Homepage

ML-Asset Management: Curation, Discovery, and Utilization

VLDB 2025 - Tutorial






Mengying Wang　Moming Duan　Yicong Huang　Chen Li　Bingsheng He　Yinghui Wu

CWRU　NUS National University of Singapore　UCI

VLDB 2025 WELCOME TO LONDON

Tutorial Roadmap

Homepage

Motivation and Background (00:00 - 00:05)

2 ML-Asset Curation (00:05 - 00:30)

Demo: ModelGo

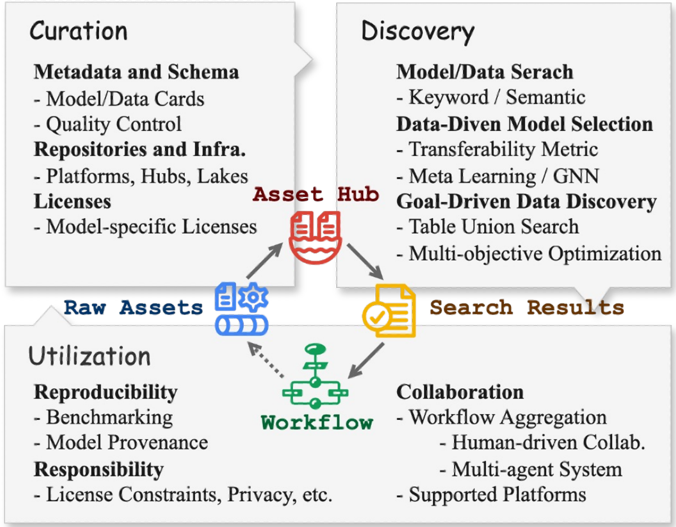
3 ML-Asset Search and Discovery (00:30 - 00:50)

Demo: CRUX

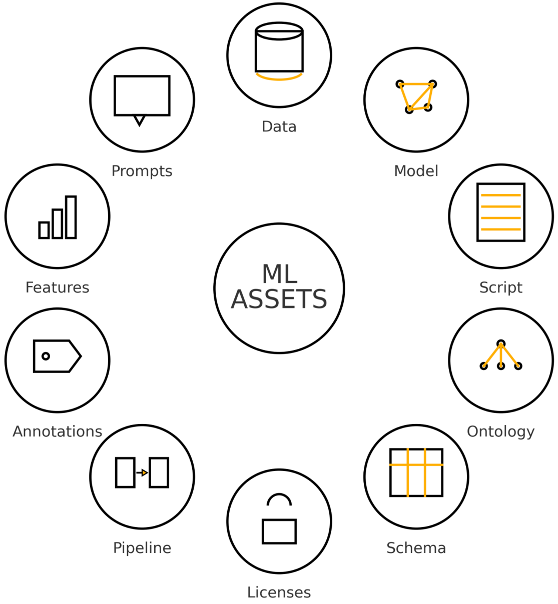
4 ML-Asset Utilization (00:50 - 01:15)

Demo: Texera

5 System Challenges and Opportunities (01:15 - 01:30)



What are “ML Assets”?



ML Assets are high-value, reusable artifacts generated and utilized throughout ML-driven workflows.

ML Assets: Explosive Growth v.s. Underutilized

Numbers from Hugging Face…

Total Number of Models

2M+

New Models Added Monthly

100k

Storage Occupied by ML Assets

17PB

>50%

Models without Model Cards

<8%

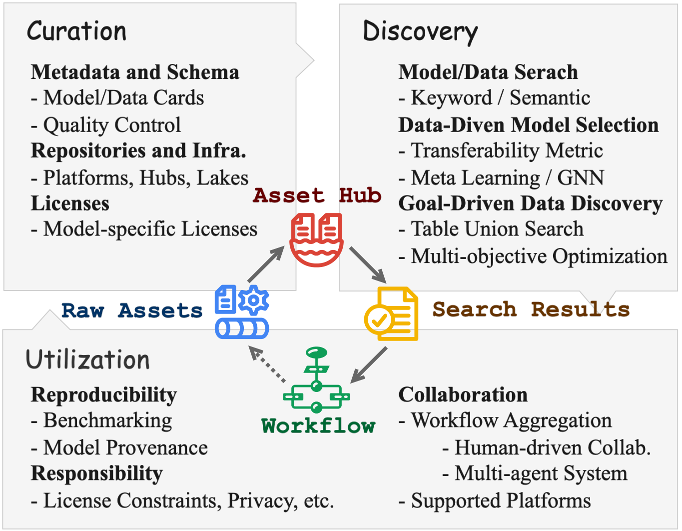
Properly licensed models

Charting and Navigating Hugging Face's Model Atlas.
Eliahu Horwitz, Nitzan Kurer, Jonathan Kahana, Liel Amar, and Yedid Hoshen. 2025.

ML Asset Management Lifecycle

Give assets an ID

Let assets be found



Use assets correctly

Tutorial Roadmap

Homepage

1 Motivation and Background (00:00 - 00:05)

ML-Asset Curation (00:05 - 00:30)

Demo: ModelGo

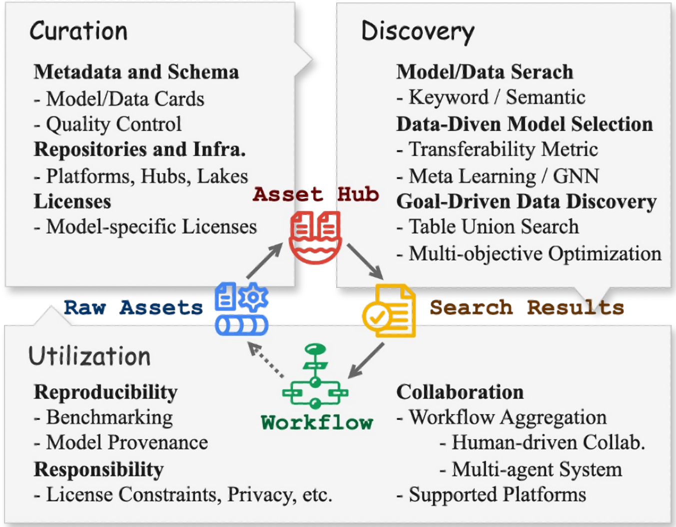
3 ML-Asset Search and Discovery (00:30 - 00:50)

Demo: CRUX

4 ML-Asset Utilization (00:50 - 01:15)

Demo: Texera

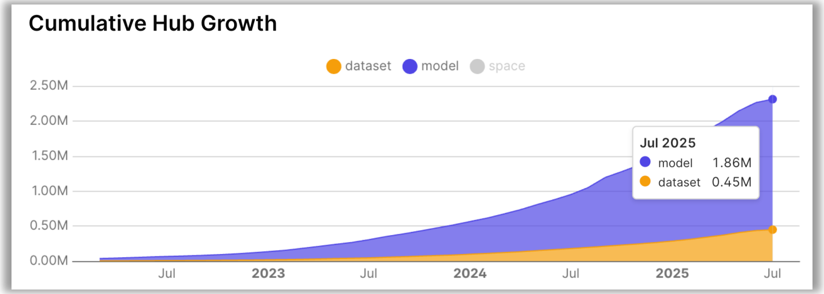
5 System Challenges and Opportunities (01:15 - 01:30)



ML-Asset Curation

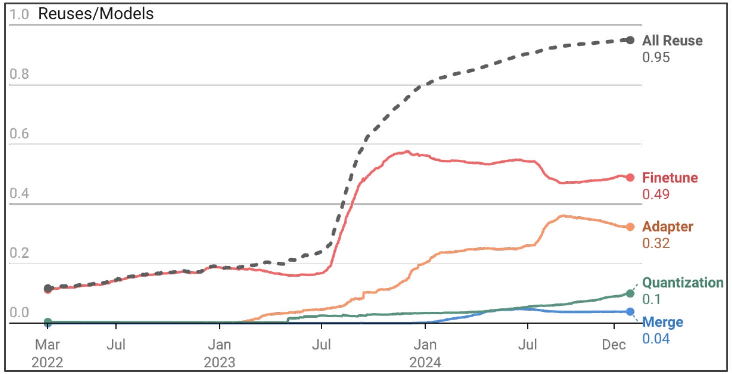
Background #1

Explosive Growth of ML Assets Driven by the Open Source Movement



Background #2

Reusing ML-Asset Is Increasingly Prevalent in ML Development



- There are over 2 million models and 449K datasets on Hugging Face.
- Pretraining is costly, but model reused is cheap (e.g., LoRA, only 0.1% – 1% parameters tuned).
- The model collaboration chains can be extended, nested, and merged.

Hub Stats Dataset, *https://huggingface.co/spaces/cfahlgren1/hub-stats*
Moming Duan, Mingzhe Du, Rui Zhao, Mengying Wang, Yinghui Wu, Nigel Shadbolt, and Bingsheng He. ***Position: Current Model Licensing Practices are Dragging Us into a Quagmire of Legal Noncompliance*** *(ICML'25 Oral)*

Moming – NUS
Assets Curation

ML-Asset Curation

Background #3

Model Platform and Dataset Platform

Platforms	# Model	# Dataset	User Contribution	Model Cards / Metadata	License Curation	Versioning / Dependency
Hugging Face	2,000 K	489 K	✅ High	✅ Support	⚠️ Self-reported	✅ Support
Kaggle	3.2 K	528 K	✅ High	✅ Support	⚠️ Self-reported	✅ Support
TensorFlow Hub	56	1.3 K	Moderate	⚠️ Basic	✅ Apache-2.0	⚠️ Limited
Pytorch Hub	75	NA	GitHub-based	⚠️ Basic	⚠️ Third-party	⚠️ Limited
OpenML	NA (Flow)	24.1 K	✅ High	✅ Support	⚠️ Self-reported	✅ Support
OpenVINO	248	NA	Verified	⚠️ Basic	✅ Curated	❌

- Community-powered platforms (e.g., Hugging face, Kaggle) host the largest number of ML assets, but rely on self-reported information for asset descriptions **-> Need for manual curation**

Moming – NUS
Assets Curation

ML-Asset Curation

Background #4

Licensing of Assets in ML Projects

ML Project	Task	Data License	Software License	Model License	Dataset	Risk Resource
Stable Diffusion v1-5	Text to Image	CC-BY-4.0	CreativeML-OpenRAIL-M	CreativeML-OpenRAIL-M	LAION-5B	Common Crawl
BLOOM	Text Generation	*Mixture*	*Unknown*	BigScience-BLOOM-RAIL-1.0	*Crowdsourced*	Common Crawl, Wikipedia, etc.
OrangeMixs	Text to Image	*Mixture*	*Unknown*	CreativeML-OpenRAIL-M	*Crowdsourced*	Danbooru
ControlNet	Text to Image	*Unknown*	Apache-2.0	OpenRAIL	*Unknown*	n/a
Openjourney	Text to Image	CC-BY-NC-4.0	*Unknown*	CreativeML-OpenRAIL-M	Midjourney Gen	Midjourney Gen
ChatGLM-6B	Text Generation	*Mixture*	Apache-2.0	*Custom*	the Pile, Wudao, *Crowdsourced*	PubMed, Wikipedia, arXiv, GitHub, etc.
Llama2	Text Generation	*Unknown*	Llama2 Community License	Llama2 Community License	*Unknown*	n/a
StarCoder	Text Generation	*Mixture*	Apache-2.0	BigCode-OpenRAIL-M	The Stack	none
Falcon-40B	Text Generation	ODC-By	Apache-2.0	Apache-2.0	RefinedWeb	Wikipedia, Reddit, StackOverflow, etc.
Waifu Diffusion	Text to Image	*Mixture*	*Unknown*	CreativeML-OpenRAIL-M	*Unknown*	n/a
Dolly-v2-12B	Text Generation	CC-BY-SA-3.0&4.0	MIT	MIT	databricks-dolly -15k, the Pile	PubMed, Wikipedia, arXiv, GitHub, etc.
Dreamlike Photoreal	Text to Image	*Unknown*	*Unknown*	*Modified* CreativeML-OpenRAIL-M	*Unknown*	n/a
Counterfeit	Text to Image	*Unknow*	*Unknown*	CreativeML-OpenRAIL-M	*Unknown*	n/a
GPT-2	Text Generation	*Mixture*	*Modified* MIT	*Modified* MIT	*Crowdsourced*	WordPress, GitHub, wikiHow, IMDb, etc.
GPT-J-6B	Text Generation	*Mixture*	Apache-2.0	Apache-2.0	the Pile	PubMed, Wikipedia, arXiv, GitHub, etc.
LLaMA-7B	Text Generation	*Mixture*	*Custom*	*Custom*	*Crowdsourced*	GitHub, arXiv, etc.
BERT	Fill Mask	*Mixture*	Apache-2.0	Apache-2.0	Book Corpus, Wikipedia (en)	Wikipedia (en)
Whisper	ASR	*Unknown*	MIT	MIT	*Unknown*	n/a
MPT	Text Generation	*Mixture*	Apache-2.0	Apache-2.0	*Crowdsourced*	Common Crawl, Wikipedia, etc.
Mistral-7B	Text Generation	*Unknow*	Apache-2.0	Apache-2.0	*Unknow*	n/a

License of Data **License of Software** **License of Model**

HuggingFace Models. *https://huggingface.co/models*

Model Licenses:
Open Responsible AI (RAIL), Llama2/3/… Community, Gemma, ModelGo, "Custom"

Software Licenses:
Apache-2.0, MIT, GPL-3.0, AFL-3.0, BSD-3-Clause …

Data Licenses:
Creative Commons, ODC-By, ODBL, DbCL-1.0, …

- A machine learning project may involve all three types of licenses.

