

LF AI & Data

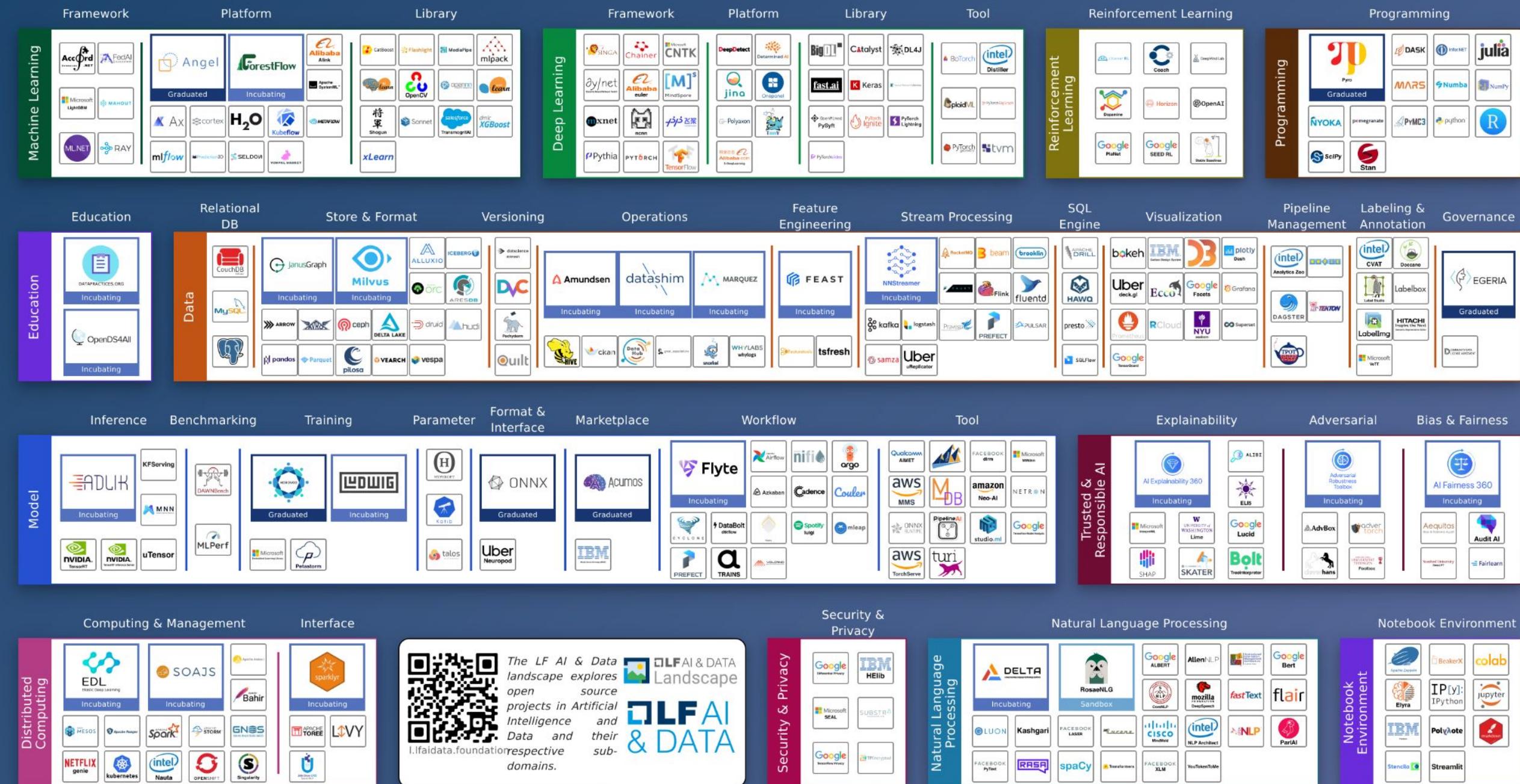
April 14, 2021

Ibrahim Haddad, Ph.D.
Executive Director, LF AI & Data
Ibrahim@LinuxFoundation.org



Ecosystem Overview

 DLF AI & DATA



Open Source offers *unique* benefits to AI & Data

Fairness

Methods to detect and mitigate bias in datasets and models, including bias against known protected populations

Robustness

Methods to detect alterations/tampering with datasets and models, including alterations from known adversarial attacks

Explainability

Methods to enhance understandability and interpretability by persona/roles in process of AI model outcomes/decision recommendations, incl ranking and debating results and decision options

