset License Analyzer and License set.* We demonstrate the following: ModelGo Analyzer [12, 14], an ontology-based tool for automated license compliance analysis in ML projects. It evaluates licensing-related issues such as rights granting, term conflicts, and incompatibility between licenses. ModelGo Licenses [11]: a new Creative Commons-style model-specific license set designed for general model publication. It supports flexible licensing options to meet diverse model sharing needs.

***Texera**: Collaborative Workflow Composition.* Texera [66, 75] is an open-source platform designed to support collaborative data science and AI/ML. It offers a GUI-based workflow interface that enables analysts with diverse technical backgrounds to contribute effectively. Analysts can collaboratively edit workflows, interact with live executions [35], and jointly debug a workflow execution [25] in real time. Texera also supports reproducibility and determinstic replays [49] by preserving execution configurations and histories.

4 BIOGRAPHY

Mengying Wang is a Ph.D. candidate in Computer Science at Case Western Reserve University (CWRU), advised by Dr. Yinghui Wu. Her research interests include ML-asset management agentic workflow, knowledge graph discovery, and graph RAG. **Moming Duan** is a Research Fellow at Institute of Data Science, National University of Singapore (NUS). His research interests include AI Governance and Licensing and Federated Learning. **Yicong Huang** is a Ph.D. candidate in the Department of Computer Science, University of California, Irvine (UCI), advised by Dr. Chen Li. His research interests lie in data management and ML systems. He is a main contributor of the Texera project. **Chen Li** is a Professor in the Department of Computer Science at UCI. His research interests are in data management, including data-intensive computing, databases, query processing, ML systems, and data science. His current focus is building open-source systems for big data and AI/ML. **Bingsheng He** is a Professor at School of Computing, NUS. His current research interests include database and machine learning systems, and high performance computing. **Yinghui Wu** is an Associate Professor in the Department of Computer and Data Sciences, CWRU. His area is in data management, data science, and graph data analysis.

ACKNOWLEDGMENTS

Wang and Wu are supported by NSF under OAC-2104007. Huang and Li are supported by NSF under award III-2107150 and NIH under award 1U01AG076791-01. Duan and He are supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative.

REFERENCES

[1] Hugging Face AI. 2025. Hugging Face – The AI Community Building the Future. Retrieved Apr 13, 2025 from https://huggingface.co/

[2] Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Luca Foschini, Joan Giner-Miguelez, Pieter Gijsbers, Sujata Goswami, Nitisha Jain, Michalis Karamousadakis, Michael Kuchnik, et al. 2024. Croissant: A metadata format for ml-ready datasets. *Advances in Neural Information Processing Systems* 37 (2024), 82133–82148.

[3] Becky Arnold, Louise Bowler, Sarah Gibson, Patricia Herterich, Rosie Higman, Anna Krystalli, Alexander Morley, Martin O'Reilly, Kirstie Whitaker, et al. 2019. The turing way: a handbook for reproducible data science. *Zenodo* (2019).

[4] Karen Bennet, Gopi Krishnan Rajbahadur, Arthit Suriyawongkul, and Kate Stewart. 2024. *Implementing AI Bill of Materials (AI BOM) with SPDX 3.0: A Comprehensive Guide to Creating AI and Dataset Bill of Materials.* Technical Report. The Linux Foundation. https://doi.org/10.70828/RNED4427 Accessed: January 2025.

[5] Kris Cardinaels, Michael Meire, and Erik Duval. 2005. Automating metadata generation: the simple indexing interface. In *Proceedings of the 14th international conference on World Wide Web*. 548–556.

[6] Beatrice Casey, Kaia Damian, Andrew Cotaj, and Joanna Santos. 2025. An Empirical Study of Safetensors' Usage Trends and Developers' Perceptions. *arXiv preprint arXiv:2501.02170* (2025).

[7] Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, and Paul Groth. 2020. Dataset search: a survey. *The VLDB Journal* 29, 1 (2020), 251–272.

[8] Tingyang Chen, Dazhuo Qiu, Yinghui Wu, Arijit Khan, Xiangyu Ke, and Yunjun Gao. 2024. View-based explanations for graph neural networks. *PACMMOD* 2, 1 (2024), 1–27.

