The LF AI & Data Foundation is a project of The Linux Foundation that supports open source innovation in artificial intelligence, machine learning, deep learning and data open source projects. The LF AI & Data Foundation was created to support numerous technical projects within this important space.

With the LF AI & Data Foundation, members are working to create a neutral space for harmonization and acceleration of separate technical projects focused on AI, ML, DL and Data technologies.

For more information, please view the How to Get Involved deck.

Questions? Please email info@lfaidata.foundation.

**LF AI & Data Foundation**

- Web Site: [https://lfaidata.foundation/](https://lfaidata.foundation/)
- Landscape: [https://landscape.lfaidata.foundation/](https://landscape.lfaidata.foundation/)
- GitHub: [https://github.com/lfai](https://github.com/lfai)
- Mail Lists: [https://lists.lfaidata.foundation/g/main](https://lists.lfaidata.foundation/g/main)
- Twitter: @LFAIData_Fdn
- Artwork: [https://artwork.lfaidata.foundation](https://artwork.lfaidata.foundation)
- Presentations: Google Slides or MS Powerpoint
- Email: info@lfaidata.foundation
- Slack: [https://slack.lfaidata.foundation/](https://slack.lfaidata.foundation/)

**Current Projects**

<table>
<thead>
<tr>
<th>Project</th>
<th>Status</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Help Us Improve the Wiki</td>
<td></td>
<td>This Wiki is owned by the LF AI Foundation Community. Contributions are always welcomed to help improve it! In the upper right of this page, select Log In to contribute. You will need a Linux Foundation ID (created at <a href="https://identity.linuxfoundation.org/">https://identity.linuxfoundation.org/</a>) to log in. For a Confluence overview, click here.</td>
</tr>
</tbody>
</table>

Welcome to the LF AI & Data Foundation wiki, where you will find information with a cross project focus. For individual projects, follow the links below.

**LF AI & Data Committees**

- **Technical Advisory Council**
  - Wiki: Technical Advisory Council Home
  - Email: tac-general@lists.lfaidata.foundation

- **Outreach Committee**
  - Wiki: Outreach Committee Home
  - Email: outreach-committee@lists.lfaidata.foundation
SANDBOX
1chipML is an open source library for basic numerical crunching and machine learning for microcontrollers. As Internet of Things and Edge Computing are becoming a ubiquitous reality, we need a reliable and open framework to use on limited and low power demanding hardware.
GitHub: https://github.com/1chipML/1chipML

GRADUATE
Acumos is an Open Source Platform, which supports design, integration and deployment of AI models. Furthermore, Acumos supports an AI marketplace that empowers data scientists to publish adaptive AI models, shielding them from the need to create fully integrated solutions.
GitHub: https://github.com/acumos

INCUBATION
Adlik is an end-to-end optimizing framework for deep learning models. The goal of Adlik is to accelerate deep learning inference process both on cloud and embedded environment.
GitHub: https://github.com/Adlik

INCUBATION
Amundsen is a data discovery and metadata engine for improving the productivity of data analysts, data scientists and engineers when interacting with data.
GitHub: https://github.com/amundsen-io
GRADUATE

Angel is a high-performance distributed machine learning platform based on the philosophy of Parameter Server. It is tuned for performance with big data from Tencent and has a wide range of applicability and stability, demonstrating increasing advantage in handling dimensionality.

GitHub: 
https://github.com/Angel-ML/angel

GRADUATE

Adversarial Robustness Toolbox (ART) provides tools that enable developers and researchers to evaluate, defend, and verify Machine Learning models and applications against adversarial threats.

GitHub: 
https://github.com/Trusted-AI/adversarial-robustness-toolbox

INCUBATION

AI Explainability 360 is an open source toolkit that can help users better understand the ways that machine learning models predict labels using a wide variety of techniques throughout the AI application lifecycle.

GitHub: 
https://github.com/Trusted-AI/AIX360

INCUBATION

AI Fairness 360 is an extensible open source toolkit that can help users understand and mitigate bias in machine learning models throughout the AI application lifecycle.