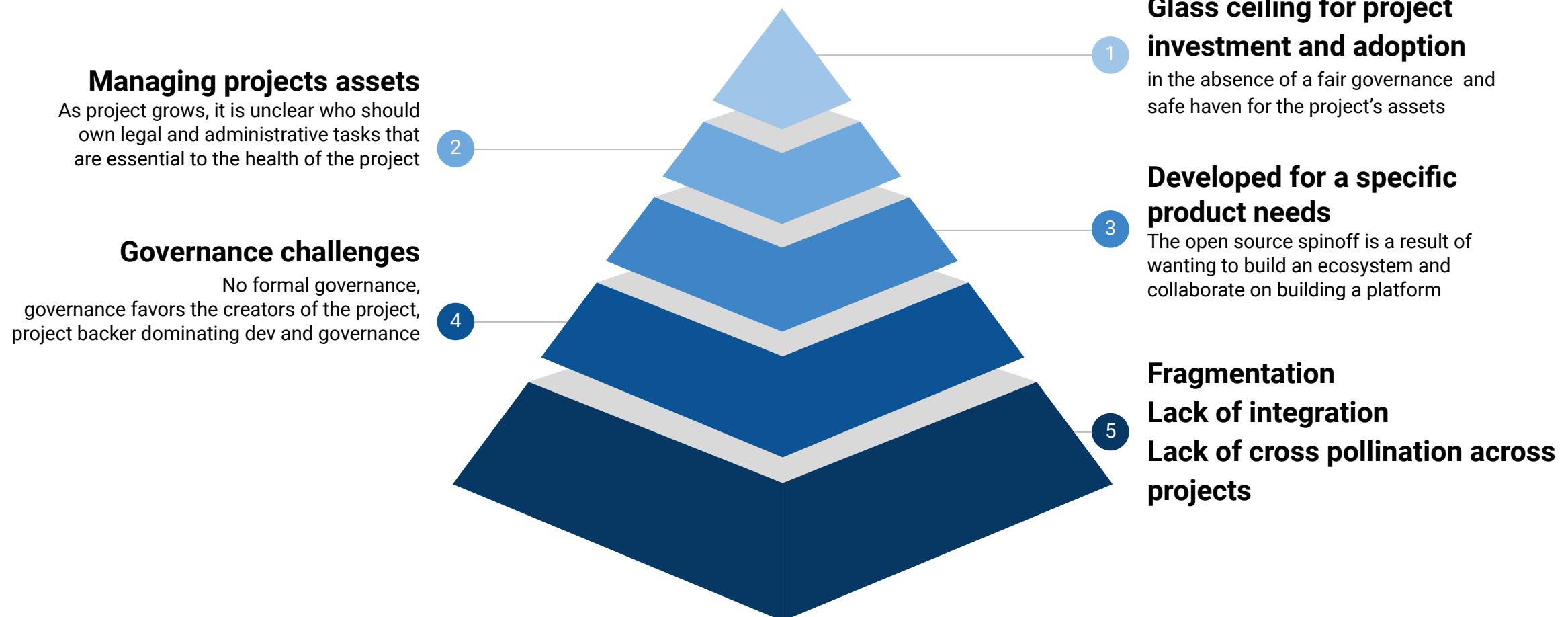
Lineage

Methods to ensure provenance of datasets and AI models, including reproducibility of generated datasets and AI models

Open Data

Methods to clean, sort, tagg and track provenance and a governance structure for doing these things

Open Source AI & Data - Ecosystem Challenges



Enter LF AI & Data

 **LF** AI & DATA

What is LF AI & Data?

