THE LINUX FOUNDATION | Research

Artificial Intelligence and Data in Open Source

March 2022

Ibrahim Haddad, Ph.D.
Executive Director, *LF AI & Data Foundation*

With a foreword by Dr. Seth Dobrin, VP Data and AI, Chief Data Officer Cloud and Cognitive Software, *IBM*

In partnership with:

LF AI & DATA

Contents

Foreword

Over the course of the last decade, the landscape of AI and data technologies has begun to rely heavily on the open source community. We have seen a shift both in enterprise adoption of open source software as well as an increased reliance on open source software in the base of proprietary software solutions that provide value-added capabilities on top of the open source distribution. These value-added capabilities can range from the creation and management of 'enterprise grade' distributions that began, in the early 20-teens, with Python, R, and Hadoop leading the charge. The primary governing foundations for these distributions are the Apache Foundation and the Linux Foundation as these governing bodies provide a governance and license structure that is more conducive to enterprise adoption.

The Linux Foundation and its AI and Data umbrella foundation (LF AI & Data) are critical to our (IBM's) strategy as we rely heavily on the consumption of LF AI & Data projects, the contribution to LF AI & Data projects, and the founding of new AI and data projects and committees hosted in LF AI & Data. This is exemplified in our standing as one of the leading contributors to LF AI & Data and a leading contributor to the Linux Foundation. As one of the major AI and data vendors, we see a responsibility to make certain technologies available freely to the world in AI as it relates to the robustness of AI, the fairness of AI, and the explainability of AI.

Additionally, as certain capabilities become less of a differentiator, but still remain vital to our and other entities, we leverage the open source community and specifically LF AI & Data to continue to create new value and scale by open sourcing them. On top of that, we have replicated the open source concept internally to IBM by building new and existing capabilities as 'Inner Source' which is essentially proprietary code that is managed internally to IBM as if it were open source software. Specific to AI, we have an inner source effort we call Watson Core which forms a scalable AI foundation for all AI capabilities across the company. IBM is not the only company building an inner source model for application development. If you adopt inner source practices as an extension of open source, the impact is even greater than what is laid out in this paper.

If you are reading this paper, you likely have an interest in the open source model. Many enterprises resist incorporating open source into their organizations because of various concerns. This paper provides data to help counter most arguments. Additionally, if you are a consumer of open source software, I encourage you to be a good citizen and participate fully in the community by committing back to these projects. This makes the communities more vibrant and helps with the retention of your employees.

Dr. Seth Dobrin, *VP Data and AI, Chief Data Officer Cloud and Cognitive Software, IBM*

CHALLENGES

Qualified open source talent is still in short supply. **Participation in open source projects provides visibility into talented developers and is a great recruitment tool.**

CHALLENGES

Organizations and governments perpetually need **additional hosting infrastructures, computational power, smarter algorithms and advanced tools** to manipulate, sort, tag, label, etc., significantly large data sets.

CHALLENGES

Nurturing trust in AI-enabled products and services is a challenge. **Open source methodology, transparency and accountability enables the development of trustworthy AI systems and processes.**

CHALLENGES

Ensuring data privacy, security, and governance can be challenging, **exacerbated by varying legislation across countries and geographies.**

CHALLENGES

With the ascent of machine learning and its reliance on high quality data, **there is a need for data agreements which facilitate data sharing and create a predictable path for training ML models.**

CHALLENGES

Implementing and verifying trusted and responsible AI systems and processes **is critical for any AI-enabled system.**

OPPORTUNITIES

Trust and responsibility **should be core principles of AI.**

OPPORTUNITIES

The LF AI & Data Trusted AI committee focuses on policies, guidelines, and the development of technical projects to **ensure the implementation of fair and trustworthy AI systems.**

OPPORTUNITIES

Edge computing is enabling a paradigm that moves AI and ML to where the data generation and computation takes place: edge devices. **Marrying AI with edge computing enables enhanced performance and real-time decision making.**

OPPORTUNITIES

Embedding AI in chip design is one of the most significant market opportunities in hardware today, **along with specialized chips to implement unique requirements for powering the metaverse.**

OPPORTUNITIES

Both academia and industry are innovating to usher in a **new era of smarter, faster, and most efficient algorithms.**

OPPORTUNITIES

CDLA licenses enable wider sharing and usage of open data, **particularly to bring clarity to the use of open data for artificial intelligence and machine learning models.**

Abstract

Over the past two decades, companies have adopted open source software (OSS) across multiple industries and technology verticals. This phenomenal enterprise adoption of OSS has increased (1) the use of open source in products and services, (2) levels of contributions to existing projects, (3) the creation of projects fostering collaboration, and (4) the development of new technologies. How did we get here? Simply put, more of us realized that collaborating on common enabling technologies was the fastest path to better and more cost-effective software solutions than any organization could deliver on its own.