GitHub: 
https://github.com/Trusted-AI/AIF360
Artigraph is a tool to improve the authorship, management, and quality of data. It emphasizes that the core deliverable of a data pipeline or workflow is the data, not the tasks. Artigraph aims to shift tooling focus towards managing the entire data lifecycle (lineage, metadata, schema, storage formats and systems, etc).

GitHub: https://github.com/artigraph/artigraph

BeyondML is a framework for developing sparse neural networks that can perform multiple tasks across multiple data domains. This framework provides value to the community by:

- simplifying the development and deployment of advanced machine learning capabilities for use on low-end devices and in dynamic environments characteristic of the resource-constrained edge
- reducing in the complexity and cost of deploying ML models or systems to cloud platforms
- reducing in the carbon footprint of deployed ML models

GitHub: https://github.com/Beyond-ML-Labs
Within the Bitol project, the primary objective is to tackle multiple challenges, such as data normalization, ensuring the relevance of documentation, establishing service-level expectations, simplifying data and tool integration, and promoting a data product-oriented approach.

These efforts offer several advantages, including stimulating innovation and streamlining integration processes.

Github: https://github.com/bitol-sandbox

CLAIMED (Component Library for AI, Machine Learning, ETL and Data Science) is a runtime and programming language agnostic Data & AI component framework abstracting away all complexity for advanced MLOps and TrustedAI. Donated by the IBM Center for Open Source Data and AI Technologies, the project aims to democratize Data & AI by providing empowerment for Data Scientists to take on MLOps and TrustedAI, as well as Citizen Data Scientists to use no-code/low-code tooling.

Github: https://github.com/claimed-framework/
Datashim is enabling and accelerating data access for Kubernetes/OpenShift workloads in a transparent and declarative way. It’s open source since September of 2019 and it is growing to support use cases related to data in AI projects.

GitHub: https://github.com/IBM/dataset-lifecycle-framework

DataPractices.org was pioneered by data.world as a “Manifesto for Data Practices” of four values and principles that illustrate the most effective, ethical, and modern approach to data teamwork. As a member of the foundation, data.practices.org will expand to offer open courseware and establish a collaborative approach to defining and refining data best practices.

GitHub: https://github.com/datadotworld/data-practices-site

DeepCausality is a hypergeometric computational causality library that enables fast and deterministic context-aware causal reasoning over complex multi-stage causality models. DeepCausality adds only minimal overhead and thus is suitable for real-time applications without additional acceleration hardware.

GitHub: https://github.com/deepcausality-rs/deep_causality
SANDBOX


INCUBATION

DELTA is a deep learning based end-to-end natural language and speech processing platform. DELTA aims to provide easy and fast experiences for using, deploying, developing natural language processing and speech models for both academia and industry cases. DELTA is mainly implemented using TensorFlow and Python 3. GitHub: https://github.com/didi/delta

INCUBATION

DocArray is a library for nested, unstructured, multimodal data in transit, including text, image, audio, video, 3D mesh, etc. It allows deep-learning engineers to efficiently process, embed, search, recommend, store, and transfer the multi-modal data with a Pythonic API. GitHub: https://github.com/docarray
INCUBATION

Elastic Deep Learning (EDL) optimizes the global utilization of the cluster running deep learning jobs and the waiting time of job submitters. It includes two parts: a Kubernetes controller for the elastic scheduling of distributed deep learning jobs, and a fault-tolerable deep learning framework.

GitHub: https://github.com/PaddlePaddle/edl

GRADUATE

Egeria is an open source project dedicated to making metadata open and automatically exchanged between tools and platforms, no matter which vendor they come from.

GitHub: https://github.com/odpi/egeria

SANDBOX

Elyra is an open-source low code/no code framework for creating reproducible, scalable and component-based data science pipelines. It allows senior data scientists to create reusable components easily. Citizen data scientists can reuse their code without programming skills. MLOps engineers are provided with tested and maintainable deliverables, and scale on Kubeflow, Airflow and others.