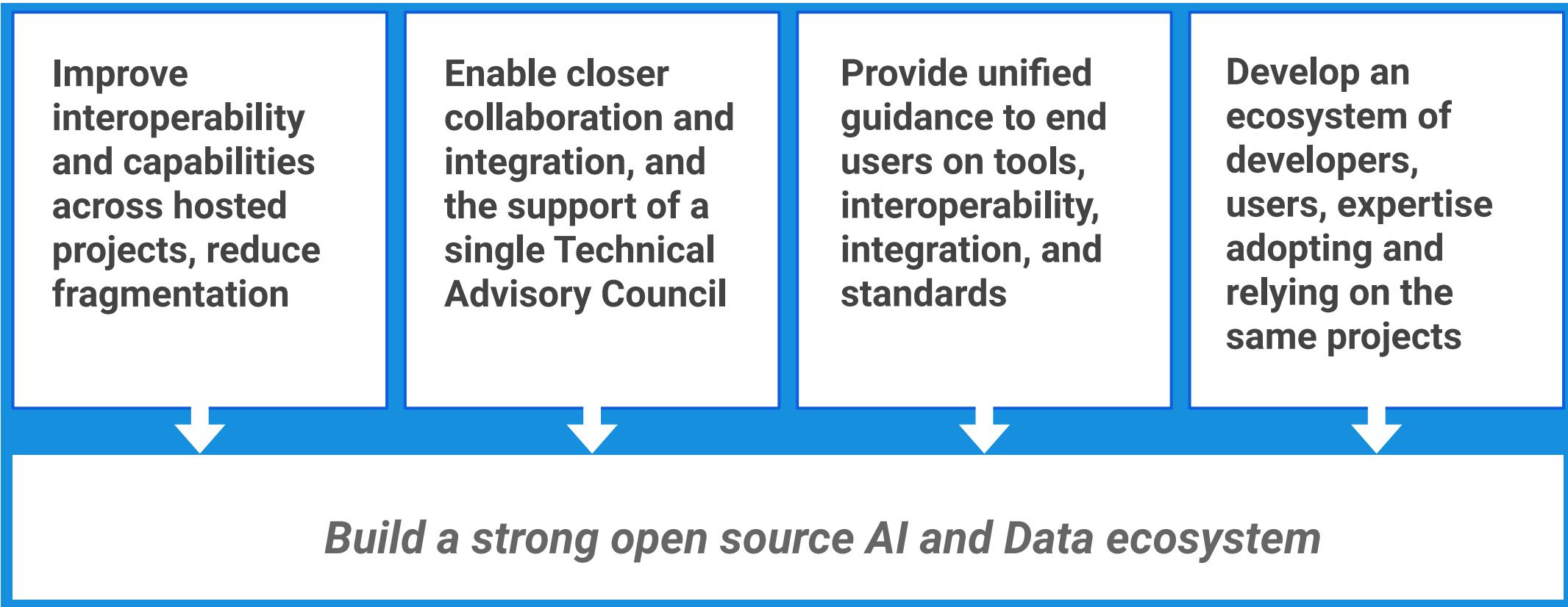
- › Non-profit
- › Open source collaborative effort focused on AI & Data
- › Hosted at the Linux Foundation
- › Focused on accelerating development innovation and a faster trajectory to adoption and growth

LF AI & Data

Collaborating to support an open and growing ecosystem of open source AI, data and analytics projects, by accelerating innovation, enabling collaboration and the creation of new opportunities for all the members of the community

<https://lfaidata.foundation>

Goals of LF AI & Data



LF AI & Data Growth

- › Launched in 03/2018 with 9 members and 1 hosted project
- › +3 projects in 2018
- › +5 projects in 2019
- › +15 projects in 2020
- › +4 projects (Jan - Apr) 2021
- › Currently supported by 46 member companies
- › >2000 subscribers to our main mailing list