ML-Asset Curation

What are OSS/Model/Data Licenses & Where to Find Them

A license is a legal agreement that specifies how others can use, modify, and distribute a work.

(Under IP law and contract law)

License Categories

According to the level of freedom and restrictions they impose on users.

Public Domain
Dedicated to the public, free of copyright, use and share w/o conditions. (CC0, Unlicense, ODC-By)

Permissive
Open source, grant broad rights, primarily requiring attribution and disclaiming warranties. (MIT, Apache-2.0)

Copyleft
Derivative works must be distributed under the same license, ensuring continued “freedom.” (GPL, CC-BY-SA)

Proprietary
Most restrictive, w/o gaining ownership or the right, for commercial software or data, significant limitations, often revocable, often include end-user license agreements. (Llama2, Stability AI, Gemma, NVIDIA EULA)

Moming – NUS
Assets Curation

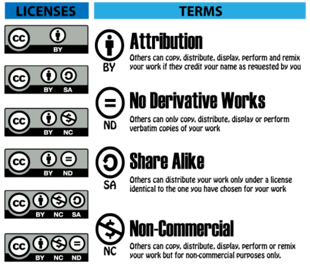
ML-Asset Curation

What are OSS/Model/Data Licenses & Where to Find Them

A license is a legal agreement that specifies how others can use, modify, and distribute a work.

Frequently Used Licenses

open source initiative Approved License®



RESPONSIBLE AI
LICENSES

Gemma

FFM Llama 2

OSI-Approved Licenses
Common on GitHub (e.g., Apache-2.0, MIT, GPL-3.0), these Open Source Software (OSS) licenses are widely supported by the community.

Creative Commons Licenses
Widely adopted for web content (articles, music, video) and datasets, allowing flexible control over use and distribution of original works and derivative works.

Open Responsible AI Licenses
Widely used for AI model sharing on Hugging Face (e.g., Stable Diffusion v1, StarCoder). Include restrictions for responsible model use, incompatible with Free Software (GPLs).

Proprietary Model License Agreement
Widely used for industrial AI models, restricts commercial use and others, revocable, enforced via contract consent.

ML-Asset Curation

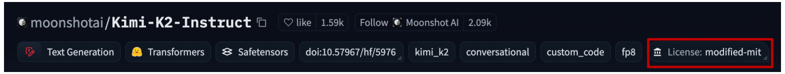
What are OSS/Model/Data Licenses & Where to Find Them

A license is a legal agreement that specifies how others can use, modify, and distribute a work.

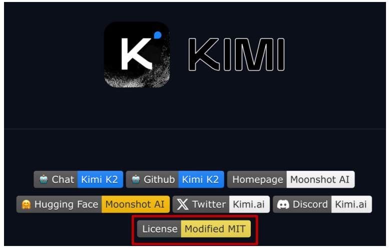
You may find the license info in:

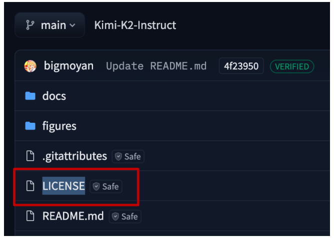
- Path (/LICENSE, /LICENSE.txt)
- Model/Data Cards (README.md, e.g., license: mit)
- Publications or Official Websites
- SPDX header (e.g., # SPDX-License-Identifier: GPL-3.0-or-later)

HF 👉 License Filter

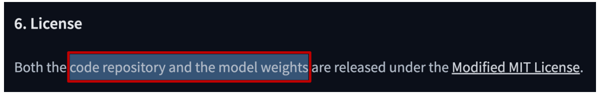


👈 HF License Tag & File





⚠️ **Always check the README**
Models and code might have separate licenses (e.g., chatGLM-6B).



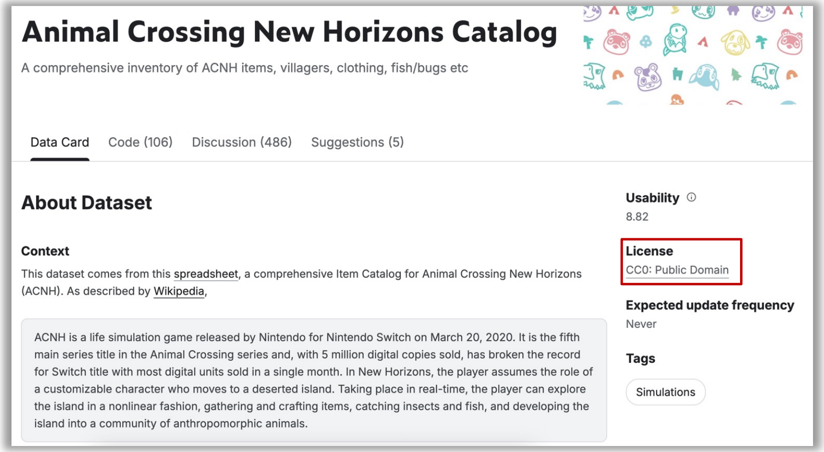
👈 Model README

Moming – NUS
Assets Curation

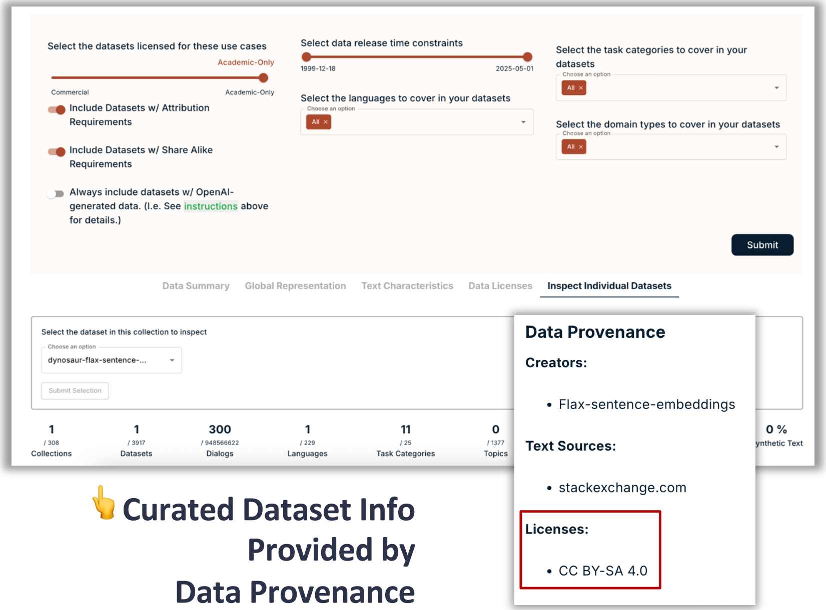
ML-Asset Curation

What are OSS/Model/Data Licenses & Where to Find Them

A license is a legal agreement that specifies how others can use, modify, and distribute a work.



👆 Kaggle License Tag



👆 Curated Dataset Info Provided by Data Provenance initiative

Kaggle Datasets, https://www.kaggle.com/datasets
Data Provenance Explorer, https://www.dataprovenance.org/data-provenance-explorer
Longpre, Shayne, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff et al. ***A large-scale audit of dataset licensing and attribution in AI****. Nature Machine Intelligence 2024.*

ML-Asset Curation

How to Read These License

First Important Thing: It Is a License or a License Agreement?

License

- A license is a **grant of permission**—a legal right to do something that would otherwise be restricted (e.g., copy, use, or distribute a software or model).
- It can be **unilateral**, meaning it doesn't require the recipient's explicit agreement to be enforceable.

License Agreement

- A license agreement is a **contractual document** that defines the terms and conditions under which a license is granted.
- It often requires both parties to **agree**, making it a mutual agreement.

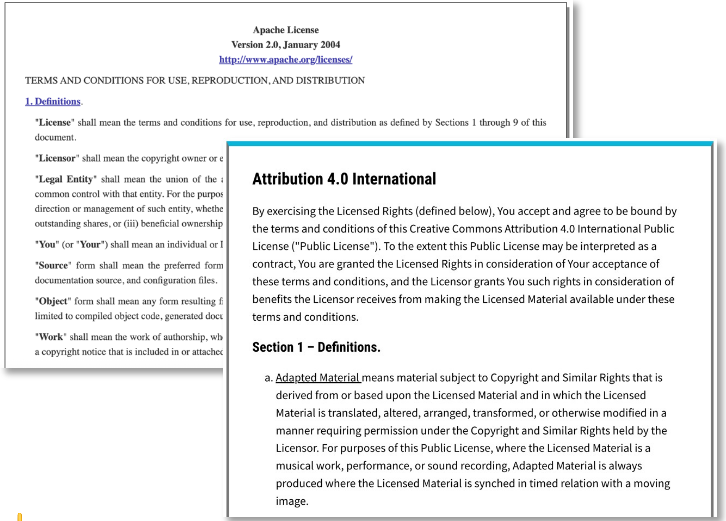
Aspects	License	License Agreement
Legal form	Permission (can be unilateral)	Contract (mutual agreement)
Typical context	Open source, public use	Commercial, proprietary, restricted use
Requires acceptance?	Not always (e.g., OSS)	Yes, typically signed or clicked to agree
Example	MIT, Apache-2.0, CCs, OpenRAILs	Llama3.*, Gemma, Stable Diffusion3.5

Moming – NUS
Assets Curation

ML-Asset Curation

How to Read These License

First Important Thing: It Is a License or a License Agreement?



Apache License
Version 2.0, January 2004
http://www.apache.org/licenses/

TERMS AND CONDITIONS FOR USE, REPRODUCTION, AND DISTRIBUTION

1. Definitions.

"License" shall mean the terms and conditions for use, reproduction, and distribution as defined by Sections 1 through 9 of this document.

"Licensor" shall mean the copyright owner or e

"Legal Entity" shall mean the union of the a common control with that entity. For the purpos direction or management of such entity, whethe outstanding shares, or (iii) beneficial ownership

"You" (or "Your") shall mean an individual or l

"Source" form shall mean the preferred form documentation source, and configuration files.

"Object" form shall mean any form resulting f limited to compiled object code, generated docu

"Work" shall mean the work of authorship, wh a copyright notice that is included in or attache

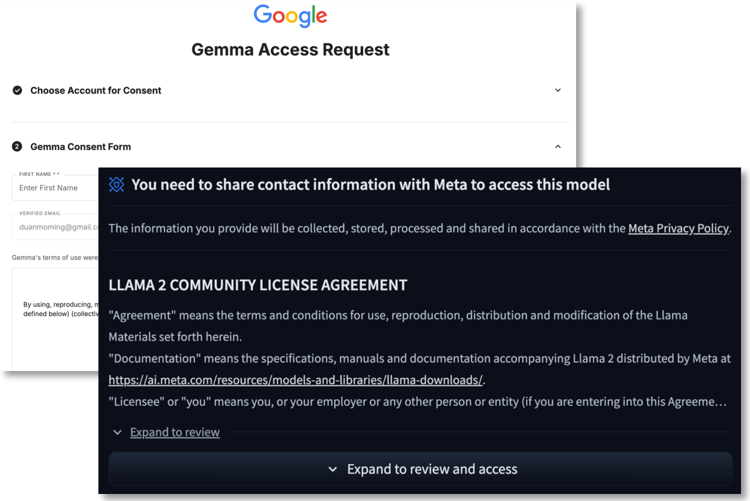
Attribution 4.0 International

By exercising the Licensed Rights (defined below), You accept and agree to be bound by the terms and conditions of this Creative Commons Attribution 4.0 International Public License ("Public License"). To the extent this Public License may be interpreted as a contract, You are granted the Licensed Rights in consideration of Your acceptance of these terms and conditions, and the Licensor grants You such rights in consideration of benefits the Licensor receives from making the Licensed Material available under these terms and conditions.

Section 1 – Definitions.

a. Adapted Material means material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor. For purposes of this Public License, where the Licensed Material is a musical work, performance, or sound recording, Adapted Material is always produced where the Licensed Material is synched in timed relation with a moving image.

👆 **Licenses**
Grant Rights Under IP Law
No explicit consent needed



Google
Gemma Access Request

Choose Account for Consent

Gemma Consent Form

FIRST NAME
Enter First Name

VERIFIED EMAIL
duanmoming@gmail.c

Gemma's terms of use were

By using, reproducing, m
defined below) (collectiv

You need to share contact information with Meta to access this model

The information you provide will be collected, stored, processed and shared in accordance with the Meta Privacy Policy.

LLAMA 2 COMMUNITY LICENSE AGREEMENT

"Agreement" means the terms and conditions for use, reproduction, distribution and modification of the Llama Materials set forth herein.

"Documentation" means the specifications, manuals and documentation accompanying Llama 2 distributed by Meta at https://ai.meta.com/resources/models-and-libraries/llama-downloads/.

"Licensee" or "you" means you, or your employer or any other person or entity (if you are entering into this Agreeme…

Expand to review

Expand to review and access

👆 **License Agreements**
Bound users by contract law
Require explicit user consent

Moming – NUS
Assets Curation

ML-Asset Curation

Why License Curation Matters?