Github: https://github.com/elyra-ai/elyra
INCUBATION

**FATE**

INCUBATION

**Feast**

INCUBATION
Feathr is an enterprise-grade, high-performance feature store. Feathr automatically computes your feature values and joins them to your training data, using point-in-time-correct semantics to avoid data leakage, and supports materializing and deploying your features for use online in production.

GitHub: https://github.com/linkedin/feathr

FlagAI (Fast Large-scale General AI models) is a fast, easy-to-use and extensible toolkit for large-scale model. Our goal is to support training, fine-tuning, and deployment of large-scale models on various downstream tasks with multi-modality. Currently, we are focusing on NLP models and tasks. In the near future, we will support other modalities.

GitHub: https://github.com/BAAI-Open/FlagAI

Flyte is a production-grade, declarative, structured and highly scalable cloud-native workflow orchestration platform. It allows users to describe their ML/Data pipelines using Python, Java or (in the future) other languages and Flyte manages the data flow, parallelization, scaling and orchestration of these pipelines. Flyte builds on top of Docker containers and Kubernetes.

GitHub: https://github.com/flyteorg/flyte
ForestFlow is a scalable policy-based cloud-native machine learning model server. ForestFlow strives to strike a balance between the flexibility it offers data scientists and the adoption of standards while reducing friction between Data Science, Engineering and Operations teams. GitHub: https://github.com/ForestFlow/ForestFlow

Horovod, a distributed training framework for TensorFlow, Keras and PyTorch, improves speed, scale and resource allocation in machine learning training activities. Uber uses Horovod for self-driving vehicles, fraud detection, and trip forecasting. It is also being used by Alibaba, Amazon and NVIDIA. Contributors to the project outside Uber include Amazon, IBM, Intel and NVIDIA. GitHub: https://github.com/horovod/horovod
Intersectional Fairness (ISF) is a bias detection and mitigation technology for intersectional bias, which combinations of multiple protected attributes cause. ISF leverages the existing single-attribute bias mitigation methods to make a machine-learning model fair regarding intersectional bias.

Approaches applicable to ISF are pre-, in-, and post-processing. Currently, ISF supports Adversarial Debias, Equal Odds, Massaging, and Reject Option Classification.

GitHub: https://github.com/intersectional-fairness/isf
The goal of the Interoperability Initiative is to enable voice and conversational AI to work like the web. Before us is a future where users can freely find and make use of any conversational assistant and language model that addresses their goals, just as they do now with web pages.

Our path toward achieving this goal is to define, develop, and promote standards—beginning with an open, universal application programming interface (API) which enable conversational assistants to follow standards to freely interoperate with each other, using APIs to connect, communicate, and transfer content and control across assistants, platforms, and language models.

We encourage you to learn more about our interoperability work by following the links below, where you can learn about our approach, process, future work, and how to get involved.

JanusGraph is a scalable graph database optimized for storing and querying graphs containing hundreds of billions of vertices and edges distributed across a multi-machine cluster.

GitHub: https://github.com/janusgraph/janusgraph
Kedro

Kedro is an open-source Python framework for creating reproducible, maintainable and modular data science code. It borrows concepts from software engineering best-practice and applies them to machine-learning code; applied concepts include modularity, separation of concerns and versioning.

GitHub: https://github.com/kedro-org

Kompute

Kompute is a general purpose GPU compute framework for cross vendor graphics cards (AMD, Qualcomm, NVIDIA & friends). Blazing fast, mobile-enabled, asynchronous and optimized for advanced GPU data processing use cases.