Very active and growing developer community

Jan 1- Dec 31, 2020

8.92K

Contributors

217K

Contributions

100.03K

Commits

358

Repositories

6.93K

Emails

34.65K

PRs/Changesets

2.08K

Slack messages

22.54K

Total issues

Active and growing developer community

Cumm. Jan 1- Dec 31, 2020 vs. Jan 1, 2020 to Apr 2, 2021

8.92K

Contributors

34.65K

PRs/Changesets

100.03K

Commits

22.54K

Total issues

358

Repositories

2.08K

Slack messages

+7.25%

+7.21%

+11.03%

+5.19%

+3.07%

+68.75%

9.61K

Contributors

37.15K

PRs/Changesets

111.07K

Commits

23.71K

Total issues

369

Repositories

3.51K

Slack messages

Structure and Governance

Foundation Governance

Funding effort to support hosted technical projects



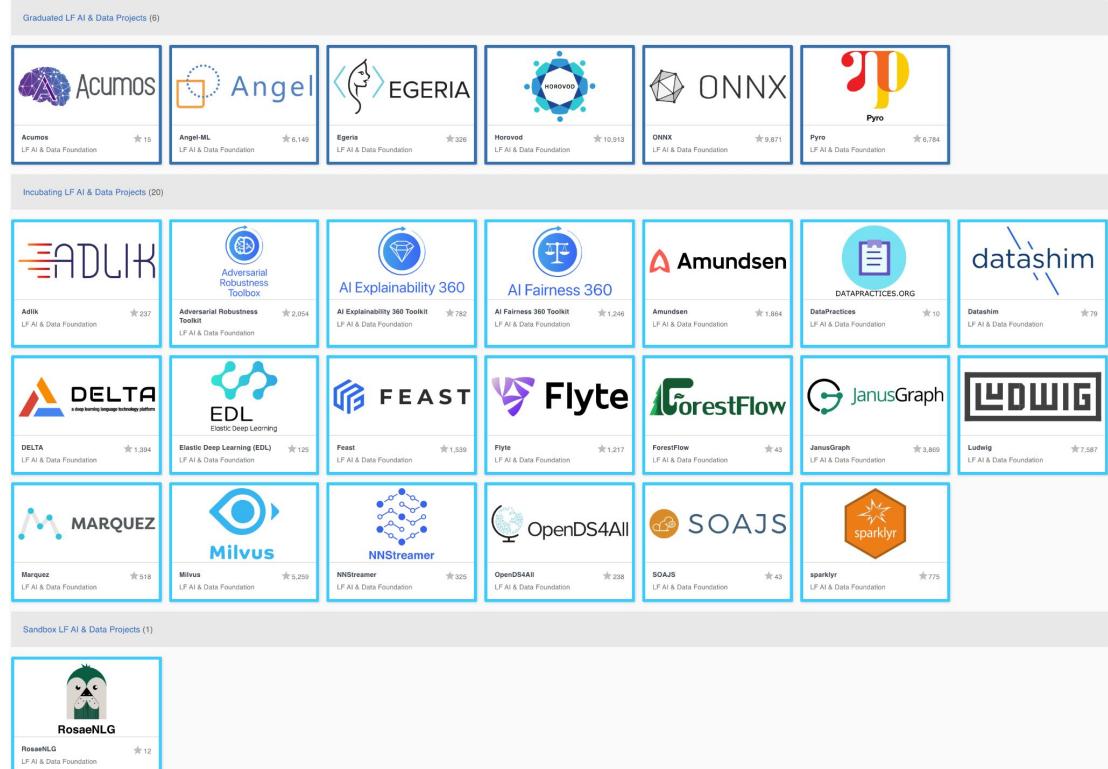
Technical Coordination

Technical Advisory Council



Hosted Projects

Projects have their own independent technical governance



Members

<https://landscape.lfai.foundation/card-mode?project=company>



Projects (27)

<https://landscape.lfai.foundation/card-mode?project=company>

Graduated LF AI & Data Projects (6)

**Acumos**
Acumos
LF AI & Data Foundation ★ 15

**Angel**
Angel-ML
LF AI & Data Foundation ★ 6,149

**EGERIA**
Egeria
LF AI & Data Foundation ★ 326

**HOROVOD**
Horovod
LF AI & Data Foundation ★ 10,913

**ONNX**
ONNX
LF AI & Data Foundation ★ 9,871

**Pyro**
Pyro
LF AI & Data Foundation ★ 6,784

Incubating LF AI & Data Projects (20)