2024 Ethics Reviewers Guidelines

The role of ethics review is to assess NeurIPS submissions for risks in at least one of the following areas:

- Research involving human subjects
- Data privacy, copyright, and consent
- Data quality and representativeness
- Safety and security
- Discrimination, bias, and fairness
- Deception and harassment
- Environmental Impact
- Human rights (including surveillance)

NeurIPS calls for ethics reviewers to audit submission materials, including license issues. 👇

Data-related concerns:

The points listed below apply to all datasets used for submissions, both for publicly available data and internal datasets.

- *Privacy:* Datasets should minimize the exposure of any personally identifiable information, unless informed consent from those individuals is provided to do so.
- *Consent:* Any paper that chooses to create a dataset with real data of real people should ask for the explicit consent of participants, or explain why they were unable to do so.
- *Deprecated datasets:* Authors should take care to confirm with dataset creators that a dataset is still available for use. Datasets taken down by the original author (ie. deemed obsolete, or otherwise discontinued), should no longer be used, unless it is for the purposes of audit or critical assessment. For some indication of known depreciated datasets, please refer to the NeurIPS list of deprecated datasets.
- *Copyright and Fair Use:* While the norms of fair use and copyright in machine learning research are still evolving, authors must respect the terms of datasets that have defined licenses (e.g. CC 4.0, MIT, etc).
- *Representative evaluation practice:* When collecting new datasets or making decisions about which datasets to use, authors should assess and communicate the degree to which their datasets are representative of their intended population. Claims of diverse or universal representation should be substantiated by concrete evidence or examples.

Impact Mitigation Measures

We propose some reflection and actions taken to mitigate potential harmful consequences from the research project.

- *Data and model documentation:* Researchers should communicate the details of the dataset or the model as part of their submissions via structured templates.
- *Data and model licenses:* If releasing data or models, authors should also provide licenses for them. These should include the intended use and limitations of these artifacts, in order to prevent misuse or inappropriate use.
- *Secure and privacy-preserving data storage & distribution:* Authors should leverage privacy protocols, encryption and anonymization to reduce the risk of data leakage or theft. Stronger measures should be employed for more sensitive data (e.g., biometric or medical data).

***NeurIPS Code of Ethics**, https://neurips.cc/public/EthicsGuidelines*
***This model IS NOT Apache 2.0 as you derived it from Gemma, which is proprietary model**, https://huggingface.co/FreedomIntelligence/Apollo2-9B/discussions/2*

JLouisBiz Apr 10

According to your model card, you derived this fine tuned model from google/gemma-2-9b which is proprietary model.

And you are trying to "re-license" it, but do you have written permission from Google for re-licensing to Apache 2.0? I don't think so.

This way you are putting users at risks...

Re-licensing software, especially proprietary models like Google's Gemma, without permission is a serious issue. Here's why:

1. **Legal Violation**: Unauthorized re-licensing can be a violation of copyright laws, leading to potential legal consequences.
2. **Breach of Contract**: If you have an agreement with Google, re-licensing the model without their permission might constitute a breach of contract.
3. **Trust and Credibility**: It can damage your reputation and credibility, as users may lose trust in your ability to manage intellectual property rights responsibly.
4. **User Risks**: As pointed out, users can be put at risk. They might rely on software that doesn't have clear ownership or licensing, which can lead to unexpected changes, discontinuation, or legal liability.

If you're looking to use a model like Gemma, it's crucial to seek permission from the original owners, like Google. Alternatively, you might explore other the truly free software models that are freely available and have clear licensing agreements, such as those under Apache 2.0. which you know which are those!

So my question is, Why? Why? Why?

Do you understand that re-licensing is not possible?

Do you understand that people fetching it will think it is Apache 2.0 but are put in legal danger?

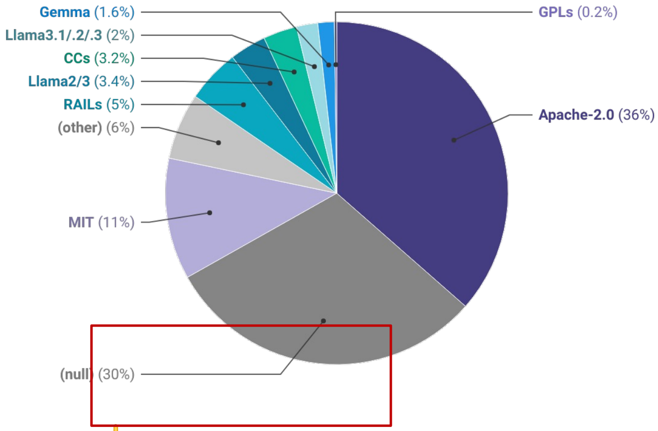
👆 Apollo2, derived from Gemma but relicensed as Apache-2.0, raises user concerns about legal compliance.

Moming – NUS
Assets Curation

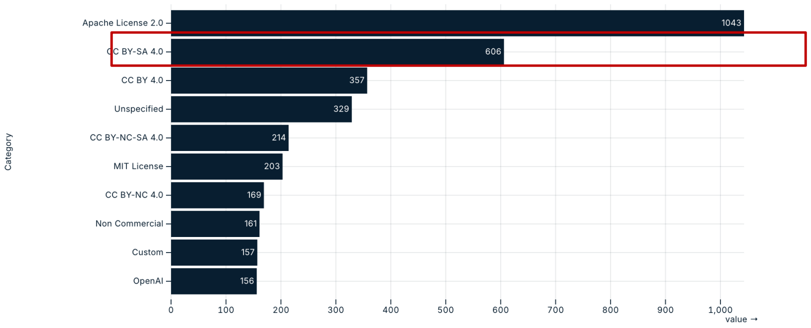
ML-Asset Curation

Ensure Compliance in ML-Assets

#1: Choose an Appropriate License



👆 ~1/3 of models on Hugging Face don't declare a license.



👆 Most popular dataset license: Apache-2.0 (an OSS license)

Moming Duan, Mingzhe Du, Rui Zhao, Mengying Wang, Yinghui Wu, Nigel Shadbolt, and Bingsheng He. ***Position: Current Model Licensing Practices are Dragging Us into a Quagmire of Legal Noncompliance*** *(ICML'25 Oral)*
Data Provenance Explorer - Dataset Explorer, *https://www.dataprovenance.org/data-provenance-explorer/dataset-explorer*

ML-Asset Curation

Ensure Compliance in ML-Assets

#1: Choose an Appropriate License

If you're publishing an original works
- Code -> OSS licenses, Dataset -> Data licenses, Models -> Model or OSS licenses.

If you're publishing derivatives based on open works (e.g., under OS, free content licenses)
- If the original work is under a Copyleft license -> Use the **same** license (e.g., (A/L)GPL, CC-BY(-NC)-SA)
- If the original work is under an Open Domain license -> Free to **relicense** (e.g., CC0, Unlicense, ODC-By)

If your work contains parts of proprietary works (Separable)
- License only your own contribution. **Do not use GPLs.**

If you're publishing derivatives of proprietary works (e.g., under license agreements, ToU)
- Just use the same license and version. **Do not relicense it!** Do not claim copyright ownership!

Moming – NUS
Assets Curation

ML-Asset Curation

Ensure Compliance in ML-Assets

#2: Comply with Restrictions and Obligations

If the original license is an Open Domain license
- No restrictions or obligations. You can claim copyright on your own contributions.

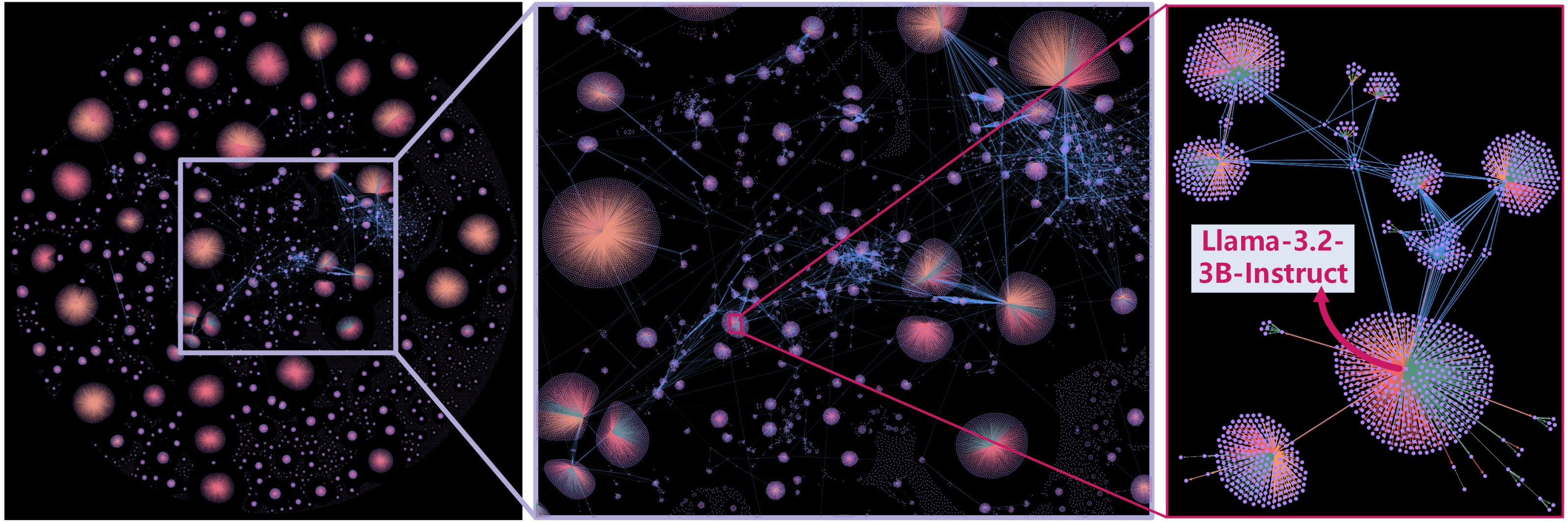
Else If the original license is an OSS, free software, or free content license
- Better to retain the original license files, headers, and notices.
- Indicate that your work is a modification based on the original work under that license.
- You can claim copyright on your own contributions.

Else If the original license is a proprietary license or Open Responsible AI (OpenRAIL) license
- Do not publish proprietary content.
- Provide the official link to indicate where it can be accessed.
- Always retain original agreements, attribution, copyright, trademark notices.
- Comply with all use policies and restrictions.
- Avoid commercial use unless explicitly permitted.

ML-Asset Curation

Ensure Compliance in ML-Assets

#3: Compliance Analysis for ML Supply Chain (Most Challenging)
Identify all dependencies in the ML project and analyze their compliance.



👆 *Visualization of Model Dependencies on* 🤗
Finetune (50%), Adapter (33%), Quantization (10%), and Merge (4%)

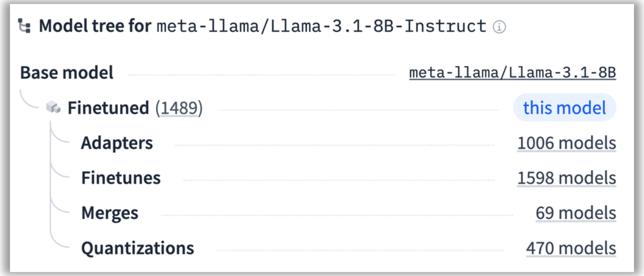
Moming Duan, Mingzhe Du, Rui Zhao, Mengying Wang, Yinghui Wu, Nigel Shadbolt, and Bingsheng He. ***Position: Current Model Licensing Practices are Dragging Us into a Quagmire of Legal Noncompliance*** *(ICML '25 Oral)*

ML-Asset Curation

Ensure Compliance in ML-Assets

#3: Compliance Analysis for ML Supply Chain (Most Challenging)

Identify all dependencies in the ML project and analyze their compliance.



👆 **Only four types of dependencies are labeled on** 🤗

👈

Model Atlas: A Project for Model Tree Heritage Recovery

⚠️ Lack effective tools to fully recover ML supply chain.

Eliahu Horwitz, Nitzan Kurer, Jonathan Kahana, Liel Amar, and Yedid Hoshen. ***We Should Chart an Atlas of All the World's Models****, Arxiv 2025*
Hugging Face Spaces Eliahu/Model-Atlas*,* *https://huggingface.co/spaces/Eliahu/Model-Atlas*

ML-Asset Curation

Ensure Compliance in ML-Assets

#3: Compliance Analysis for ML Supply Chain (Most Challenging)
Identify all dependencies in the ML project and analyze their compliance.

1. Compositional Analysis

Identify which parts of the original work are nested or embedded in the new work (consider recursive effects).

2. Definition Analysis

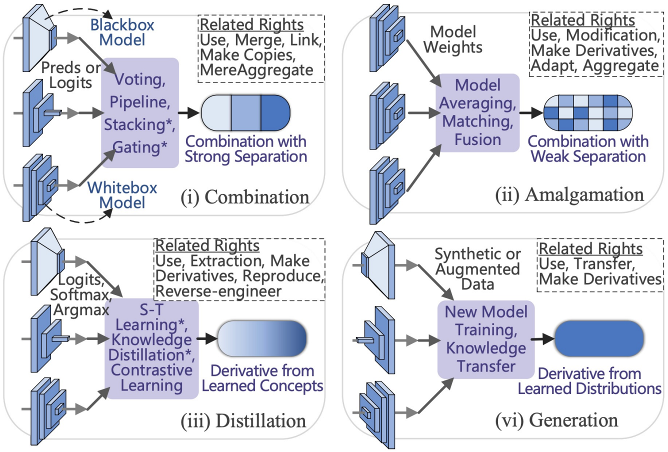
The new work is definited as "what" (Derivative? Independent?) according to the original work's license.