[9] Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. 2022. Behavioral use licensing for responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 778–788. https://doi.org/10.1145/3531146.3533143

[10] Behrouz Derakhshan, Alireza Rezaei Mahdiraji, Ziawasch Abedjan, Tilmann Rabl, and Volker Markl. 2020. Optimizing machine learning workloads in collaborative environments. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 1701–1716.

[11] Moming Duan. 2025. ModelGo Licenses - A Standard Way for Model Publishing. Retrieved Apr 13, 2025 from https://www.modelgo.li/

[12] Moming Duan, Mingzhe Du, Rui Zhao, Mengying Wang, Yinghui Wu, Nigel Shadbolt, and Bingsheng He. 2025. Position: Current Model Licensing Practices are Dragging Us into a Quagmire of Legal Noncompliance. In *Forty-second International Conference on Machine Learning Position Paper Track*.

[13] Moming Duan, Qinbin Li, and Bingsheng He. 2024. ModelGo: A practical tool for machine learning license analysis. In *Proceedings of the ACM Web Conference 2024 (WWW)*. 1158–1169. https://doi.org/10.1145/3589334.3645520

[14] Moming Duan, Rui Zhao, Linshan Jiang, Nigel Shadbolt, and Bingsheng He. 2024. " They've Stolen My GPL-Licensed Model!": Toward Standardized and Transparent Model Licensing. *arXiv preprint arXiv:2412.11483* (2024).

[15] Grace Fan, Jin Wang, Yuliang Li, Dan Zhang, and Renée J Miller. 2023. Semantics-Aware Dataset Discovery from Data Lakes with Contextualized Column-Based Representation Learning. *Proceedings of the VLDB Endowment* 16, 7 (2023).

[16] Raul Castro Fernandez, Pranav Subramaniam, and Michael J Franklin. 2020. Data market platforms: trading data assets to solve data problems. *Proceedings of the VLDB Endowment* 13, 12 (2020), 1933–1947.

[17] Free Software Fundation. 2025. Various Licenses and Comments about Them. Retrieved Apr 13, 2025 from https://www.gnu.org/licenses/license-list.en.html#Llama

[18] Sainyam Galhotra, Yue Gong, and Raul Castro Fernandez. 2023. Metam: Goal-oriented data discovery. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, 2780–2793.

[19] Moncef Garouani, Franck Ravat, and Nathalie Valles-Parlangeau. 2024. Model Lake: A New Alternative for Machine Learning Models Management and Governance. In *International Conference on Web Information Systems Engineering*. Springer, 133–144.

[20] Robert W Gomulkiewicz. 2009. Open Source License Proliferation: Helpful Diversity or Hopeless Confusion? *Washington University Journal of Law & Policy* 30, 1 (2009).

[21] Google. 2025. Gemma Terms of Use. Retrieved Apr 13, 2025 from https://ai.google.dev/gemma/terms

[22] Thomas F Gordon. 2011. Analyzing open source license compatibility issues with Carneades. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law (ICAIL)*. 51–55. https://doi.org/10.1145/2018358.2018364

[23] Ramanathan V Guha, Dan Brickley, and Steve Macbeth. 2016. Schema. org: evolution of structured data on the web. *Commun. ACM* 59, 2 (2016), 44–51.

[24] Eliahu Horwitz, Nitzan Kurer, Jonathan Kahana, Liel Amar, and Yedid Hoshen. 2025. Charting and Navigating Hugging Face's Model Atlas. *arXiv preprint arXiv:2503.10633* (2025).

[25] Yicong Huang, Zuozhi Wang, and Chen Li. 2023. Udon: Efficient Debugging of User-Defined Functions in Big Data Systems with Line-by-Line Control. *Proc. ACM Manag. Data* 1, 4 (2023), 225:1–225:26. https://doi.org/10.1145/3626712

[26] Black Duck Software Inc. 2025. Black Duck Software Composition Analysis. Retrieved Apr 13, 2025 from https://www.blackduck.com/software-composition-analysis-tools/black-duck-sca.html

[27] Michael C Jaeger, Oliver Fendt, Robert Gobeille, Maximilian Huber, Johannes Najjar, Kate Stewart, Steffen Weber, and Andreas Wurl. 2017. The FOSSology project: 10 years of license scanning. *International Free and Open Source Software Law Review* 9 (2017), 9.