**ADLIK**
Adlik
LF AI & Data Foundation ★ 237

**Adversarial Robustness Toolbox**
Adversarial Robustness Toolkit
LF AI & Data Foundation ★ 2,054

**AI Explainability 360**
AI Explainability 360 Toolkit
LF AI & Data Foundation ★ 782

**AI Fairness 360 Toolkit**
AI Fairness 360 Toolkit
LF AI & Data Foundation ★ 1,246

**Amundsen**
Amundsen
LF AI & Data Foundation ★ 1,864

**DATAPRACTICES.ORG**
DataPractices
LF AI & Data Foundation ★ 10

**datashim**
datashim
LF AI & Data Foundation ★ 79

**DELTA**
DELTA
LF AI & Data Foundation ★ 1,394

**EDL**
Elastic Deep Learning (EDL)
LF AI & Data Foundation ★ 125

**FEAST**
Feast
LF AI & Data Foundation ★ 1,539

**Flyte**
Flyte
LF AI & Data Foundation ★ 1,217

**ForestFlow**
ForestFlow
LF AI & Data Foundation ★ 43

**JanusGraph**
JanusGraph
LF AI & Data Foundation ★ 3,869

**LUDWIG**
Ludwig
LF AI & Data Foundation ★ 7,587

**MARQUEZ**
Marquez
LF AI & Data Foundation ★ 518

**Milvus**
Milvus
LF AI & Data Foundation ★ 5,259

**NNStreamer**
NNStreamer
LF AI & Data Foundation ★ 325

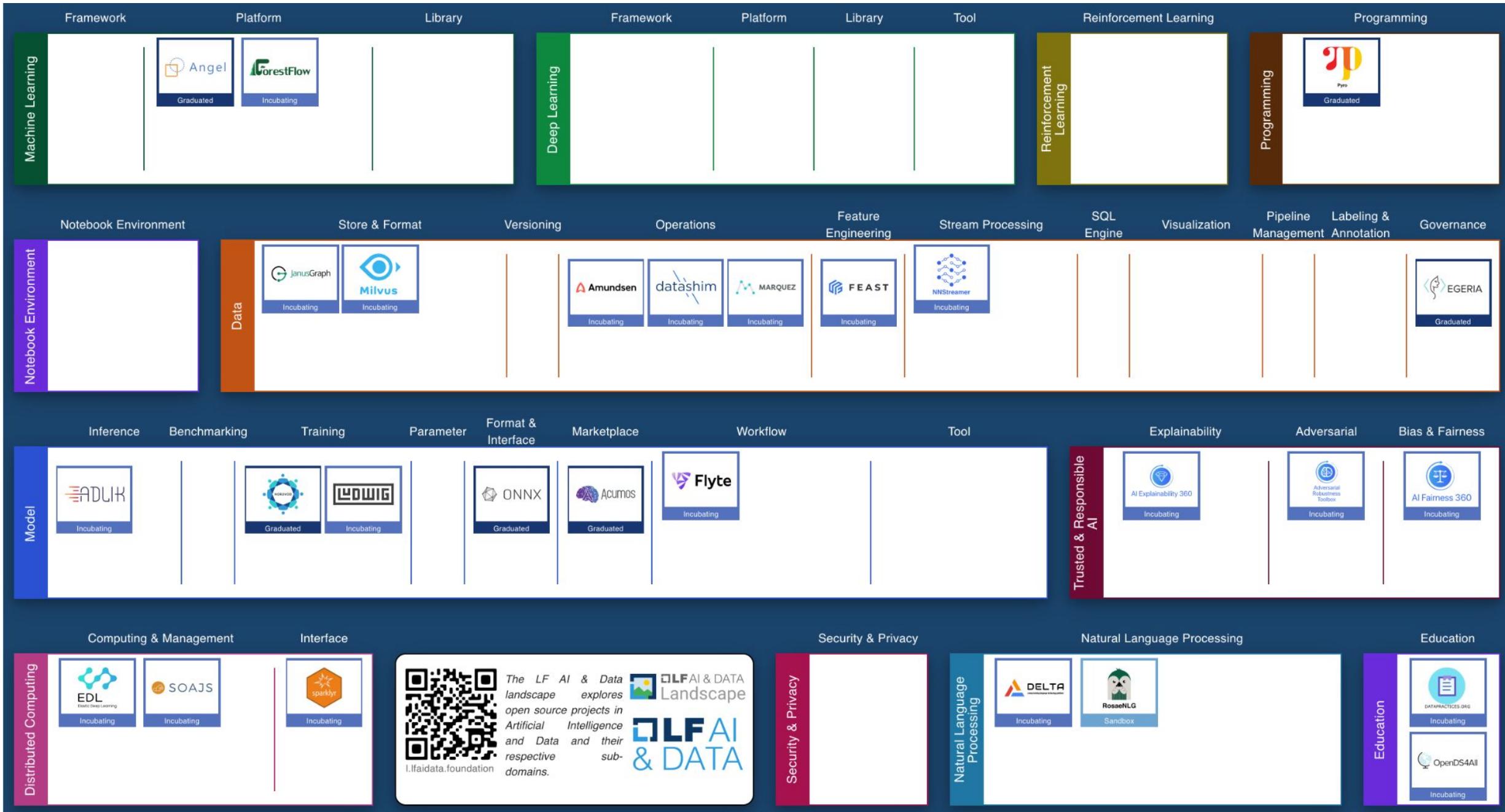
**OpenDS4All**
OpenDS4All
LF AI & Data Foundation ★ 238

**SOAJS**
SOAJS
LF AI & Data Foundation ★ 43

**sparklyr**
sparklyr
LF AI & Data Foundation ★ 775

Sandbox LF AI & Data Projects (1)

**RosaenLG**
RosaenLG
LF AI & Data Foundation ★ 12





Acumos AI

Open source framework to build, share and deploy AI applications

Acumos is an open source platform, which supports design, integration and deployment of AI models. Furthermore, it offers an AI marketplace that empowers data scientists to publish adaptive AI models, while shielding them from the need to custom develop fully integrated solutions.

[Learn More](#)



Adlik

Open source toolkit for accelerating deep learning inference

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 environments.

[Learn More](#)



AI Explainability 360

Open source toolkit that can help users better understand the ways that machine learning models predict labels

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.