3. Rights Granting Analysis

Verify if you have rights (e.g., copy, share, adapt) to reuse (e.g., finetune, MoE, distill) the asset, and if these rights are revocable.

4. Conditions and Restrictions

Identify conditions and restrictions that remain when publishing or using your new work (e.g., attribution, source disclosure).



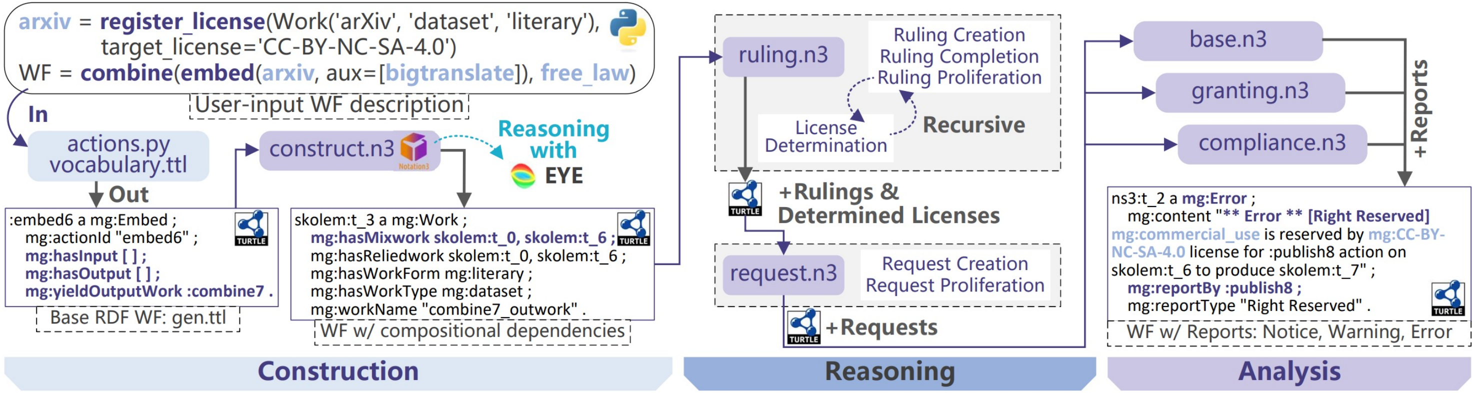
👆 **Different ML reuse methods require different rights and create different compositional dependencies.**

Moming – NUS
Assets Curation

ML-Asset Curation

Ensure Compliance in ML-Assets

#3: Compliance Analysis for ML Supply Chain (Most Challenging)
Identify all dependencies in the ML project and analyze their compliance.

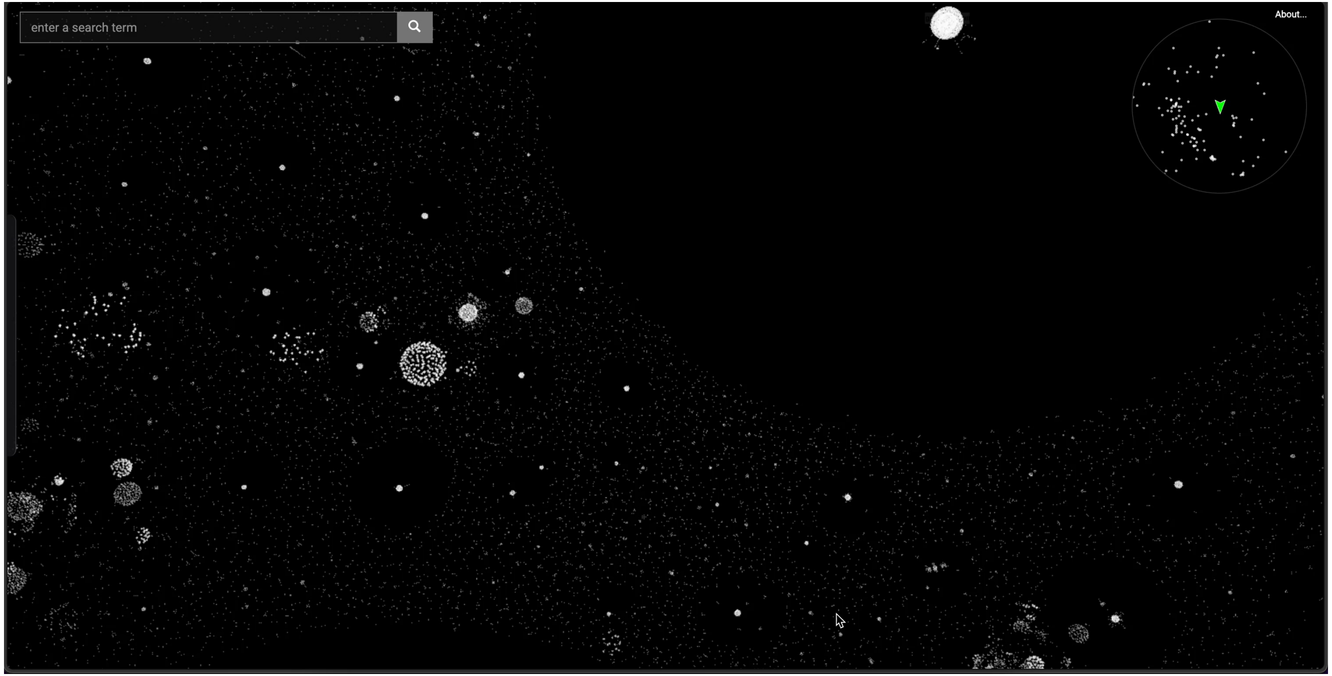


ModelGo Analyzer: Automatic Compliance Analysis Tool for ML Workflow
(More Curation Tools Needed)

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

Moming – NUS
Assets Curation

ML-Asset Curation



Model Galaxy: A Project for Model Dependencies Visualization

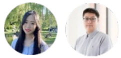
Tutorial Roadmap

Homepage

1 Motivation and Background (00:00 - 00:05)

2 ML-Asset Curation (00:05 - 00:30)

Demo: ModelGo



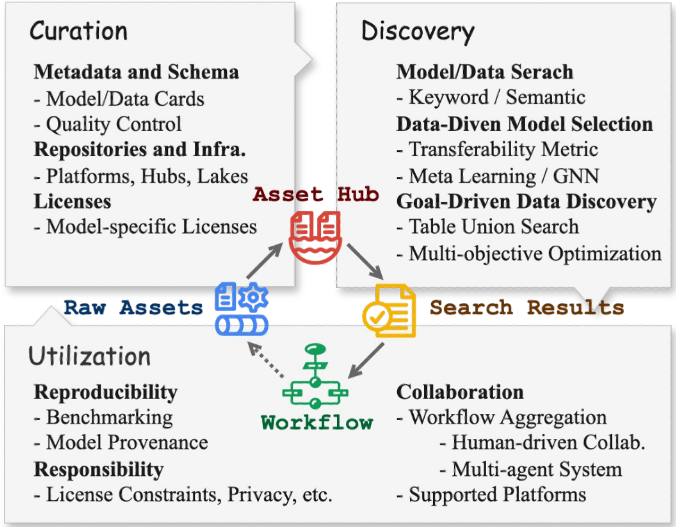
ML-Asset Search and Discovery (00:30 - 00:50)

Demo: CRUX

4 ML-Asset Utilization (00:50 - 01:15)

Demo: Texera

5 System Challenges and Opportunities (01:15 - 01:30)



Mengying – CWRU
Assets Discovery

Why Asset Search Matters

💰 **Valuable Assets**

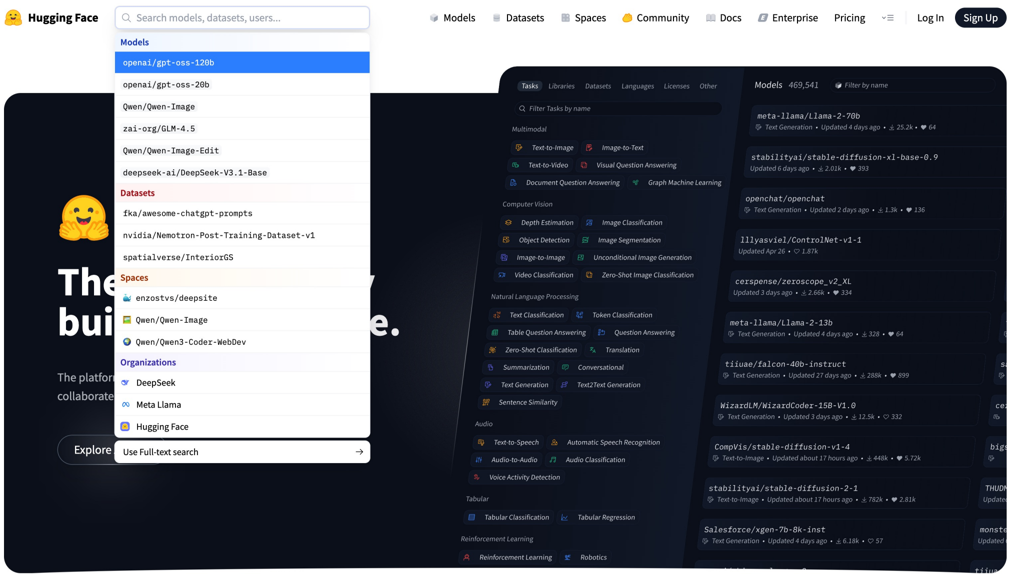
ML assets (models and datasets) are costly to create and maintain.

🚀 **Explosion in Volume:**

Rapidly growing number of assets (e.g., Hugging Face: 2M+ models, 100K added monthly).

⚠️ **Underutilized Resources:**

Over half remain unused due to poor discoverability and insufficient metadata.



Mengying – CWRU
Assets Discovery

Model & Dataset Search

Keyword & Tag-Based Search

Basic filtering on platforms like Hugging Face and Kaggle.

- Faceted search over structured metadata;
- Exact matching on model properties;
- Limited by metadata quality.

Semantic & Vector-Based Retrieval

Embedding models in unified vector spaces.

- Similarity-based search capabilities;
- Vector databases for fast retrieval;
- Contextual understanding of assets.

Graph-Based Discovery

Leveraging relationships and interactions between assets.

- Asset knowledge graphs;
- Performance-based recommendations;
- Exploiting connections among various asset types.

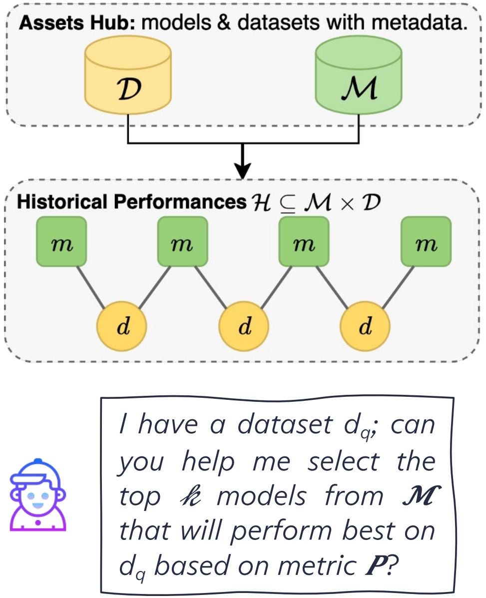
👍 Offer entry points for ML-asset discovery
😟 But tackle various asset types independently and neglect valuable interactions.

Data-driven Model Selection

Problem Definition[1]

Given a collection of models and associated metadata, recommend models with potentially high performance for a 'query' dataset.

- Input:
  - a set of datasets $\mathcal{D}$ with metadata,
  - a set of pre-trained models $\mathcal{M}$ with metadata,
  - a (limited) amount of historical performance $\mathcal{H}$,
  - a model performance measure $\boldsymbol{P}$,
  - integer k, and
  - an example dataset d_q(a "query");
- Output:
  - a set of k pre-trained models from $\mathcal{M}$ with expected good performance $\boldsymbol{P}$ over d_q.



1. Selecting Top-k Data Science Models by Example Dataset, CIKM 2023
Mengying Wang, Sheng Guan, Hanchao Ma, Yiyang Bian, Haolai Che, Abhishek Daundkar, Alp Sehirlioglu, Yinghui Wu

Mengying – CWRU
Assets Discovery

Data-driven Model Selection

Model Selection with AutoML

CASH: Combined Algorithm Selection and Hyper-parameter tuning

- Brute-force methods
  - *Grid Search*
  - *Random Search*
- Learning a "policy" from past attempts
  - Gradient-Based Optimization
  - Bayesian Optimization
  - Genetic Algorithm
  - Reinforcement Learning
- Learning from metadata
  - Meta Learning

🤔 More like model "generation", not "selection".
🫥 Not transparent and customizable.
😮‍💨 Rely on model hyperpara. and perform., high risk to overfitting.
😐 Some approaches don't have guarantee for global optimization.