[28] Kaggle. 2025. Kaggle: Your Home for Data Science. Retrieved Apr 13, 2025 from https://www.kaggle.com/

[29] Feiyang Kang, Hoang Anh Just, Yifan Sun, Himanshu Jahagirdar, Yuanzhi Zhang, Rongxing Du, Anit Kumar Sahu, and Ruoxi Jia. 2024. Get more for less: Principled Data Selection for Warming Up Fine-Tuning in LLMs. In *The Twelfth International Conference on Learning Representations*.

[30] Georgia M Kapitsaki, Frederik Kramer, and Nikolaos D Tselikas. 2017. Automating the license compatibility process in open source software with SPDX. *Journal of Systems and Software (JSS)* 131 (2017), 386–401. https://doi.org/10.1016/j.jss.2016.06.064

[31] Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. 2023. Similarity of neural network models: A survey of functional and representational measures. *Comput. Surveys* (2023).

[32] Annmary Justine Koomthanam, Aalap Tripathy, Sergey Serebryakov, Gyanaranjan Nayak, Martin Foltin, and Suparna Bhattacharya. 2024. Common metadata framework: Integrated framework for trustworthy artificial intelligence pipelines. *IEEE Internet Computing* 28, 3 (2024), 37–44.

[33] Miloš Kotlar, Marija Punt, Zaharije Radivojević, Miloš Cvetanović, and Veljko Milutinović. 2021. Novel meta-features for automated machine learning model selection in anomaly detection. *IEEE Access* 9 (2021), 89675–89687.

[34] Arun Kumar, Robert McCann, Jeffrey Naughton, and Jignesh M Patel. 2016. Model selection management systems: The next frontier of advanced analytics. *ACM SIGMOD Record* 44, 4 (2016), 17–22.

[35] Avinash Kumar, Zuozhi Wang, Shengquan Ni, and Chen Li. 2020. Amber: A Debuggable Dataflow System Based on the Actor Model. *Proc. VLDB Endow.* 13, 5 (2020), 740–753. https://doi.org/10.14778/3377369.3377381

[36] Emanuele La Malfa, Aleksandar Petrov, Simon Frieder, Christoph Weinhuber, Ryan Burnell, Raza Nazar, Anthony Cohn, Nigel Shadbolt, and Michael Wooldridge. 2024. Language-Models-as-a-Service: Overview of a new paradigm and its challenges. *Journal of Artificial Intelligence Research* 80 (2024), 1497–1523.

[37] Ziyu Li, Henk Kant, Rihan Hai, Asterios Katsifodimos, Marco Brambilla, and Alessandro Bozzon. 2023. Metadata representations for queryable repositories of machine learning models. *IEEE Access* 11 (2023), 125616–125630.

[38] Ziyu Li, Wenbo Sun, Danning Zhan, Yan Kang, Lydia Chen, Alessandro Bozzon, and Rihan Hai. 2024. Amalur: The convergence of data integration and machine learning. *IEEE Transactions on Knowledge and Data Engineering* 36, 12 (2024), 7353–7367.

[39] Ziyu Li, Hilco Van Der Wilk, Danning Zhan, Megha Khosla, Alessandro Bozzon, and Rihan Hai. 2024. Model Selection with Model Zoo via Graph Learning. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 1296–1309.

[40] Xiaozhen Liu, Zuozhi Wang, Shengquan Ni, Sadeem Alsudais, Yicong Huang, Avinash Kumar, and Chen Li. 2022. Demonstration of collaborative and interactive workflow-based data analytics in texera. *Proceedings of the VLDB Endowment* 15, 12 (2022), 3738–3741.

[41] Rachel Longjohn, Markelle Kelly, Sameer Singh, and Padhraic Smyth. 2024. Benchmark data repositories for better benchmarking. *Advances in Neural Information Processing Systems* 37 (2024), 86435–86457.