[Learn More](#)



AI Fairness 360

Open source toolkit that can help users understand and mitigate bias in machine learning models throughout the AI application lifecycle

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.

[Learn More](#)



Adversarial Robustness Toolbox

Adversarial Robustness Toolbox

Open source tools to evaluate, defend, certify and verify Machine Learning models and applications against adversarial threats

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.

[Learn More](#)



Amundsen

Amundsen is a data discovery and metadata engine for improving the productivity of data analysts, data scientists and engineers when interacting with data

Amundsen is a data discovery and metadata engine for improving the productivity of data analysts, data scientists and engineers when interacting with data.

[Learn More](#)



Angel ML

Open source high-performance distributed machine learning platform

Angel is a high-performance distributed machine learning platform. 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 model.

[Learn More](#)



DataPractices

An effort to increase awareness and data literacy in the data ecosystem

DataPractices is a "Manifesto for Data Practices," comprised of values and principles to illustrate the most effective, modern, and ethical approach to data teamwork.

[Learn More](#)



Datashim

Open source enablement and acceleration of data access for Kubernetes/Openshift workloads in a transparent and declarative way

Datashim is enabling and accelerating data access for Kubernetes/Openshift workloads in a transparent and declarative way. Opensourced since September of 2019 and is growing to support use-cases related to data access in AI projects.

[Learn More](#)



Delta

DELTA is a deep learning based end-to-end natural language and speech processing platform

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.

[Learn More](#)



Elastic Deep Learning

Open source deep learning framework to build cluster cloud services

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.

[Learn More](#)



Egeria

The open standard that simplifies sharing and exchanging metadata.

Egeria is the world's first open source metadata standard. It provides open APIs, event formats, types and integration logic so organizations can share data management and governance across the entire enterprise without reformatting or restricting the data to a single format, platform, or vendor product.

[Learn More](#)



FEAST

Open source feature store for machine 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.
- Ensures model performance by tracking, validating, and monitoring features in production.

[Learn More](#)



Flyte

Open source acceleration for machine learning and data workflows to production

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.

[Learn More](#)



ForestFlow

An open source scalable policy-based cloud-native machine learning model server

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.

[Learn More](#)



Horovod

Open source distributed training framework for TensorFlow, Keras and PyTorch

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.

[Learn More](#)



JanusGraph

Distributed, open source, massively scalable graph database

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.

[Learn More](#)



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

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.

[Learn More](#)



Marquez

Open source metadata service for the collection, aggregation, and visualization of a data ecosystem's metadata

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.

[Learn More](#)



Milvus

Open source similarity search engine for massive-scale feature vectors

Milvus is built with heterogeneous computing architecture for the best cost efficiency. Milvus can be used in a wide variety of scenarios to boost AI application development.

[Learn More](#)



NNStreamer

Gstreamer plugins supporting ease and efficiency with neural network models and pipelines

NNStreamer 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.

[Learn More](#)



ONNX

Open source 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.

[Learn More](#)



OpenDS4All

Enables the creation of educational Data Science programs.

OpenDS4All is an open source project built to accelerate the creation of data science curricula at academic institutions.

[Learn More](#)



Pyro

Open source universal probabilistic programming language

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.

[Learn More](#)



SOAJS

Open source microservices and API management platform

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.

[Learn More](#)



Sparklyr

Open source and modern interface to scale data science and machine learning workflows using Apache Spark™, R, and a rich extension ecosystem

sparklyr is an open-source and modern interface to scale data science and machine learning workflows using Apache Spark™, R, and a rich extension ecosystem. It enables using Apache Spark with ease using R by providing access to core functionality like installing, connecting and managing Spark and using Spark's MLlib, Spark Structured Streaming and Spark Pipelines from R.

[Learn More](#)

Companies hosting projects in LF AI & Data

<https://landscape.lfai.foundation/format=hosting>



FACEBOOK

Herron Tech



Tech
Mahindra

Tencent 腾讯

Uber

wework

ZILLIZ
Reinvent Data Science

ZTE

Incubating Projects

- › TAC committee is responsible to voting in new incubation projects.
- › The TAC meets bi-weekly and generally booked 4-6 weeks in advance.
- › [Proposing projects for incubation](#)

What does it mean to be a Foundation project?