Today, more leading-edge software development occurs inside open source communities than ever before. Proprietary projects have increasing difficulty keeping up with the rapid pace of development that open source achieves.

Artificial intelligence (AI) is no different from any other technology domain; OSS dominates. In this ecosystem, we can identify over 300 critical open source projects offering over 500 million lines of code, contributed by over 35,000 developers who work side by side to advance the state of technology in an open, collaborative, and transparent way. The characteristics of the open source model make it ideal for cooperating on enabling technologies regardless of domain or industry.

This paper reviews critical challenges in the open source AI ecosystem, discusses common characteristics across AI and data projects, and presents the role of the LF AI & Data Foundation in empowering innovators and accelerating open source development.

The Drive to Open Source Leadership

The availability of enterprise-grade open source software (OSS) is changing how organizations develop, maintain, and deliver products. A transparent development community plus access to public source code enables organizations to think differently about procuring, implementing, testing, deploying, and maintaining software. Using and adopting OSS can offer many benefits, including reduced development costs, faster product development, higher code quality standards, and more.

The open source methodology offers key and unique benefits to the domains of AI and data, specifically in areas of fairness, robustness, explainability, lineage, availability of data, and governance **(FIGURE 1).**

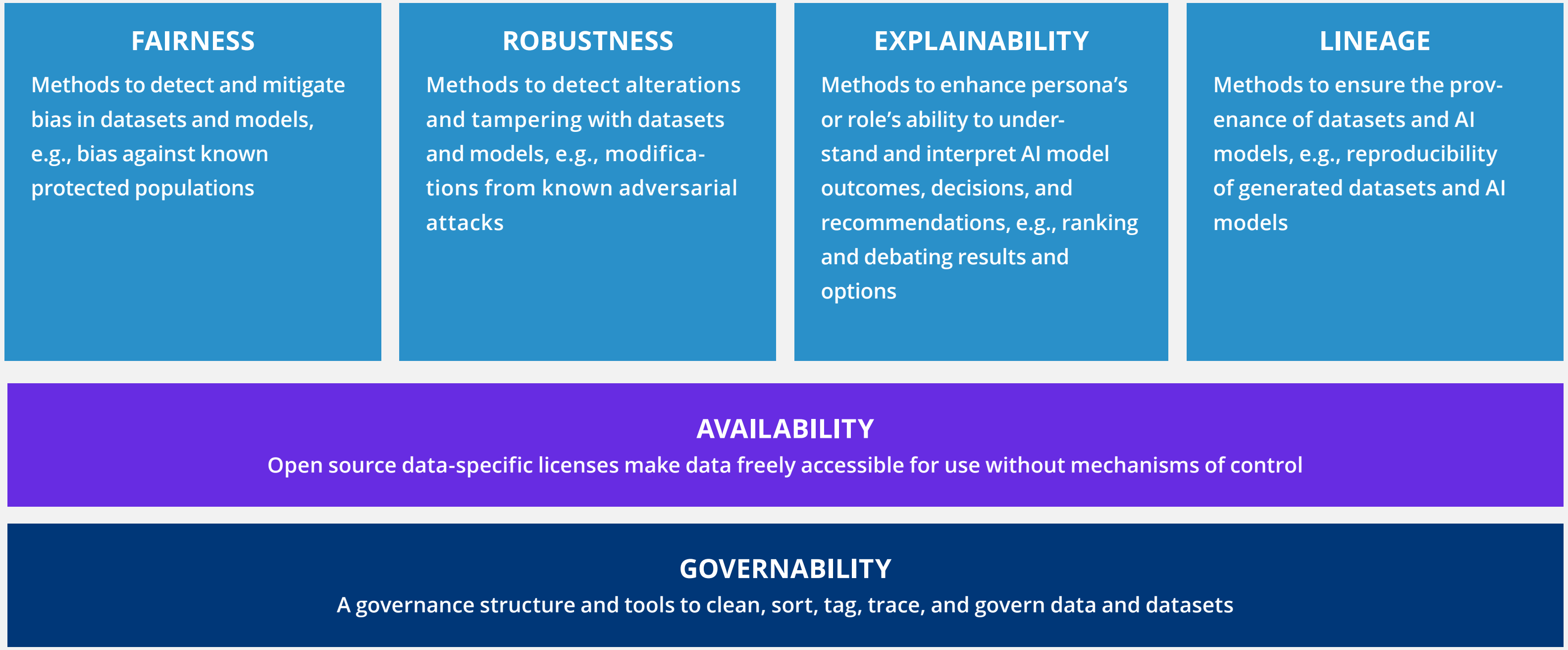
Academia’s Role in AI Research & Development