NAS (for NN model): Neural Architecture Search

- Training a surrogate model
- Predicting its learning curve

Data-driven Model Selection

Model Selection for Transfer Learning

Step 1: Initial Screening

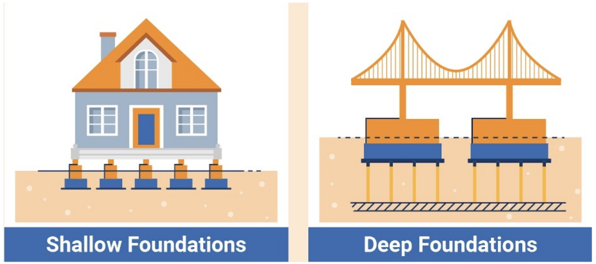
Task type, dataset size, model architecture, and the deep learning framework…For Example,

U-Net - accurately locate a specific area;

👉medical image segmentation;

YOLO - perform faster inference in real-time tasks;

👉object detection in autonomous driving;



The right pre-trained model is like a strong and proper foundation. (Image source: BigRentz)

Step 2: Compute "transferability"

Through source-target relations:

- **Distribution Similarity.** "How alike are the source and target data?"
- **Prediction Output Similarity.** "Do source and target datasets produce similar prediction patterns?"
- **Feature-Label Compatibility.** "How well do pre-trained model features match target dataset labels?".(e.g.,LogME[1])

Can be leveraged to our problem, but ……

🤯 Domain-specific expertise required.

🫠 High-cost prediction or inference.

1. LogME: Practical Assessment of Pre-trained Models for Transfer Learning, ICML 2021
Kaichao You, Yong Liu, Jianmin Wang, Mingsheng Long

Mengying – CWRU
Assets Discovery

Data-driven Model Selection

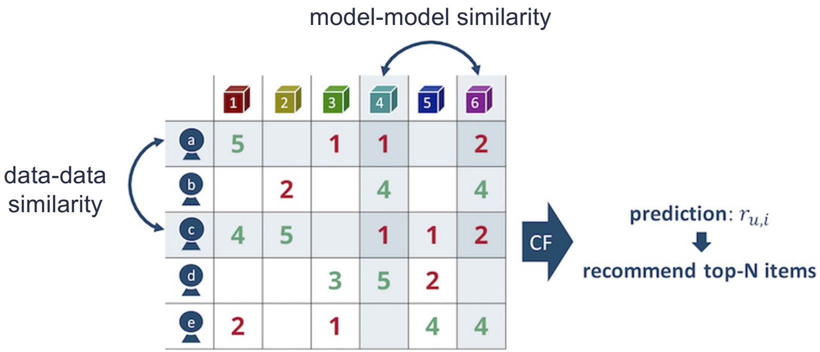
Model Selection by Recommendations

Collaborative Filtering (e.g., Matchbox[1])

- Heavily relies on dataset-model interactions.
- Cannot cope with “cold-start”.
- Hindered by the “ramp-up” issue.

Image source: takuti.github.io

“Cold-start” problem: make recommendations for new datasets that have no interaction records.
“Ramp-up” issue: Recommendation quality is limited until sufficient interaction history accumulates between datasets and models.



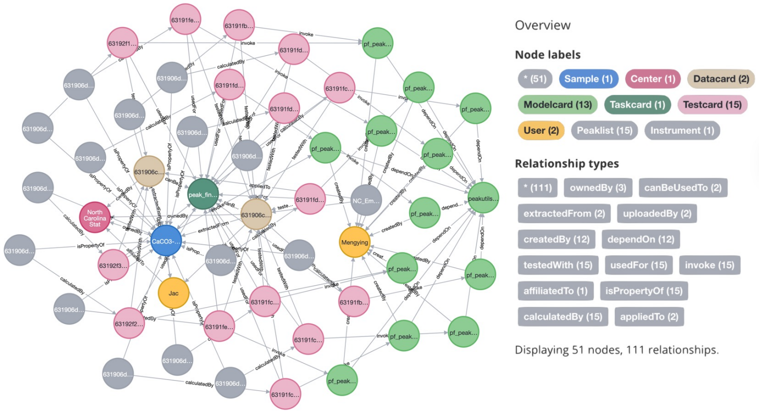
1. Matchbox: Large Scale Bayesian Recommendations, WWW 2009
David Stern, Ralf Herbrich, Thore Graepel

Data-driven Model Selection

Model Selection by Recommendations

Graph-learning Based Recommendation (e.g., ModsNet[1])

- Handles cold-start by leveraging graph metadata and structure.
- Captures rich signals from heterogeneous relations via message passing.
- Rapid ramp-up & explainable through influential paths and graph propagation.

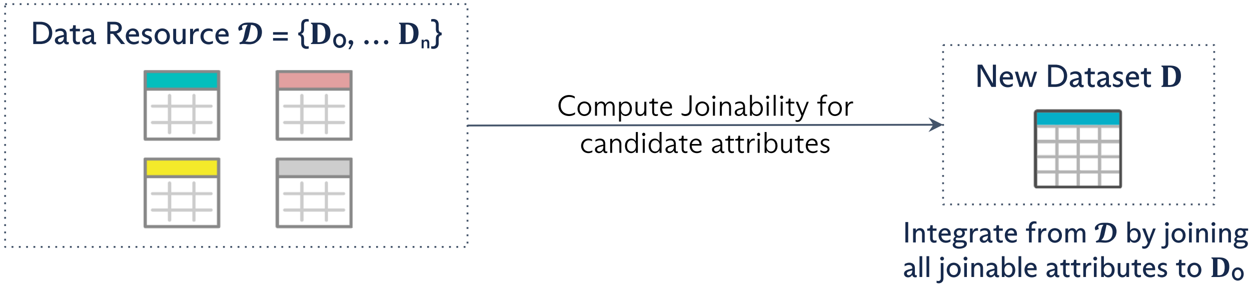


1. ModsNet: Performance-aware Top-k Model Search using Exemplar Datasets, VLDB 2024
Mengying Wang*, Hanchao Ma*, Sheng Guan, Yiyang Bian, Haolai Che, Abhishek Daundkar, Alp Sehirlioglu, Yinghui Wu

Mengying – CWRU
Assets Discovery

Model-driven Data Discovery

Table Union Search

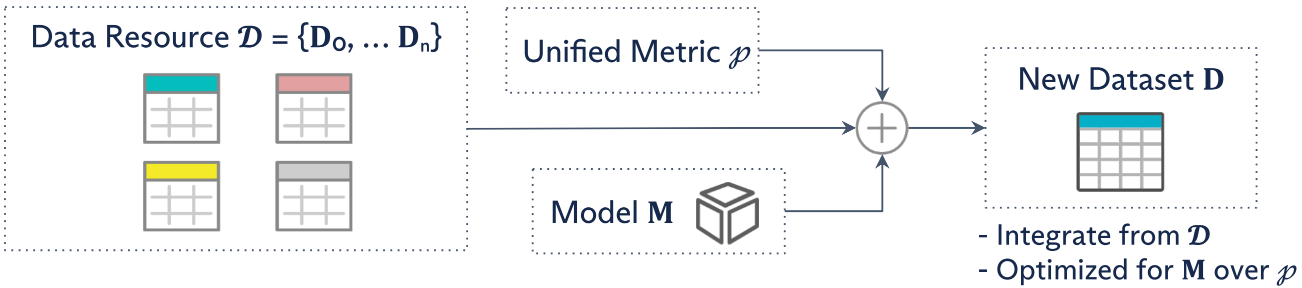


- **Motivation:** Improve dataset completeness and semantic compatibility by integrating relevant tables.
- **Key Method:** Construct semantic table graphs using column embeddings (e.g., LLM-based embeddings), and identify tables for integration through relationship-based semantic matching.
- **Reference:** SANTOS: Relationship-based Semantic Table Union Search, SIGMOD 2023

  Aamod Khatiwada, Grace Fan, Roee Shraga, Zixuan Chen, Wolfgang Gatterbauer, Renée J. Miller, Mirek Riedewald

Model-driven Data Discovery

Goal-driven Data Discovery

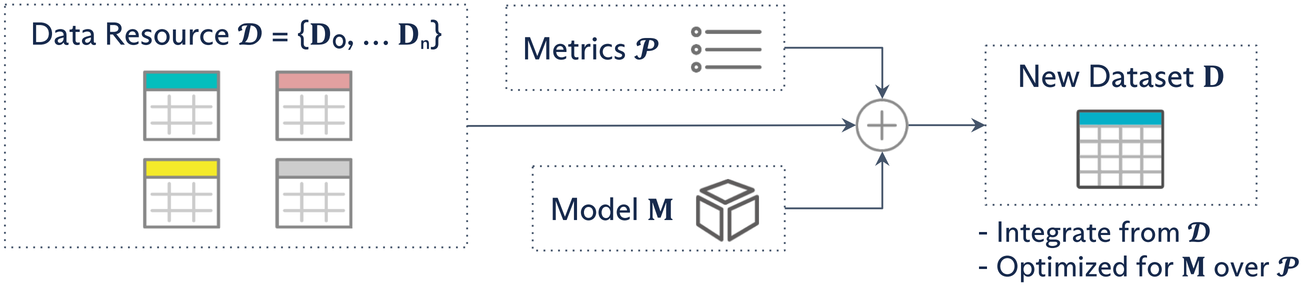


- **Motivation:** Construct datasets explicitly optimized for a single metric tailored to a specific ML task.
- **Key Method:** Iteratively generate candidate datasets, evaluate using a defined utility function (such as combining model performance and cost), and select the best-performing datasets.
- **Reference:** METAM: Goal-Oriented Data Discovery, ICDE 2023

  Sainyam Galhotra, Yue Gong, Raul Castro Fernandez

Model-driven Data Discovery

Multi-objective Data Discovery



- **Motivation:** Simultaneously optimize multiple objectives (e.g., performance, cost) for dataset discovery.
- **Key Method:** Use *multi-objective optimization* (Skyline/Pareto methods) to identify datasets that provide optimal trade-offs among multiple objectives, resulting in a Pareto-optimal set.
- **Reference:** MODis: Generating Skyline Datasets for Data Science (EDBT 25)

  Mengying Wang, Hanchao Ma, Yiyang Bian, Yangxin Fan, Yinghui Wu

Mengying – CWRU
Assets Discovery

Discovery Challenges and Opportunities

Challenge 1: Cold-Start Problem

Limited metadata restricts effective asset retrieval, especially for new or under-documented assets.

Leverage LLM or other methods to retrieve metadata from raw assets automatically.

Challenge 2: Discovery at Scale

Efficiently querying massive ML-asset repositories is computationally expensive.

Scalable infrastructure (distributed vector DBs, caching, hybrid indexes, etc.).

Challenge 3: Semantic Understanding

Interactive discovery requires systems to understand both sides (model and data) at a semantic level.

- Unified representation across multimodal assets;
- advanced NLP and interactive discovery (RAG, conversational AI).

Mengying – CWRU
Assets Discovery

Demo Walkthrough (CRUX[1])

1. **Asset Ingestion**
   Uploading models and datasets with structured metadata forms;
2. **Knowledge Graph Visualization**
   Exploring connections between assets, such as models, datasets, and tests;
3. **Model Recommendation**
   Selecting suitable models from the model repository for a specific(new) dataset;
4. **Data Discovery**
   Generating datasets for a specific model or script by discovering a data repository.

CRUX PROJECT

CWRU

CaseDB

CRUX

U.S. National Science Foundation

1. CRUX: Crowdsourced Materials Science Resource and Workflow Exploration, CIKM 2022
Mengying Wang, Hanchao Ma, Abhishek Daundkar, Sheng Guan, Yiyang Bian, Alp Sehirlioglu, Yinghui Wu

Tutorial Roadmap

Homepage

1 Motivation and Background (00:00 - 00:05)

2 ML-Asset Curation (00:05 - 00:30)

Demo: ModelGo

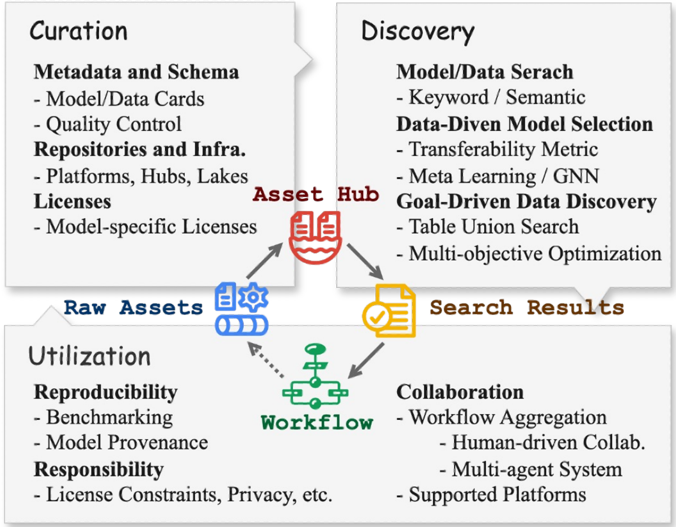
3 ML-Asset Search and Discovery (00:30 - 00:50)

Demo: CRUX

ML-Asset Utilization (00:50 - 01:15)

Demo: Texera

5 System Challenges and Opportunities (01:15 - 01:30)



Mengying – CWRU
Assets Utilization

ML-Asset Utilization

-- Turn curated & discoverable assets into reliable, compliant workflows.

Collaboration[1]

- Modular, versioned **DAG workflows** built from assets; share & reuse (e.g., Texera, Davos).
- Automation trend: **agentic planning** assembles pipelines by reasoning over asset metadata.