- › **A neutral home for the project** increases the willingness of developers from organizations to collaborate, contribute, and become committers
- › **Endorsement by members of the LF AI & Data Technical Advisory Committee** (TAC) is an independent signal of the quality of your project
- › **Open and neutral governance model**
- › **Program and project management services**
- › **Access to full-time staff** who are eager to assist your project
- › **Access to marketing and communications services** to support community and ecosystem engagement
- › **Access to a full range of events services** to help the project build and grow a community around it
- › **Legal support** ranging from hosting the trademark in a neutral organization to CLA and DCO system integration with GitHub, regular compliance scans, etc.
- › **Presence across multiple geographies** including China and Japan

Strong Marketing/PR/Community Effort

LF AI provides marketing services to support community and ecosystem engagement.

37 announcements in 2019

61 announcements in 2020

- › Events announcements and promotion
- › Project release
- › Project incubation
- › New members joining
- › Project cross collaboration
- › Invited posts

LF AI provides event and community ecosystem building services.



Joining LF AI & Data is Easy!



General Inquiries:

info@lfidata.foundation

 **LF** AI & DATA

Backup

Why incubate with us?

Building large, sustainable ecosystems requires collective resources



Events - We gathered over 45,000 attendees from over 12,000 organizations across 113 countries in 2019.



Legal - We manage IP for the world's most important tech and have one of the world's top open source legal teams in house.



Training - We trained millions of students (free and paid online), online skills certification, and on-site e-learning.



Certification - We designed and executed software and hardware testing and certification programs.



Developer Marketing - We have the largest share of voice of any open source foundation and a proven method to build large scale developer programs.



Developer Operations - We host the infrastructure that develops the world's largest software communities and provide release management, IT ops, and support.



Application Security - our projects are regularly audited and pen tested. We offer bug bounties, dependency analysis, and code scanning.

Why is LF AI & Data important?

1 OPEN & NEUTRAL

Provide an neutral environment and an open governance that fosters collaboration and cross pollination across projects and communities

2 HARMONIZATION & INTEROPERABILITY

Increase efforts to harmonize AI & Data open source projects and reduce fragmentation; increase interoperability among technical projects

3 TRUSTED & RESPONSIBLE AI

Become the home of open source trusted and responsible AI, both at the level of technical projects and also at the level of principles and guidelines