Many AI-related open source projects, platforms, frameworks, and libraries started as academic R&D at different universities. Decades of government and taxpayers’ support in the AI and data domain, tireless professors and students, and open source as a tool for collaborating with other academics on the implementation side all advanced the field significantly. Hallmarks of this success were:

- A collaborative approach to innovation
- A disciplined process of creating and validating new ideas
- Licenses applied to the source code resulting from R&D

FIGURE 1

Open source areas of benefits exclusive to AI and data



FAIRNESS

Methods to detect and mitigate bias in datasets and models, e.g., bias against known protected populations

ROBUSTNESS

Methods to detect alterations and tampering with datasets and models, e.g., modifications from known adversarial attacks

EXPLAINABILITY

Methods to enhance persona’s or role’s ability to understand and interpret AI model outcomes, decisions, and recommendations, e.g., ranking and debating results and options

LINEAGE

Methods to ensure the provenance of datasets and AI models, e.g., reproducibility of generated datasets and AI models

AVAILABILITY

Open source data-specific licenses make data freely accessible for use without mechanisms of control

GOVERNABILITY

A governance structure and tools to clean, sort, tag, trace, and govern data and datasets

- Community building in different domains, all with passion and practice
- Discussion of results at conferences with external review and feedback

> Academic and enterprise efforts culminated under the open source umbrella. Academia continues to be a laboratory for new ideas. On GitHub, students across the globe have developed hundreds of open source AI and data projects.

In 2015, enterprise interest in AI began to grow fast and aligned with academia, with reference implementations of many of the ideas in today's open source projects. The open source environment fostered collaboration between academia and industry without restrictions on that type of relationship. Companies invested incredible resources in the space and improved on the general academic approach and practices in such areas as:

- Finalizing the software produced
- Building architecture (plugins, application programming interfaces)
- Accelerating the launch of initiatives
- Cultivating a developer ecosystem around these projects
- Gathering contributions for projects from other companies (typically business partners)
- Providing access to large datasets

Academic and enterprise efforts culminated under the open source umbrella. Academia continues to be a laboratory for new ideas. On GitHub, students across the globe have developed hundreds of open source AI and data projects. Some of these projects will gain traction and eventually find their place on the landscape, become widely adopted, and attract a large developer community around them. Project members must experiment in the ongoing stream of new ideas and encourage new participants to learn the open source model and develop specific domain expertise.

The OS Model for Ongoing AI Development

The AI and data space is changing rapidly. The sweet spot is the intersection of academia, enterprise, and the open source community. In early 2018, the Linux Foundation established LF AI & Data to facilitate a vendor-neutral environment in which members could advance the open source AI and data platforms and empower generations of open source innovators. Within 3.5 years, LF AI & Data grew to host 36 projects representing a little over 11% of the overall key projects in the ecosystem.

The Linux Foundation provides a neutral, trusted hub for developers to code, manage, and scale open source technology projects. Organizations go through four primary stages on their journey with OSS (**FIGURE 2**):

1. ***Consumption:*** No one wants to reinvent the wheel for enabling technologies. Most organizations start using and incorporating OSS into their commercial products and services. They comply with and continuously release internal AI and data efforts under open source licenses. They leverage the power of collaboration, benefit from the multiplier effect of open source, and provide faster, more agile development and faster time to market.
2. ***Participation:*** Organizations expand their OSS strategies to participate in projects, community events, public outreach, and they encourage their developers to work on OSS projects critical

to their operations. As they engage more actively with communities, they increase their visibility and attract the talent they need.