GitHub: https://github.com/KomputeProject
<table>
<thead>
<tr>
<th>INCUBATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>KServe provides a Kubernetes Custom Resource Definition for serving machine learning (ML) models on arbitrary frameworks. It aims to solve production model serving use cases by providing performant, high abstraction interfaces for common ML frameworks like Tensorflow, XGBoost, ScikitL, PyTorch, and ONNX. It encapsulates the complexity of autoscaling, networking, health checking, and server configuration to bring cutting edge serving features like GPU Autoscaling, Scale to Zero, and Canary Rollouts to your ML deployments. It enables a simple, pluggable, and complete story for Production ML Serving including prediction, pre-processing, post-processing and explainability. GitHub: <a href="https://github.com/kserve">https://github.com/kserve</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SANDBOX</th>
</tr>
</thead>
<tbody>
<tr>
<td>LakeSoul is a cloud-native Lakehouse framework developed by the MetaSoul team, and supports scalable metadata management, ACID transactions, efficient and flexible upsert operation, schema evolution, and unified streaming &amp; batch processing. GitHub: <a href="https://github.com/meta-soul/LakeSoul">https://github.com/meta-soul/LakeSoul</a></td>
</tr>
</tbody>
</table>
Ludwig is a toolbox built on top of TensorFlow that allows to train and test deep learning models without the need to write code. All you need to provide is your data, a list of fields to use as inputs, and a list of fields to use as outputs, Ludwig will do the rest.

Simple commands can be used to train models both locally and in a distributed way, and to use them to predict on new data. GitHub: https://github.com/uber/ludwig

Marquez is an open source metadata service for the collection, aggregation, and visualization of a data ecosystem's metadata. It maintains the provenance of how datasets are consumed and produced, provides global visibility into job runtime and frequency of dataset access, centralization of dataset lifecycle management, and much more.

GitHub: https://github.com/MarquezProject

Milvus is an open source similarity search engine for massive-scale feature vectors. Built with heterogeneous computing architecture for the best cost efficiency. Searches over billion-scale vectors take only milliseconds with minimum computing resources. Milvus can be used in a wide variety of scenarios to boost AI development.

GitHub: https://github.com/milvus-io
INCUBATION
NNStreamer (Neural Network Support as Gstreamer Plugins) is a set of Gstreamer plugins that support ease and efficiency for Gstreamer developers adopting new neural network models and managing neural network pipelines.
GitHub: https://github.com/nnstreamer

SANDBOX
OpenBytes aims to facilitate sharing, collaboration, and community through the promotion of data standards and formats enabling contributions of data. The value of this project lies in its stimulus on academic interest and AI innovation by promoting high-quality datasets and pushing the boundaries of science.
GitHub: https://github.com/ProjectOpenBytes

SANDBOX
OpenDataology is an open-source compliance analysis project. It enables users of publicly available datasets and users who curate datasets from multiple data sources (particularly for use as a part of machine learning models) to identify the potential license compliance risks. OpenDataology consists primarily of three key components.
dataset license compliance analysis workflow to ascertain allowed rights and the required obligation associated with using a publicly available dataset. A growing database and web portal document rights obligations (after license compliance analysis conducted associated with datasets and the sources analyzed in the project). Database document metadata collected and used conduct compliance workflow.
An online license generation toolkit creators datasets can use generate custom licenses, depending on the exact rights obligations that want to (instead of having to rely on limited dataset specific licenses).

GitHub: https://github.com/OpenDataology

OpenDS4All is a project created to accelerate the creation of data science curricula at academic institutions. Our goal is to provide recommendations, slide sets, sample Jupyter notebooks, and other materials for creating, customizing, and delivering data science and data engineering education.

GitHub: https://github.com/odpi/OpenDS4All

OpenFL is a Python 3 library for federated learning that enables organizations to collaboratively train a model without sharing sensitive information.

GitHub: https://github.com/intel/openfl
OpenLineage proposes an open standard and API for lineage collection that data processing engines can implement to publish at runtime details of the data sources it is reading, the types of processing it is performing, and the destination of the results. [GitHub](https://github.com/OpenLineage)

ONNX is an open format to represent deep learning models. With ONNX, AI developers can more easily move models between state-of-the-art tools and choose the combination that is best for them. ONNX is developed and supported by a community of partners. [GitHub](https://github.com/onnx)

Pyro is a universal probabilistic programming language (PPL) written in Python and supported by PyTorch on the backend. Pyro enables flexible and expressive deep probabilistic modeling, unifying the best of modern deep learning and Bayesian modeling. [GitHub](https://github.com/pyro-ppl/pyro)
The Recommenders repository provides examples and best practices for building recommendation systems, provided as Jupyter notebooks. The `recommenders` module contains functions to simplify common tasks used when developing and evaluating recommender systems.

