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.

**Current Projects**

<table>
<thead>
<tr>
<th>Project</th>
<th>Status</th>
<th>Description</th>
</tr>
</thead>
</table>
**SANDBOX**

1chipML is an open source library for basic numerical crunching and machine learning for microcontrollers. As the 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](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, while shielding them from the need to custom develop fully integrated solutions.

GitHub: [https://github.com/acumos](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](https://github.com/Adlik)
| **Amundsen** | 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](https://github.com/amundsen-io) |
| **Angel** | 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 higher dimension models. GitHub: [https://github.com/Angel-ML/angel](https://github.com/Angel-ML/angel) |
| **Adversarial Robustness Toolbox (ART)** | Adversarial Robustness Toolbox (ART) provides tools that enable developers and researchers to evaluate, defend, certify and verify Machine Learning models and applications against the adversarial threats. GitHub: [https://github.com/Trusted-AI/adversarial-robustness-toolbox](https://github.com/Trusted-AI/adversarial-robustness-toolbox) |
**AI Explainability 360**
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](https://github.com/Trusted-AI/AIX360)

**AI Fairness 360**
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](https://github.com/Trusted-AI/AIF360)

**Artigraph**
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](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 of models to cloud platforms
- reducing in the carbon footprint of deployed ML models

GitHub: https://github.com/Beyond-ML-Labs

Datashim is enabling and accelerating data access for Kubernetes /OpenShift workloads in a transparent and declarative way. It's opensource since September of 2019 and it is growing to support use-cases related to data access 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 12 principles that illustrate the most effective, ethical, and modern approach to data teamwork. As a member of the foundation, datapractices.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

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, and developing natural language processing and speech models for both academia and industry use cases. DELTA is mainly implemented using TensorFlow and Python 3.

GitHub: https://github.com/didi/delta
<table>
<thead>
<tr>
<th>INCUBATION</th>
</tr>
</thead>
</table>
| Elastic Deep Learning (EDL) optimizes the global utilization of the cluster running deep learning job 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](https://github.com/PaddlePaddle/edl) |

<table>
<thead>
<tr>
<th>GRADUATE</th>
</tr>
</thead>
</table>
| 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](https://github.com/odpi/egeria) |
FATE (Federated AI Technology Enabler) is the world's first industrial grade federated learning open source framework to enable enterprises and institutions to collaborate on data while protecting data security and privacy. It implements secure computation protocols based on homomorphic encryption and multi-party computation (MPC). Supporting various federated learning scenarios, FATE now provides a host of federated learning algorithms, including logistic regression, tree-based algorithms, deep learning and transfer learning.
Feast is an open source feature store for machine learning. It was developed as a collaboration between Gojek and Google in 2018. Feast aims to: -- Provide scalable and performant access to feature data for ML models during training or serving. -- Provide a consistent view of features for both training and serving. -- Enable re-use of features through discovery, documentation, and metadata tracking. -- Ensure model performance by tracking, validating, and monitoring features in production.

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 for 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
**INCUBATION**

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](https://github.com/ForestFlow/ForestFlow)

**GRADUATE**

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](https://github.com/horovod/horovod)
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 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 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
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, ScikitLearn, 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: https://github.com/kserve
**Ludwig**

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](https://github.com/uber/ludwig)

---

**Marquez**

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](https://github.com/MarquezProject)
| INCUBATION | 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 neural network models and neural network developers managing neural network pipelines and their filters. GitHub: https://github.com/nntstreamer |
OpenBytes

OpenBytes aims to facilitate wider sharing of, and collaboration with, data in the AI community through the promotion of data standards and formats and 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 further.

GitHub: https://github.com/Project-OpenBytes

OpenDataology

OpenDataology is an open source dataset license compliance analysis project. It enables users of publicly available datasets and users who curate a dataset from multiple data sources (particularly for use as a part of machine learning models) to identify the potential license compliance risks. The project is primarily comprised of three key components.
A dataset license compliance analysis workflow that ascertains the final allowed rights and the required obligations associated with using a publicly available dataset or a dataset that is curated from multiple data sources for any purpose.

A growing database and a web portal that documents the final rights and obligations (after the license compliance analysis is conducted) associated with the datasets and the data sources analyzed in our project. The database also documents the metadata collected and used to conduct the compliance workflow.
An online license generation toolkit that creators of dataset to generate custom licenses depending on the exact rights and obligations that they want to allow (instead of having to rely on existing available and limited dataset specific licenses).

GitHub: [https://github.com/OpenDataology](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](https://github.com/odpi/OpenDS4All)
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

RosaeNLG is a template-based Natural Language Generation (NLG) 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/
<table>
<thead>
<tr>
<th><strong>SOAJS</strong></th>
<th>INCUBATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td>GitHub: <a href="https://github.com/soajs">https://github.com/soajs</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Substra Framework</strong> (Logo to be updated)</th>
<th>INCUBATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td>GitHub: <a href="https://github.com/SubstraFoundation/substra">https://github.com/SubstraFoundation/substra</a></td>
</tr>
</tbody>
</table>
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

---

### Recent space activity

<table>
<thead>
<tr>
<th>Name</th>
<th>Role</th>
<th>Last Updated</th>
<th>View Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cupid Chan</td>
<td>BI &amp; AI</td>
<td>Sep 28, 2022</td>
<td>view change</td>
</tr>
<tr>
<td>Nancy Hausch</td>
<td>Technical Advisory Council (TAC)</td>
<td>Sep 28, 2022</td>
<td>view change</td>
</tr>
<tr>
<td>Erin Thacker</td>
<td>Outreach Committee</td>
<td>Sep 16, 2022</td>
<td>view change</td>
</tr>
<tr>
<td>dai fengxin</td>
<td>LF AI &amp; DATA MLOps Committee</td>
<td>Sep 14, 2022</td>
<td>view change</td>
</tr>
<tr>
<td>Alejandro Saucedo</td>
<td>MLSecOps Working Group</td>
<td>Sep 11, 2022</td>
<td>view change</td>
</tr>
</tbody>
</table>

### Space contributors

- Cupid Chan (1 day ago)
- Nancy Rausch (1 day ago)
- Erin Thacker (13 days ago)
- dai fengxin (15 days ago)
- Alejandro Saucedo (18 days ago)
- ...