[42] Hui Miao, Amit Chavan, and Amol Deshpande. 2017. Provdb: Lifecycle management of collaborative analysis workflows. In *Proceedings of the 2nd Workshop on Human-in-the-Loop Data Analytics*. 1–6.

[43] Hui Miao, Ang Li, Larry S Davis, and Amol Deshpande. 2017. Modelhub: Deep learning lifecycle management. In *2017 IEEE 33rd International Conference on data engineering (ICDE)*. IEEE, 1393–1394.

[44] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[45] Supun Nakandala, Yuhao Zhang, and Arun Kumar. 2020. Cerebro: A data system for optimized deep learning model selection. *Proceedings of the VLDB Endowment* 13, 12 (2020), 2159–2173.

[46] Mohammad Hossein Namaki, Avrilia Floratou, Fotis Psallidas, Subru Krishnan, Ashvin Agrawal, Yinghui Wu, Yiwen Zhu, and Markus Weimer. 2020. Vamsa: Automated provenance tracking in data science scripts. In *KDD*.

[47] Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. 2016. Automated quality assessment of metadata across open data portals. *Journal of Data and Information Quality (JDIQ)* 8, 1 (2016), 1–29.

[48] Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. 2020. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*. PMLR, 7294–7305.

[49] Shengquan Ni, Yicong Huang, Zuozhi Wang, and Chen Li. 2024. IcedTea: Efficient and Responsive Time-Travel Debugging in Dataflow Systems. *Proceedings of the VLDB Endowment* 18, 3 (2024), 902–914.

[50] OpenML. 2025. OpenML: A worldwide machine learning lab. Retrieved Apr 13, 2025 from https://openml.org/

[51] Koyena Pal, David Bau, and Renée J Miller. 2025. Model Lakes. In *Proceedings of the 28th International Conference on Extending Database Technology (EDBT)*.

[52] James Jie Pan, Jianguo Wang, and Guoliang Li. 2024. Survey of vector database management systems. *The VLDB Journal* 33, 5 (2024), 1591–1615.

[53] Liana Patel, Peter Kraft, Carlos Guestrin, and Matei Zaharia. 2024. Acorn: Performant and predicate-agnostic search over vector embeddings and structured data. *Proceedings of the ACM on Management of Data* 2, 3 (2024), 1–27.

[54] Jian Pei, Raul Castro Fernandez, and Xiaohui Yu. 2023. Data and ai model markets: Opportunities for data and model sharing, discovery, and integration. *Proceedings of the VLDB Endowment* 16, 12 (2023), 3872–3873.

[55] Arnab Phani, Benjamin Rath, and Matthias Boehm. 2021. Lima: Fine-grained lineage tracing and reuse in machine learning systems. In *Proceedings of the 2021 International Conference on Management of Data*. 1426–1439.

[56] Fotis Psallidas, Megan Eileen Leszczynski, Mohammad Hossein Namaki, Avrilia Floratou, Ashvin Agrawal, Konstantinos Karanasos, Subru Krishnan, Pavle Subotic, Markus Weimer, Yinghui Wu, et al. 2023. Demonstration of Geyser: Provenance Extraction and Applications over Data Science Scripts. In *SIGMOD*.

[57] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1776–1826.

[58] Dazhuo Qiu, Mengying Wang, Arijit Khan, and Yinghui Wu. 2024. Generating robust counterfactual witnesses for graph neural networks. In *ICDE*.

[59] Lawrence Rosen. 2005. *Open Source Licensing: Software Freedom and Intellectual Property Law*. Prentice Hall Professional Technical Reference, New Jersey.

[60] Lukas Rupprecht, James C Davis, Constantine Arnold, Yaniv Gur, and Deepavali Bhagwat. 2020. Improving reproducibility of data science pipelines through transparent provenance capture. *Proceedings of the VLDB Endowment* 13, 12 (2020), 3354–3368.

[61] Sebastian Schelter, Joos-Hendrik Boese, Johannes Kirschnick, Thoralf Klein, and Stephan Seufert. 2017. Automatically tracking metadata and provenance of machine learning experiments. (2017).