Reproducibility[2]

- **Benchmark** using curated datasets/baselines; version-controlled resources improve comparability.
- Capture full **model/data provenance** (data, transforms, hyperparams, code, metrics); adapt **Why-provenance** for explanations.

Responsibility[3]

- **Automated compatibility checks**: Record licenses with an AI BOM (SPDX 3.0) .
- **Close the metadata gap**: use FOSSology, Black Duck, Carneades, ModelGo, and enforce policy in CI.

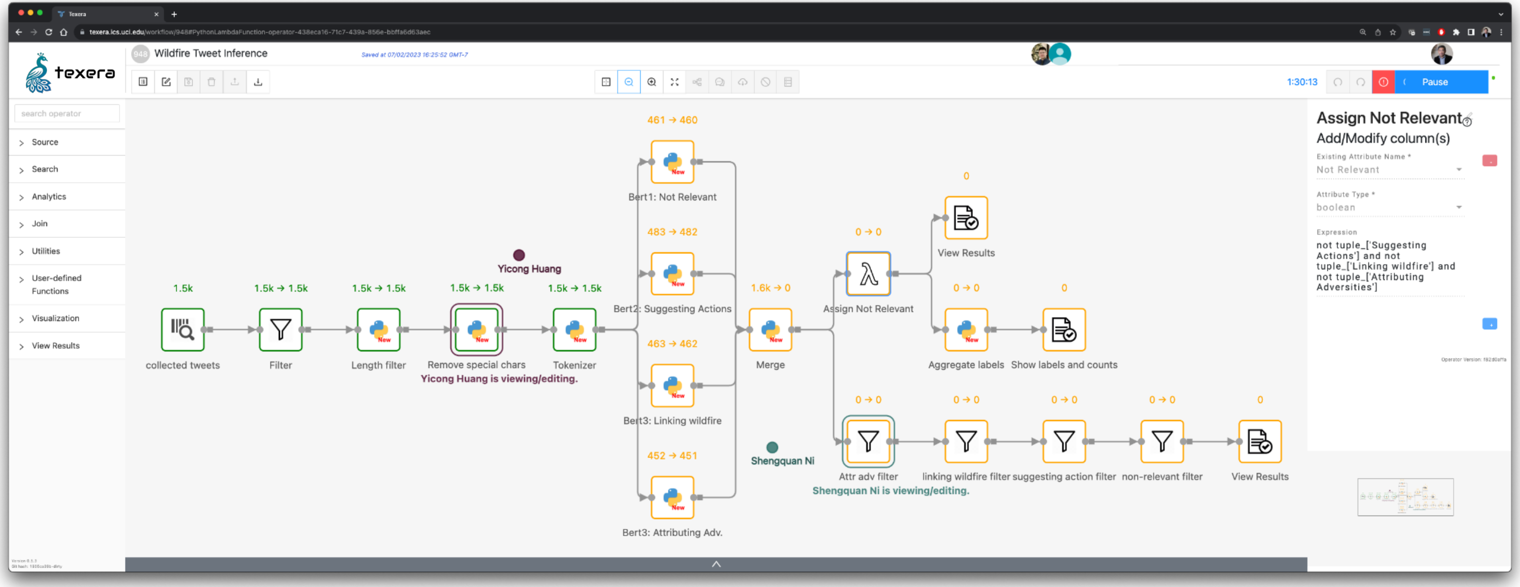
1. Texera: A System for Collaborative and Interactive Data Analytics Using Workflows
   Zuozhi Wang, Yicong Huang, Shengquan Ni, Avinash Kumar, Sadeem Alsudais, Xiaozhen Liu, Xinyuan Lin, Yunyan Ding, and Chen Li.
2. Vamsa: Automated provenance tracking in data science scripts, KDD 2019
   Mohammad Hossein Namaki, Avrilia Floratou, Fotis Psallidas, Subru Krishnan, Ashvin Agrawal, Yinghui Wu, Yiwen Zhu, and Markus Weimer.
3. ModelGo: A practical tool for machine learning license analysis, the Web 2024
Moming Duan, Qinbin Li, and Bingsheng He.

Yicong, Chen – UCI
Assets Utilization

ML-Asset Utilization

Workflow Aggregation and Automation

Texera - A System for Collaborative Data Science, AI, and ML Using Workflows



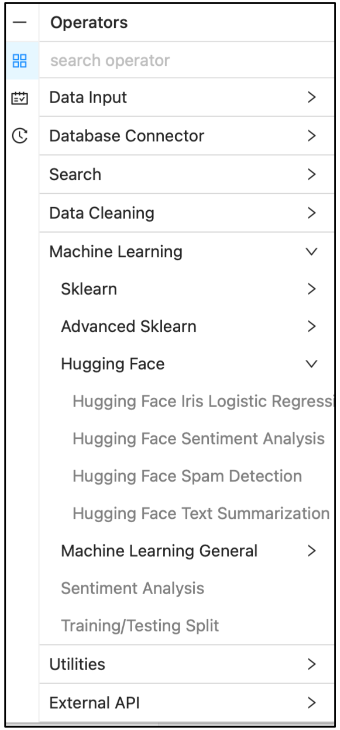
Cloud Service | Collaboration | Distributed Engine | AI/ML Access

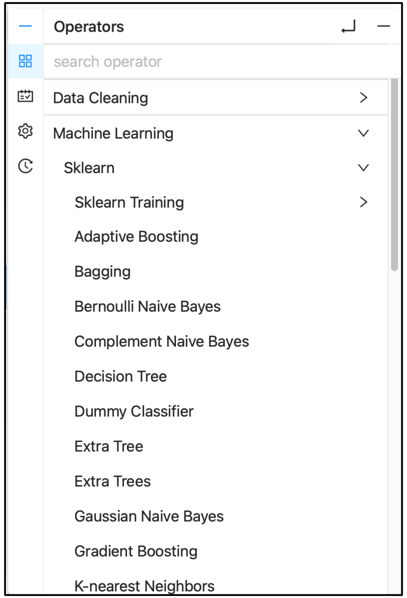
[**VLDB'24**] Texera: A System for Collaborative and Interactive Data Analytics Using Workflows, Zuozhi Wang, **Yicong Huang**, et al.