3. *Contribution:* Organizations increase their efforts, contributing code to and even hosting the projects they rely on and selectively growing the communities that support their internal efforts. They originate and host strategic open source projects to maximize value and minimize their technical debt.
4. *Leadership:* Through their prior strategies, organizations have earned the trust of their open source communities. They help other organizations to navigate project politics. They also begin capitalizing on emerging trends in technology and establishing leadership positions in projects critical to their ecosystem's ongoing success.

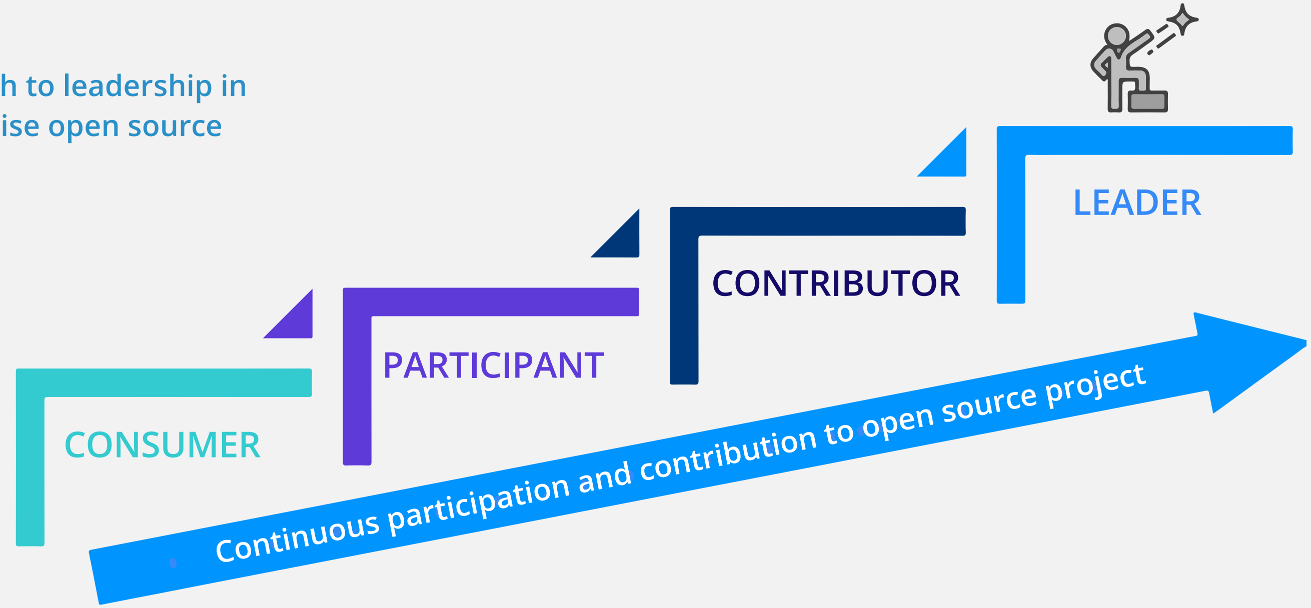
In each stage, an organization must scale its open source operation and activities to move into the next stage.

In the early stages, engineering teams typically drive open source consumption by using open source components based on their technical merits. As open source speeds up development, these teams participate selectively in critical projects, joining the conversation or contributing code. As their efforts gain traction, higher levels of the organization awaken to the merits of OSS consumption and participation, and a determined business strategy begins driving such involvement. Some organizations achieve their goals at the consumption stage and stay there. Others work to evolve, ever-improving their open source practices and pushing higher to attain specific leadership positions within the open source communities they deem critical for their products and services.

Most of these organizations are technology leaders; they have access to resources many of us can only dream of. Some have hosted multiple OSS projects. They have realized the value of mass collaboration under open source principles and methodologies. They trust LF AI & Data

FIGURE 2

The path to leadership in enterprise open source


LEADER
CONTRIBUTOR
PARTICIPANT
CONSUMER
Continuous participation and contribution to open source project

to incubate their open projects and support them with an open, fair, and transparent governance model and a set of developer-oriented resources, tools, and services. The result is a vibrant developer community, an expanding user base, active collaboration, and integration with other projects in the ecosystem, demonstrated by their experience in hosting projects.