GitHub: [https://github.com/Microsoft/Recommenders](https://github.com/Microsoft/Recommenders)

---

RosaeNLG is a template-based Natural Language Generation (NLG) tool that automates the production of relatively repetitive texts based on structured input data and textual templates, run by a NLG engine. Production usage is widespread in large corporations, especially in the financial industry.

GitHub: [https://github.com/RosaeNLG/](https://github.com/RosaeNLG/)

---

RWKV

RWKV is an RNN with transformer-level LLM performance. It can be directly trained like a GPT (parallelizable). So it combines the best of RNN and transformer excellence: great performance, fast inference, saves VRAM, fast training, “infinite” ctx_len, and free sentence embedding.

GitHub: [https://github.com/rwkv/](https://github.com/rwkv/)
Automatic machine learning, commonly referred to as AutoML, holds great promise in democratizing the utilization of machine learning by automating a large portion of the typical performance work typically performed by data scientists.

However, the vast search space of potential pipelines poses a challenge, often resulting in suboptimal or no pipelines being generated, particularly when dealing with large and complex datasets. SapientML addresses this issue by leveraging a collection of pre-existing datasets and their human-created pipelines, enabling efficient generation of high-quality pipelines for new datasets with predictive tasks. With SapientML, data scientists can rapidly create and amend AI models, as the code is provided along with detailed explanations. Furthermore, citizen data scientists can easily create the desired AI models as well.

GitHub: https://github.com/sapientml

ShaderNN is a lightweight deep learning inference framework optimized for Convolutional Neural Networks. It provides high-performance inference for deep learning applications in image and graphics processing on mobile devices.

GitHub: https://github.com/inferenceengine/shadernn
SOAJS is an open source microservices and API management platform, SOAJS eliminates the IT plumbing challenges, so you can deploy microservices significantly earlier and faster. IT initiatives such as digital transformation are simplified, accelerated, cost reduced, and risk mitigated. Our fully integrated, world-class API lifecycle management, multi-cloud orchestration, release management, and IT Ops automation capabilities eliminate your IT organization’s modernization pain.

GitHub: https://github.com/soajs

Substra is a framework offering distributed orchestration of machine learning tasks among partners while guaranteeing secure and trustless traceability of all operations. It enables privacy-preserving federated learning projects, where multiple parties collaborate on a Machine Learning objective while each one keeps their private datasets behind their own firewall.

GitHub: https://github.com/SubstraFoundation/substra
incubation

sparklyr is an R package that lets you analyze data in Spark while using familiar tools in R. sparklyr supports a complete backend for dplyr, a popular tool for working with data frame objects both in memory and out of memory. You can use dplyr to translate R code into Spark SQL.

GitHub: 
https://github.com/sparklyr/sparklyr

trustmark initiative of open voice network

incubation

Trustmark works to translate ethical principles specific to conversational AI into risk mitigation for developers, enterprise users — principles, LF EdX education, maturity model, and specifications.

sandbox

Xtreme1 is the next generation open source platform for multi-sensory training data. It accelerates the modeling process by advanced AI-powered tools, thousands of projects distilled ontologies, and plentiful data curation features.

GitHub: 
https://github.com/basicai/xtreme1

Recent space activity

Nancy Rausch
Technical Advisory Council (TAC) updated Feb 06, 2024

Nathan Southern
Outreach Committee updated Jan 24, 2024

Space contributors

- Nancy Rausch (16 days ago)
- Nathan Southern (30 days ago)
- Andreas Fehlner (49 days ago)
- Cupid Chan (63 days ago)
- Reden Martinez (112 days ago)
- ...