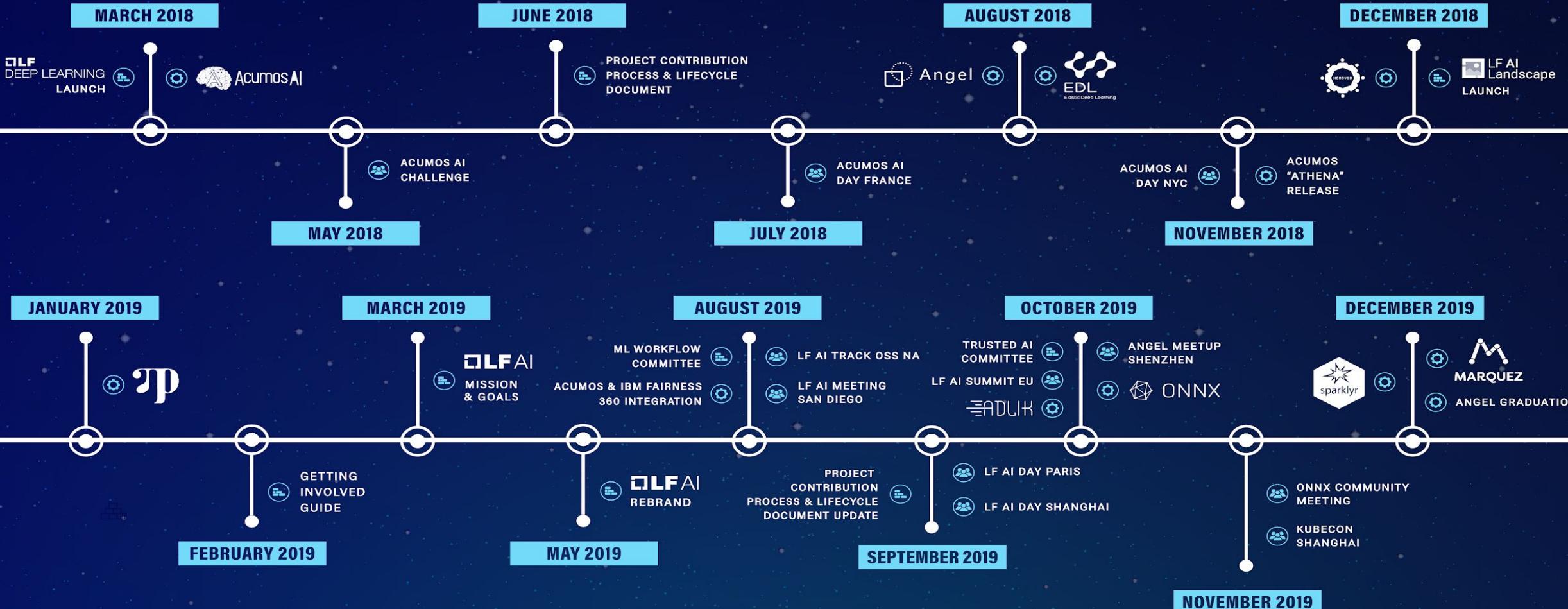
4 ECOSYSTEM DEVELOPMENT

Fund key projects; Increase awareness via a number of marketing and outreach activities; Training courses published in edX

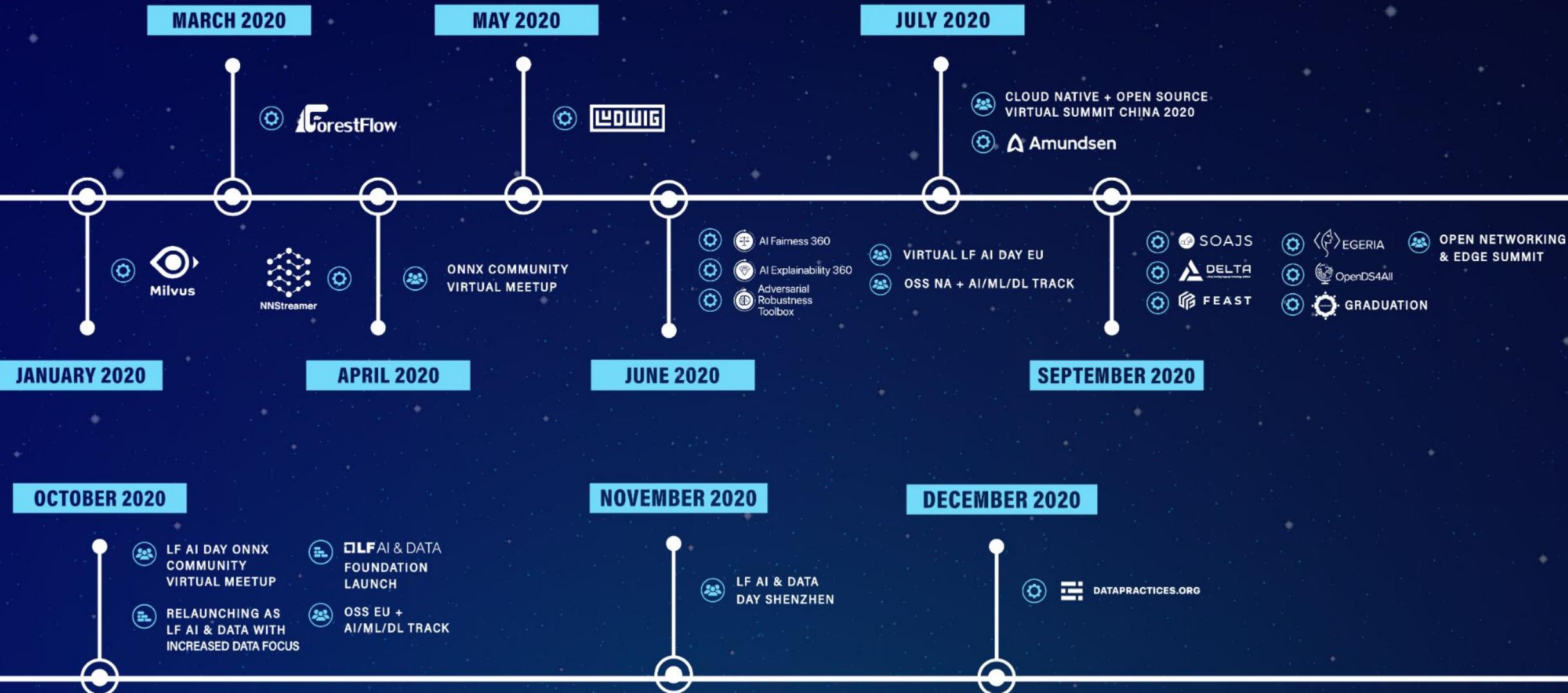
5 DATA

Develop the only vendor neutral, open source standards to enable best practices for data onec governance, connectivity, business intelligence, and analytics.

LFAI TIMELINE 2018-2019



LF AI & DATA TIMELINE 2020



LF AI & DATA TIMELINE 2021

JANUARY 2021



FEBRUARY 2021



MARCH 2021