> Organizations are infusing AI in products and services across all industries. Companies benefit from a community of other contributors helping accelerate open AI applied research. These concepts gear toward solving industry-wide challenges; no single company can address them alone.

OSS has lowered the barrier to entry to AI development. OSS libraries, frameworks, platforms, and tools make AI accessible to everyone. Organizations are infusing AI in products and services across all industries. Companies benefit from a community of other contributors helping accelerate open AI applied research. These concepts gear toward solving industry-wide challenges; no single company can address them alone. The open source methodology is the most appropriate way to work on these challenges in an open, transparent, and inclusive manner. The community develops solutions that everyone will adopt, adjust, and accommodate their specific situations and use cases. The following section explores the ecosystem of open source AI and data and shares our findings on projects, founders, locations, and licenses.

Open Source AI and Data Ecosystem

Our goal is to help developers, end-users, and others to navigate the complex ecosystem landscape. Anyone can contribute by submitting a pull request on GitHub. To that end, the LF AI & Data Foundation has created an interactive landscape (**FIGURE 3**)[1]. It represents

- 316 open source AI and data projects
- 500 million+ lines of code (LoC) and growing at an average weekly rate of one million LoC
- 35,000+ active developers contributing code
- 15 open source licenses
- 18 countries of project origin
- 102 companies founded projects
- 10 open source foundations hosting multiple projects
- 12 universities founded projects

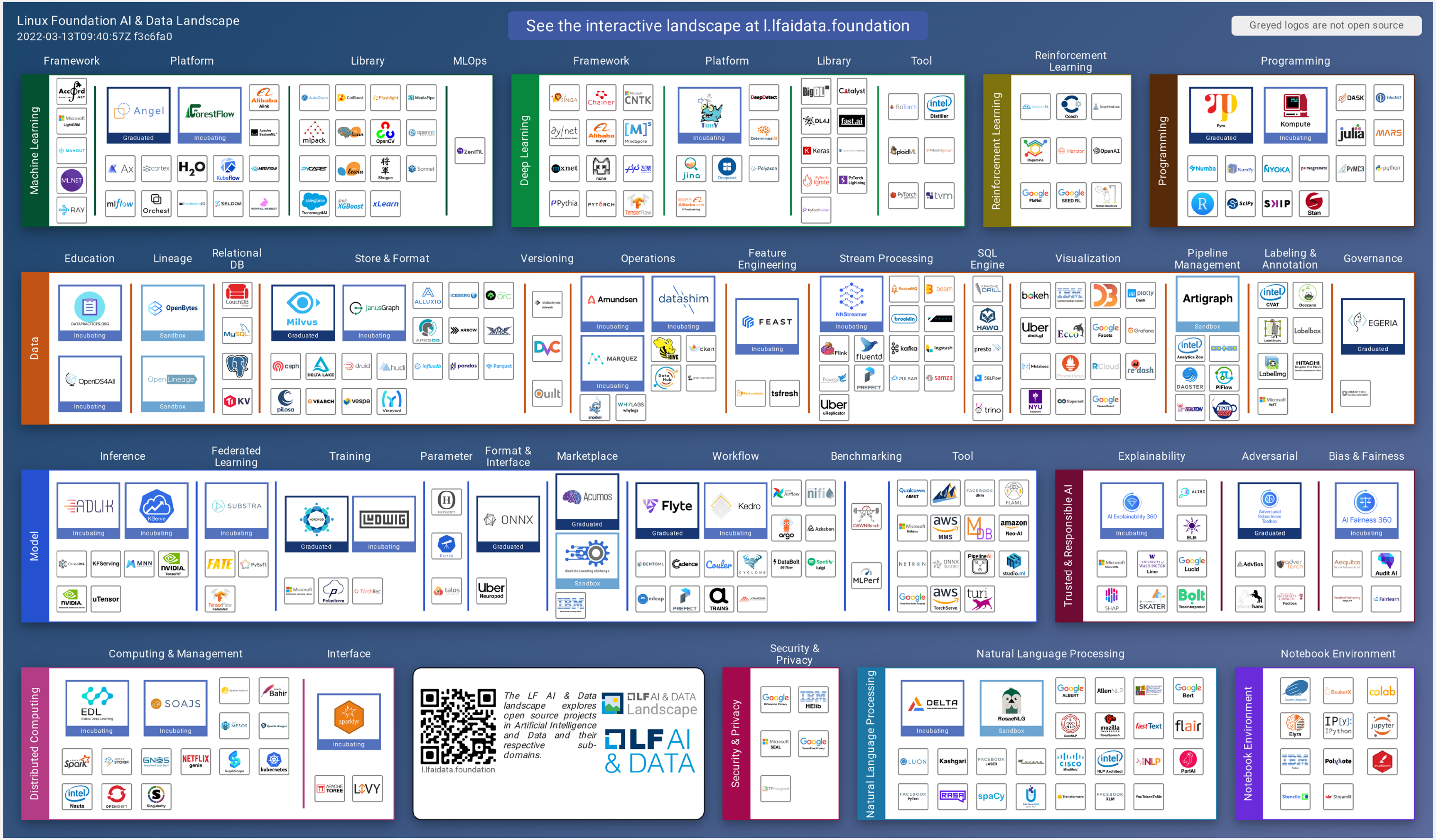
In the landscape, we organized the 316 open source projects into 11 categories:

- Machine learning (four subcategories
- Deep learning (four subcategories)
- Reinforcement learning
- Programming
- Data (13 subcategories)
- Model (eight subcategories)
- Trusted and responsible AI (three subcategories)
- Distributed computing (two subcategories)
- Security and privacy
- Natural language processing
- Notebook environment

[1] Find the interactive version at https://landscape.lfai.foundation.

FIGURE 3

Open Source AI & Data landscape, as of 3 Feb. 2022.



Each category includes several subcategories as we build out the taxonomy of the specific domain. For instance, the machine learning category consists of four subcategories: framework, platform, library, and machine learning ops (MLOps).

This growing portfolio of technical projects has led to an exponential growth in active developers across all projects. As of February 2022, 36 hosted projects have over 15,000 unique contributors involved in code development in terms of commits, pull requests, changesets, and bug reporting and resolving (**FIGURE 4**)[2].

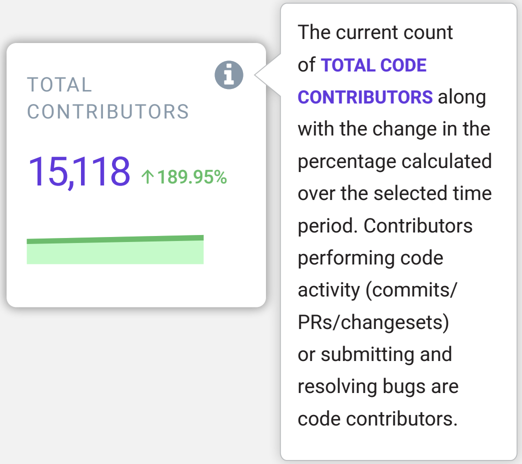
FIGURE 5 showcases growth in terms of the total number of unique commits from April 2019 to January 2022. There has been a growth of 199.17% in the total commits during the last 3 years. Active code contributors contributed over 117,320 commits to LF AI & Data hosted projects during this period. Furthermore, The commits by new contributors have increased by 309.07% during the last 3 years, a very positive trend. It's important to note that new contributors are defined as those who did their first code activity (commits/PRs/changesets) or submitted their first bug or resolved their first bug during the selected time period.

FIGURE 6 shows the distribution of license types across the projects featured on the landscape. Over 65% of projects have adopted the Apache 2.0 license, 16.5% the MIT License, just under 10% the BSD 3-Clause license, and 5% one of the GNU family of licenses.

[2] For more information, see LFX Insights, https://insights.lfx.linuxfoundation.org.
To view additional stats, visit https://insights.lfx.linuxfoundation.org/projects/lfai-f/dashboard.

FIGURE 4

Over 15,000 contributors to LF AI & Data hosted projects



TOTAL CONTRIBUTORS

15,118 ↑189.95%

The current count of **TOTAL CODE CONTRIBUTORS** along with the change in the percentage calculated over the selected time period. Contributors performing code activity (commits/PRs/changesets) or submitting and resolving bugs are code contributors.