Yicong, Chen – UCI
Assets Utilization

ML-Asset Utilization

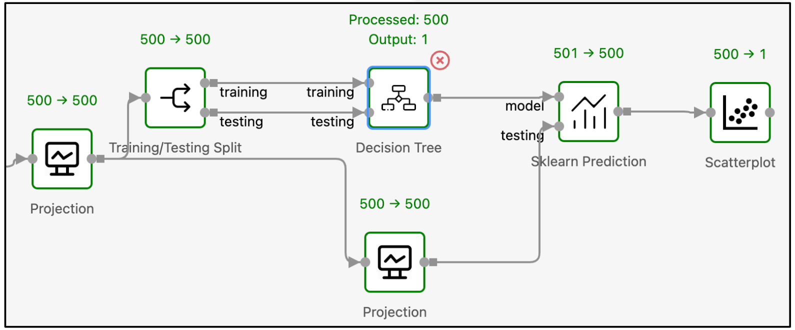
Operators as building blocks

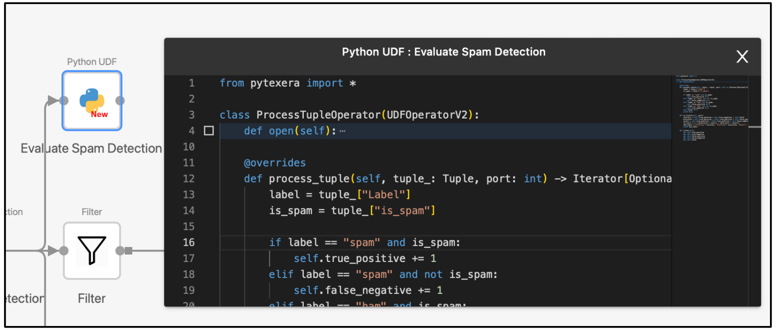




Built-in ML Operators

A Simple ML Workflow

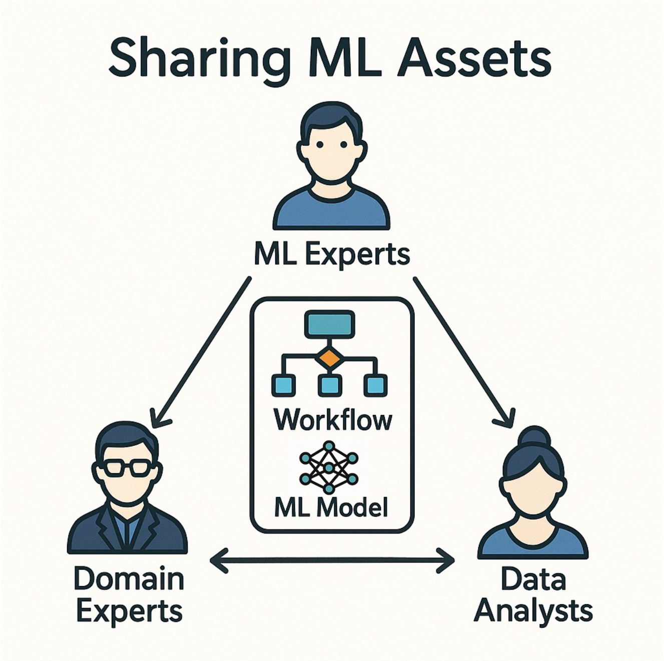


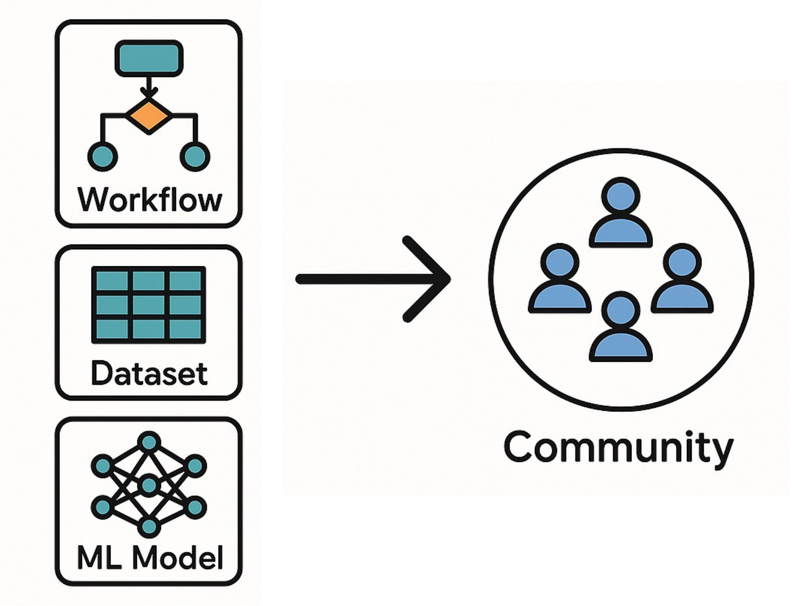


User Defined Functions (UDFs)
For customized ML operators
Java/Scala, Python, R

Collaboration

- Shared ML-Assets between Different Roles and in a Community

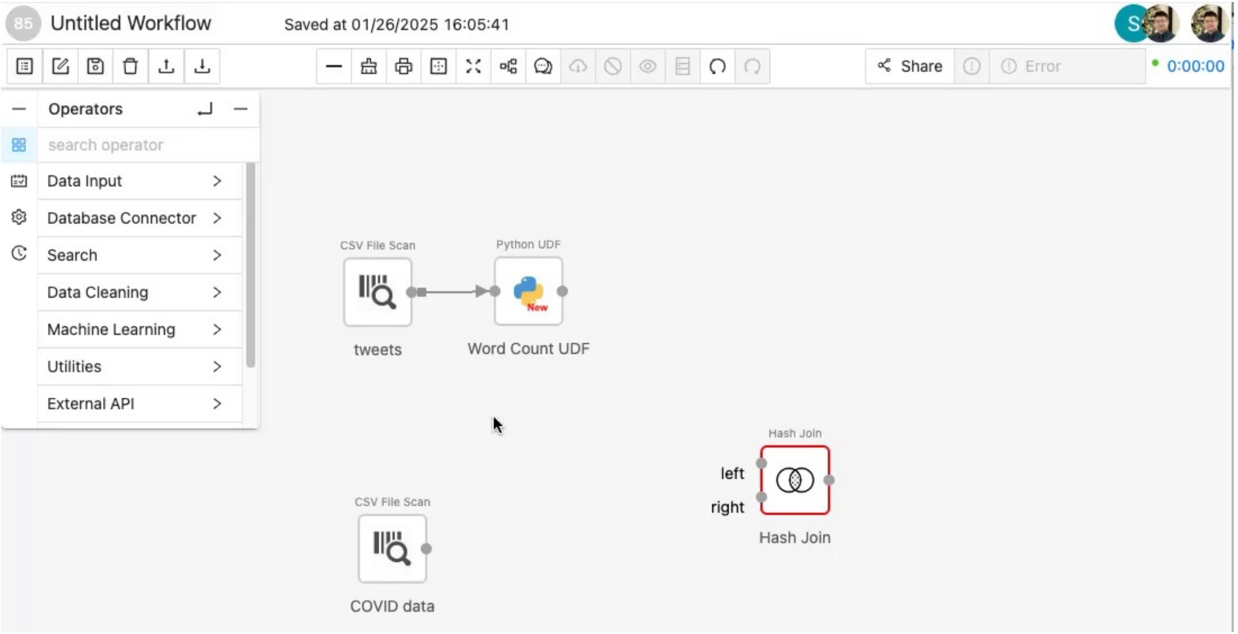




Yicong, Chen – UCI
Assets Utilization

Collaboration

Shared-Editing in a Workflow

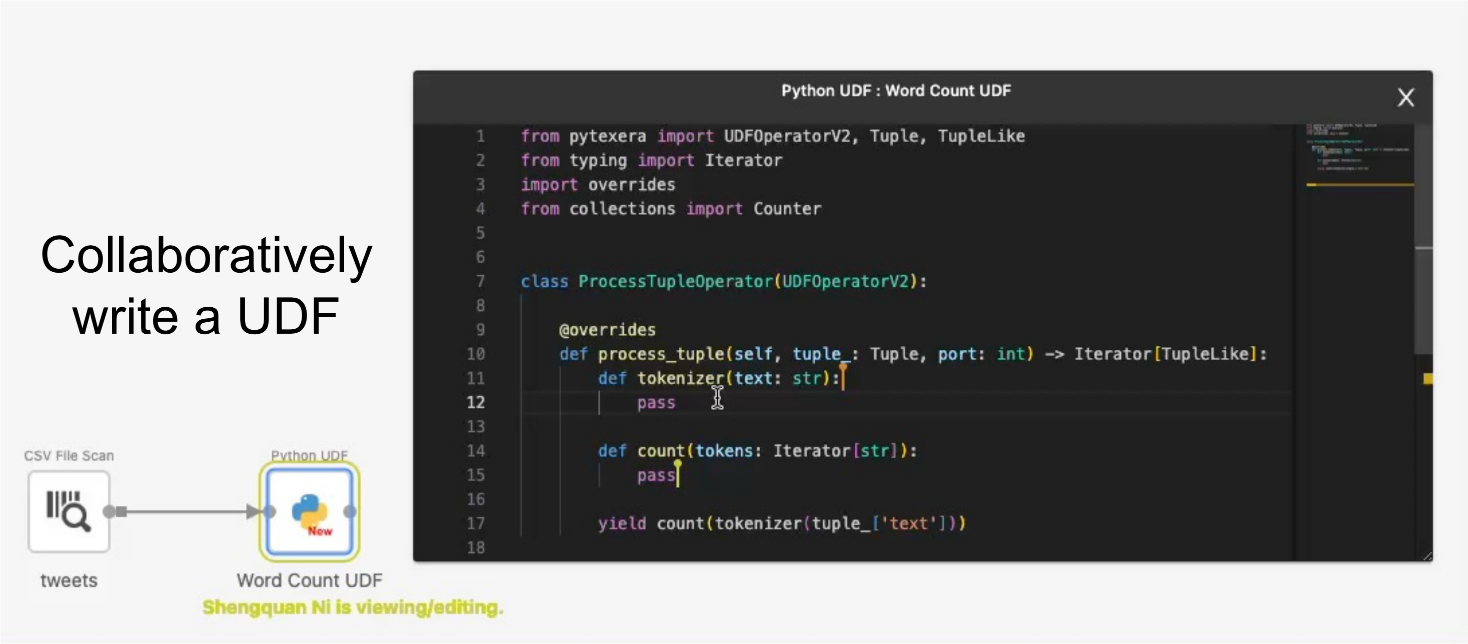


Collaboratively construct a workflow

Yicong, Chen – UCI
Assets Utilization

Collaboration

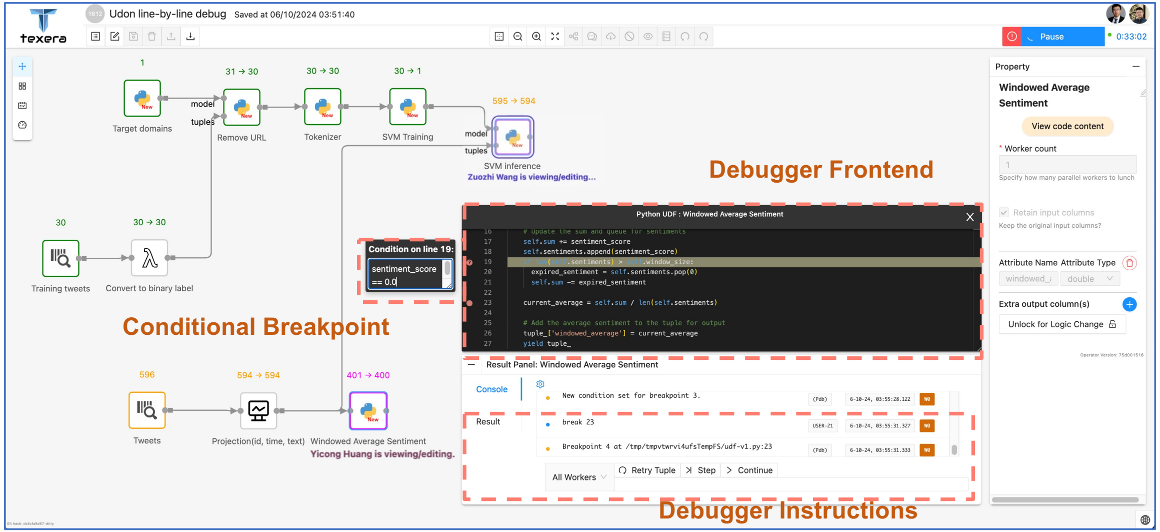
- Shared-Editing in a User-Defined Function (UDF)



Yicong, Chen – UCI
Assets Utilization

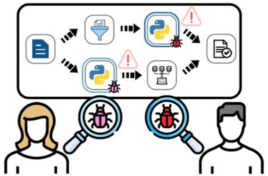
Collaboration

Shared Executions & Shared Debugging



Collaborative Debugging

- Multiple users can share the same execution and share the same debugging session.
- Collaboratively debug the same operator or work on different operators at the same time.

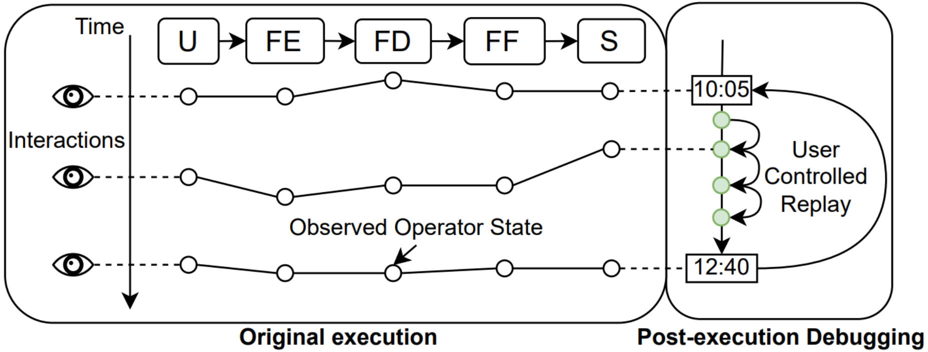


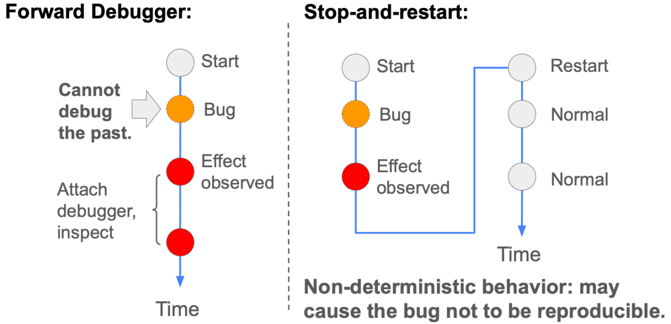
[SIGMOD'24] Udon: Efficient Debugging of User-Defined Functions in Big Data Systems, Yicong Huang, Zuozhi Wang, and Chen Li.

Yicong, Chen – UCI
Assets Utilization

Collaboration

Shared Executions & Shared Debugging





Time-Travel Debugging

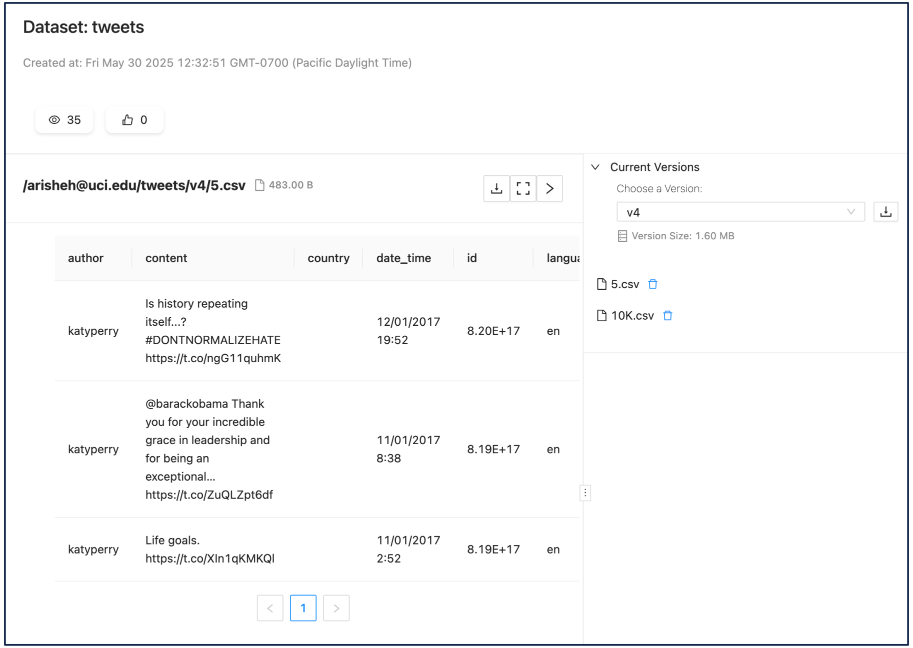
Thu 10:45am – 12:15pm, Rutherford (4F)
Poster in Room: Whittle, Fleming & Britten

[VLDB’25] IcedTea: Efficient and Responsive Time-Travel Debugging in Dataflow Systems, Shengquan Ni, Yicong Huang, Zuozhi Wang, and Chen Li.