[62] Zeyuan Shang, Emanuel Zgraggen, Benedetto Buratti, Philipp Eichmann, Navid Karimeddiny, Charlie Meyer, Wesley Runnels, and Tim Kraska. 2021. Davos: a system for interactive data-driven decision making. *Proceedings of the VLDB Endowment* 14, 12 (2021), 2893–2905.

[63] Evan R Sparks, Ameet Talwalkar, Daniel Haas, Michael J Franklin, Michael I Jordan, and Tim Kraska. 2015. Automating model search for large scale machine learning. In *Proceedings of the Sixth ACM Symposium on Cloud Computing*. 368–380.

[64] Evan R Sparks, Shivaram Venkataraman, Tomer Kaftan, Michael J Franklin, and Benjamin Recht. 2017. Keystoneml: Optimizing pipelines for large-scale advanced analytics. In *2017 IEEE 33rd international conference on data engineering (ICDE)*. IEEE, 535–546.

[65] TensorFlow. 2025. TensorFlow Hub. Retrieved Apr 13, 2025 from https://www.tensorflow.org/hub

[66] Texera – Collaborative Data Science and AI/ML Using Workflows 2025. Texera Website, https://texera.io.

[67] Jeyan Thiyagalingam, Mallikarjun Shankar, Geoffrey Fox, and Tony Hey. 2022. Scientific machine learning benchmarks. *Nature Reviews Physics* 4, 6 (2022), 413–420.

[68] 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.

[69] Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, et al. 2021. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data*. 2614–2627.

[70] Mengying Wang, Sheng Guan, Hanchao Ma, Yiyang Bian, Haolai Che, Abhishek Daundkar, Alp Sehirlioglu, and Yinghui Wu. 2023. Selecting top-k data science models by example dataset. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2686–2695.

[71] Mengying Wang, Hanchao Ma, Yiyang Bian, Yangxin Fan, and Yinghui Wu. 2025. Generating Skyline Datasets for Data Science Models. In *EDBT*.

[72] Mengying Wang, Hanchao Ma, Abhishek Daundkar, Sheng Guan, Yiyang Bian, Alpi Sehirlioglu, and Yinghui Wu. 2022. Crux: Crowdsourced materials science resource and workflow exploration. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 5014–5018.

[73] Mengying Wang, Hanchao Ma, Sheng Guan, Yiyang Bian, Haolai Che, Abhishek Daundkar, Alp Sehirlioglu, and Yinghui Wu. 2024. ModsNet: Performance-Aware Top-k Model Search Using Exemplar Datasets. *Proceedings of the VLDB Endowment* 17, 12 (2024), 4457–4460.

[74] Pengkun Wang, Zhe Zhao, HaiBin Wen, Fanfu Wang, Binwu Wang, Qingfu Zhang, and Yang Wang. 2024. Llm-autoda: Large language model-driven automatic data augmentation for long-tailed problems. *Advances in Neural Information Processing Systems* 37 (2024), 64915–64941.

[75] Zuozhi Wang, Yicong Huang, Shengquan Ni, Avinash Kumar, Sadeem Alsudais, Xiaozhen Liu, Xinyuan Lin, Yunyan Ding, and Chen Li. 2024. Texera: A System for Collaborative and Interactive Data Analytics Using Workflows. *Proceedings of the VLDB Endowment* 17, 11 (2024), 3580–3588.

[76] Yuhao Wu, Yuki Manabe, Tetsuya Kanda, Daniel M German, and Katsuro Inoue. 2015. A method to detect license inconsistencies in large-scale open source projects. In *Proceedings of IEEE/ACM 12th Working Conference on Mining Software Repositories (MSR)*. IEEE, 324–333. https://doi.org/10.1109/MSR.2015.37

[77] Yuhao Wu, Yuki Manabe, Tetsuya Kanda, Daniel M German, and Katsuro Inoue. 2017. Analysis of license inconsistency in large collections of open source projects. *Empirical Software Engineering (ESE)* 22 (2017), 1194–1222. https://doi.org/10.1007/s10664-016-9487-8

[78] Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. 2023. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems* 36 (2023), 34201–34227.

[79] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. 2021. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*. PMLR, 12133–12143.

[80] Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. 2025. Aflow: Automating agentic workflow generation. In *The Thirteenth International Conference on Learning Representations*.