FIGURE 5

Commit growth across LF AI & Data hosted projects (Apr. 2019–Mar. 2022)

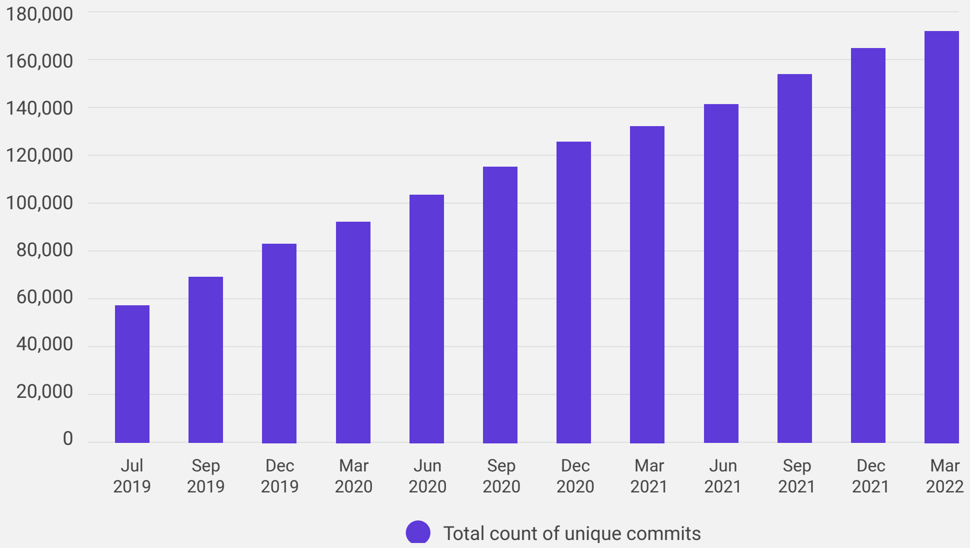
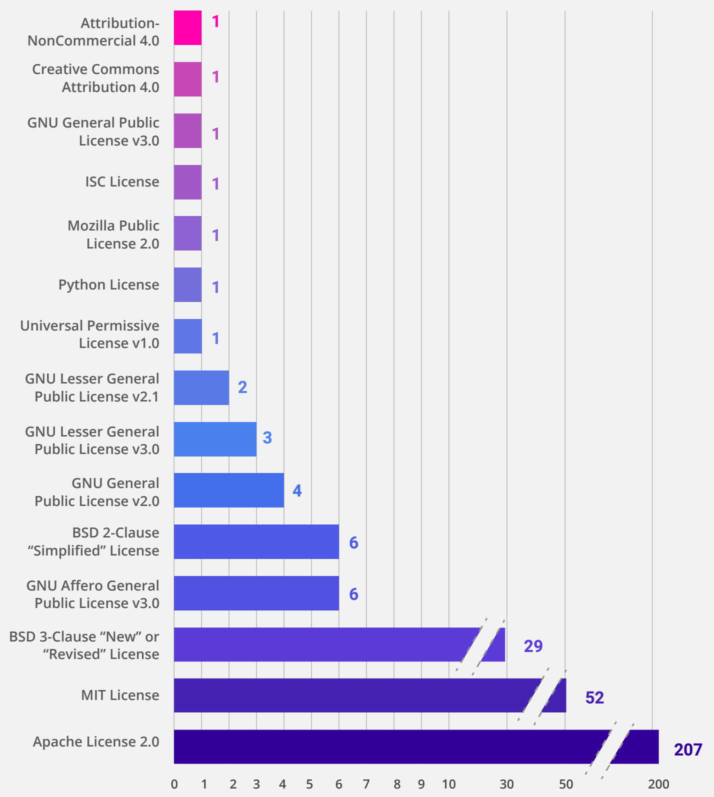


FIGURE 7 shows where projects originated. The United States has launched the most open source projects in the AI and data space, followed by China, Germany, Canada, and the United Kingdom.

Leadership in the open source AI and data ecosystem is a race among companies. **FIGURE 8** presents a cross-section of companies that founded OS AI and data projects. Several organizations such as Uber and IBM launched numerous open projects and then donated them to a hosting open source foundation. For instance, IBM contributed four projects, and Uber contributed three projects to the LF AI & Data Foundation, which hosts them.