Yicong, Chen – UCI
Assets Utilization

Reproducibility

Versioning on ML-Assets

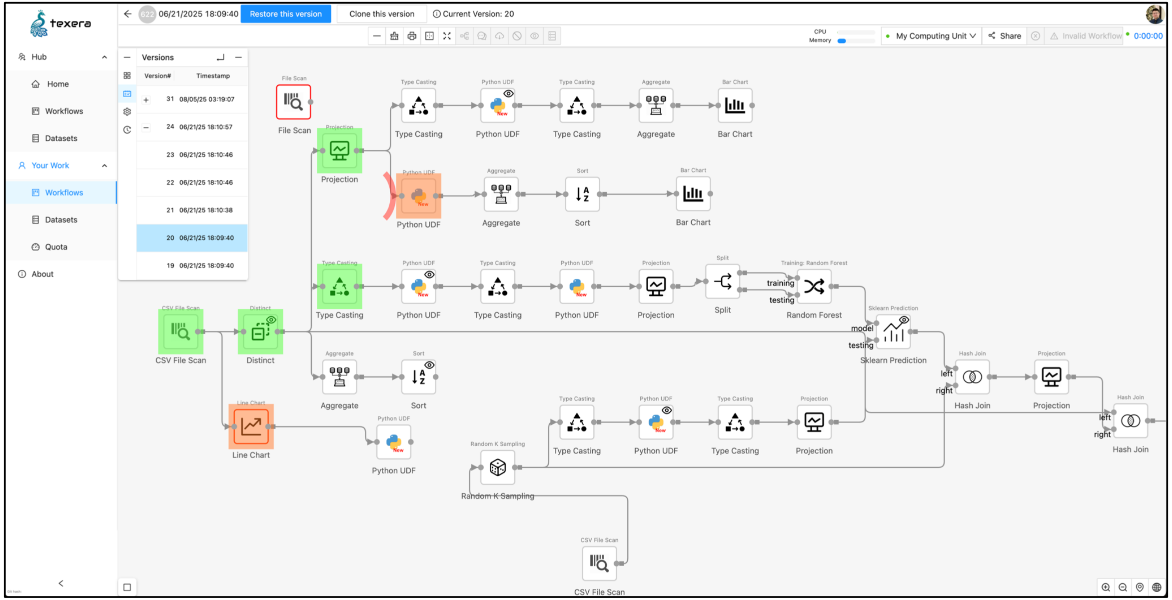


Versioned Dataset/Model Management

Yicong, Chen – UCI
Assets Utilization

Reproducibility

Versioning on ML-Assets



Diff-view between Two Workflow Versions

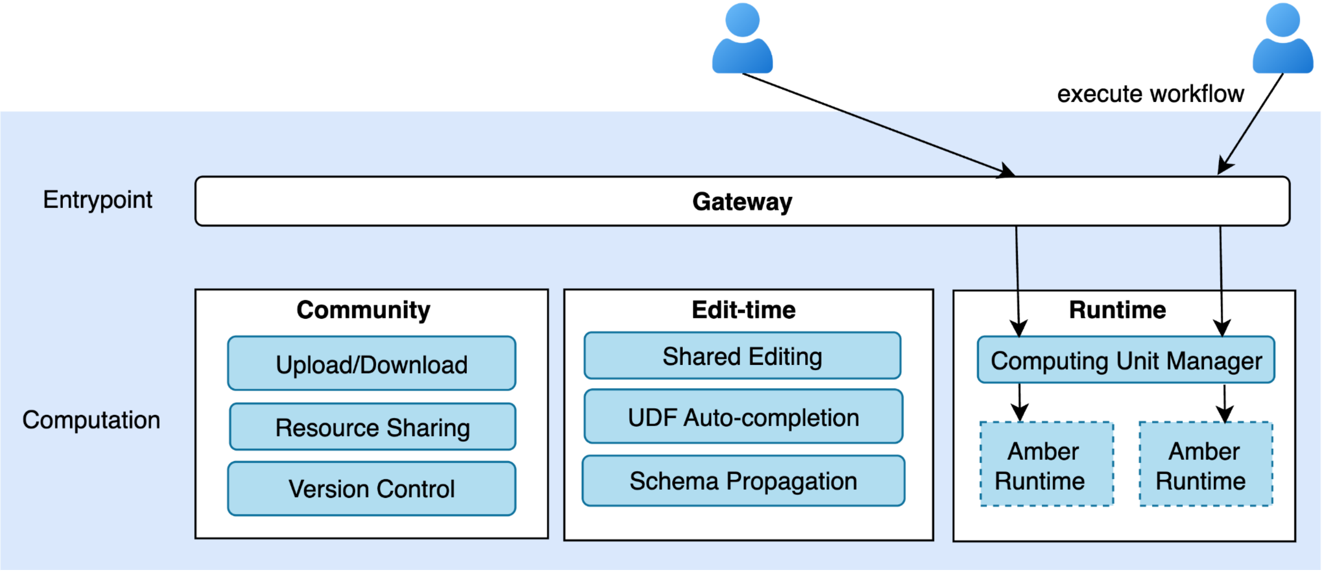
Version#	Timestamp
15	08/17/2023 02:57:54 GMT-7
14	08/17/2023 02:57:53 GMT-7
13	08/17/2023 02:57:52 GMT-7
12	08/17/2023 02:57:51 GMT-7
11	08/17/2023 02:57:50 GMT-7
10	08/17/2023 02:57:49 GMT-7
9	08/17/2023 02:57:47 GMT-7

List of Workflow Versions

Yicong, Chen – UCI
Assets Utilization

Reproducibility

Isolated ENV - Computing Units



texera

Search

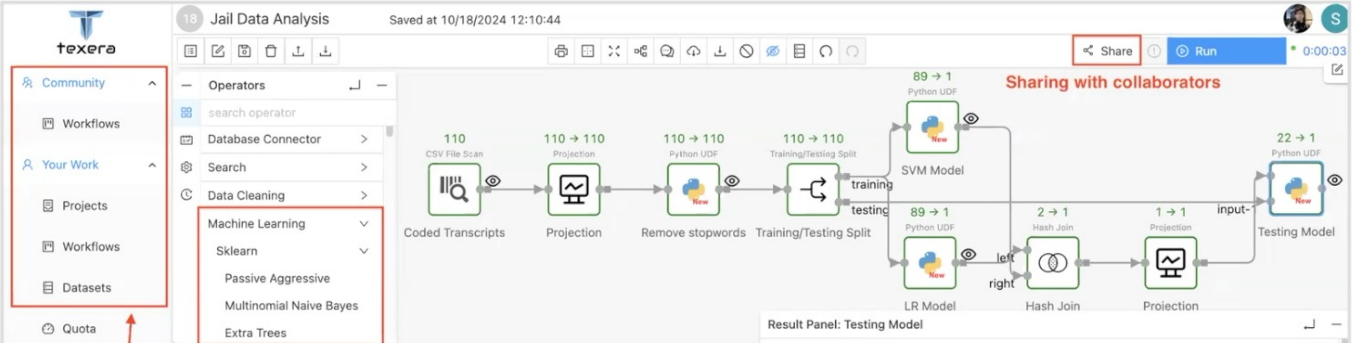
- Hub
  - Home
  - Workflows
  - Datasets
- About

About Texera

Texera is a system for users with various backgrounds to learn and practice data science, AI, and machine learning. It also allows users to share their datasets and analyze workflows.

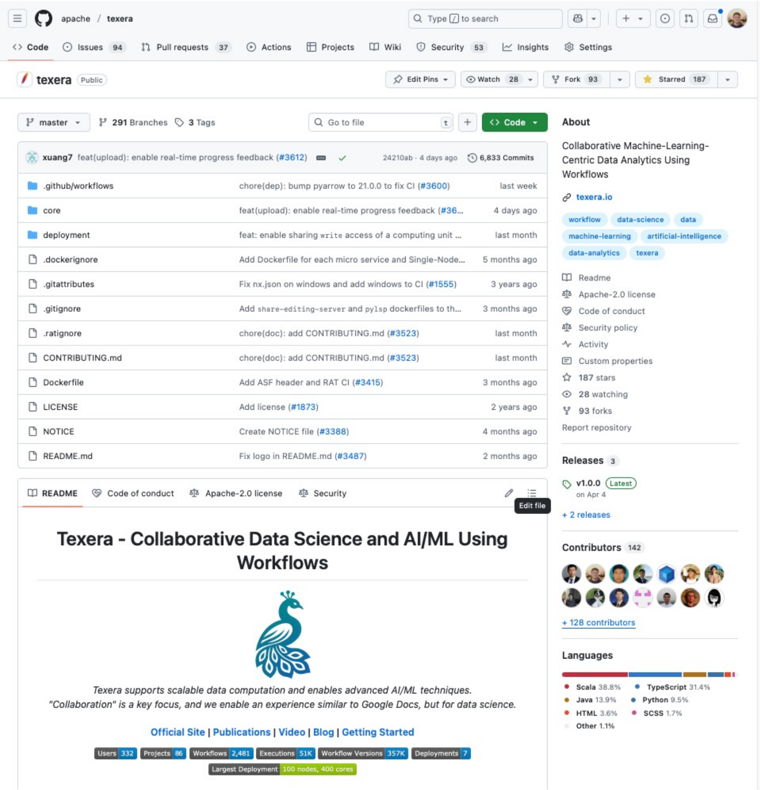
Key Features

- Supporting low-code/no-code data science using workflows
- Parallel data-processing engine running on computing clusters
- Using the Apache Pekko actor-model system
- Supporting UDFs in Python, R, Java, and Scala
- Supporting ML training and inference
- Including a rich collection of ML operators
- Interactive workflow execution model that supports pausing and resuming
- Supporting collaborations with shared editing, shared execution, and version control
- Supporting debugging, including line-by-line debugging in Python UDFs
- Supporting reproducibility of data analysis
- Region-by-region execution with full pipelining in each region
- Storing execution results using Apache Iceberg
- Supporting version-controlled file collections on S3-compatible storage managed by LakeFS
- Adopting a microservice-based architecture using Kubernetes and Docker
- Supporting computing isolation and storage isolation of multiple tenants



Yicong, Chen – UCI
Assets Utilization

Apache Texera (Incubating) – Open Source



$1.5 M

National Institute of Diabetes and Digestive and Kidney Diseases

$10 M

Texera.io

Users	600+	Projects	100+
Workflows	3,000+	Executions	90,000+
Workflow Versions	400,000+	Deployed Servers	8
Collaborating Faculty	17	Involved Undergraduates	100+
Pull Requests	2,352	Development Years	9

* Data till July 2025

Tutorial Roadmap

Homepage

1 Motivation and Background (00:00 - 00:05)

2 ML-Asset Curation (00:05 - 00:30)

Demo: ModelGo

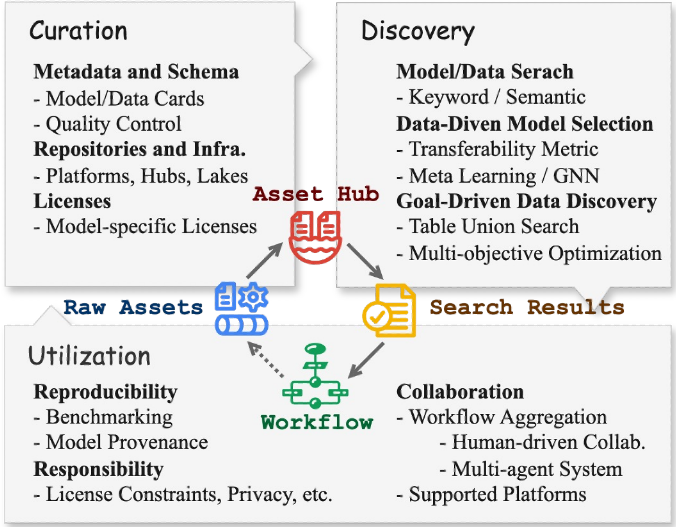
3 ML-Asset Search and Discovery (00:30 - 00:50)

Demo: CRUX

4 ML-Asset Utilization. (00:50 - 01:15)

Demo: Texera

System Challenges and Opportunities (01:15 - 01:30)



Storage, Access, and Scalability

How to reduce storage cost?

- Space-efficient data formats: e.g., Compressed binary formats (Safetensors)
- Storage essential Metadata (ModelDB)
- Distributed Storage (Model Lake)

Challenge: Security, Efficiency, Consistency

Versioning and Lineage

How to track ML assets activities? (reproducibility & auditability)

- Model Version Control (Delta-based version control systems)
- Model Provenance: ProvDB – manage and query ML workflow graphs
- Script Tracking: Vamsa/Geyser - tracking data science scripts as AST graphs
- Model Explainability and Interpretation

Challenge: Scalability, Lineage reusability

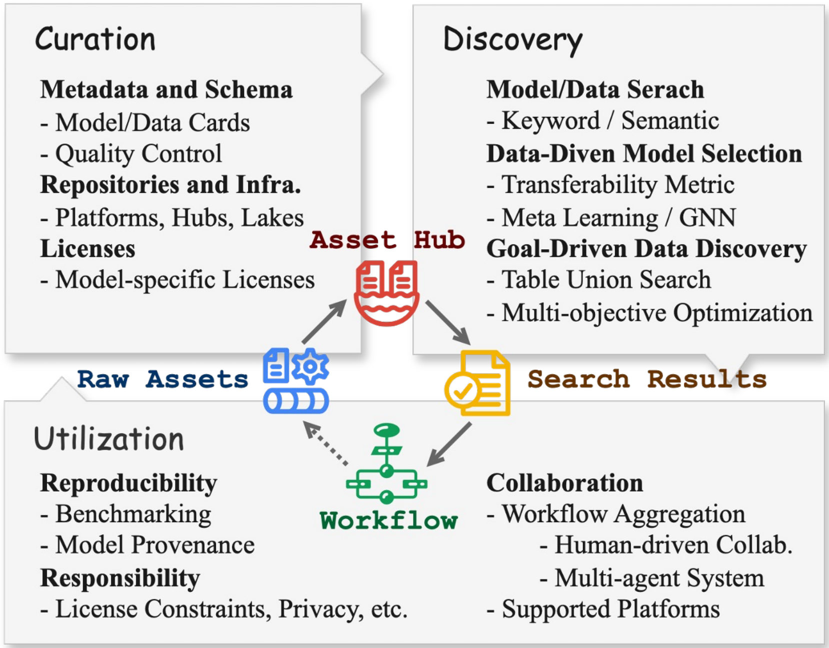
Assets Indexing and Searching

Fast access, search and recommend ML/AI assets

- Surface data types via tab views
- Hybrid indexes
- Vector databases & Vector processing for large-scale ML asset search
- Large Language Models & Retrieval Augmented Generation (RAG) for domain and feature-rich ML/AI applications

Challenge: Maintenance, Privacy and security

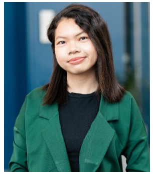
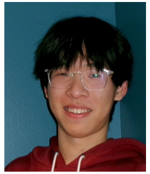
Integrated Solution for Large-scale ML/AI Infrastructure



Yinghui – CWRU
Summary

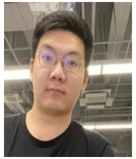
Acknowledgement

Case Western Reserve University (CRUX)

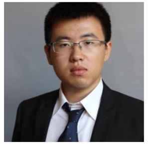
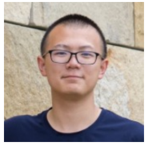
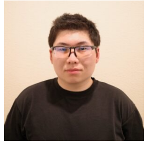
Prof.Alp Sehirlioglu, Hanchao Ma, Sheng Guan, Abhishek Daundkar, Yiyang Bian, Shrinidhi Hegde, Khanh Khuat, Nikki D'Costa, Pengjun Lu

National University of Singapore (ModelGo)

Prof. Nigel Shadbolt, Rui Zhao, Mingzhe Du

University of California, Irvine (Apache Texera)

Zuozhi Wang, Shengquan Ni, Avinash Kumar, Sadeem Alsudais, Xinyuan Lin, Xiaozhen Liu, Yunyan Ding, Jiadong Bai, Ali Risheh

Question & Answering

Homepage


Mengying Wang
CWRU


Moming Duan
NUS National University of Singapore


Yicong Huang
UCI


Chen Li
UCI


Bingsheng He
NUS National University of Singapore


Yinghui Wu
CWRU