FIGURE 6

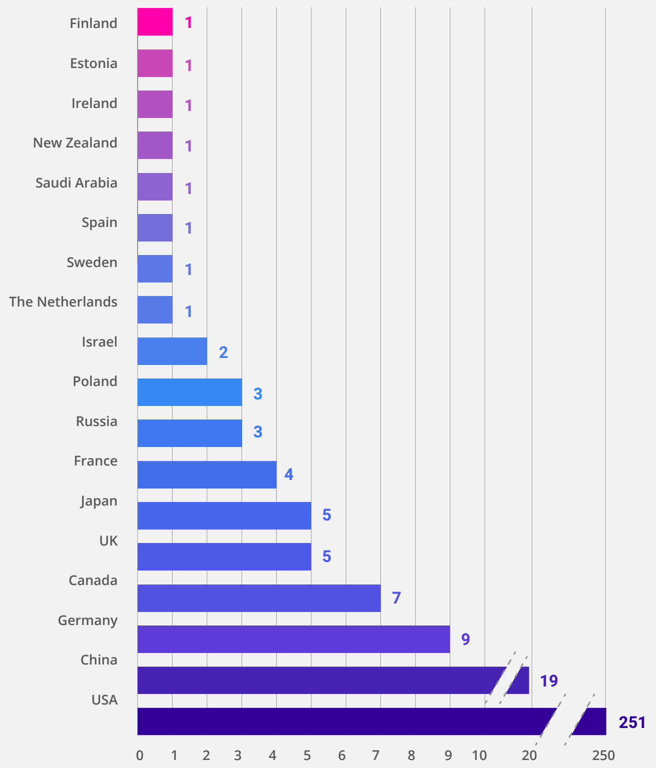
Distribution of licenses across all projects on the landscape



The United States has launched the most open source projects in the AI and data space, followed by China, Germany, Canada, and the United Kingdom.

FIGURE 7

Country of origin of open source projects on the landscape



As of February 2022, 27 organizations were hosting 36 projects in LF AI & Data **(FIGURE 9)**[3].

The following list captures the 12 universities who founded open source AI and data projects featured on the landscape:

- Carnegie Mellon University
- Georgia Institute of Technology
- New York University
- Stanford University
- Technical University of Dortmund
- University of California, Berkeley
- University of Chicago
- University of Pennsylvania
- University of Tuebingen
- University of Washington
- University of Waterloo
- INRIA (French Institute for Research in Computer Science and Automation)

[3] See online hosting list here: https://landscape.lfai.foundation/hosting.

FIGURE 8

Founders of OS projects in AI and data

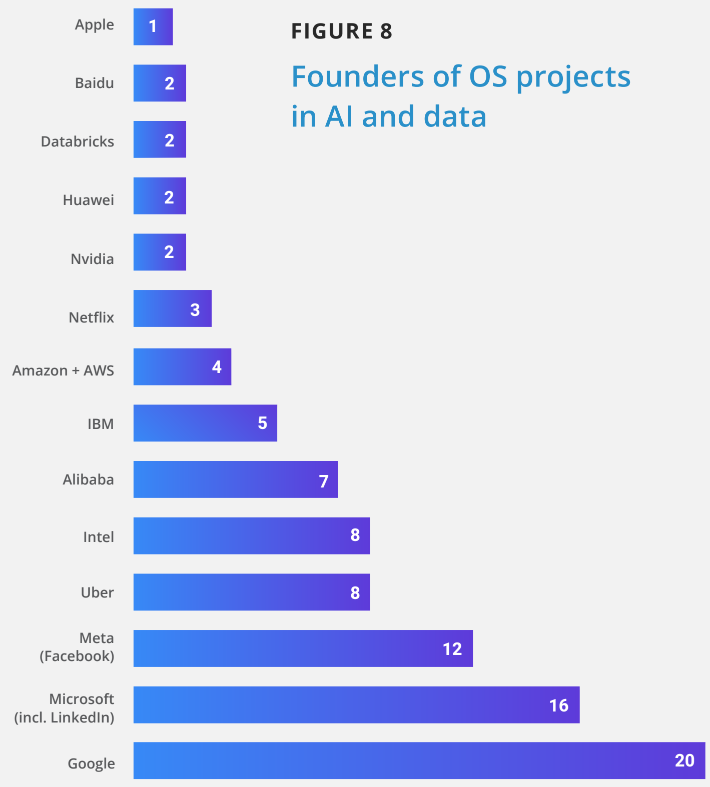
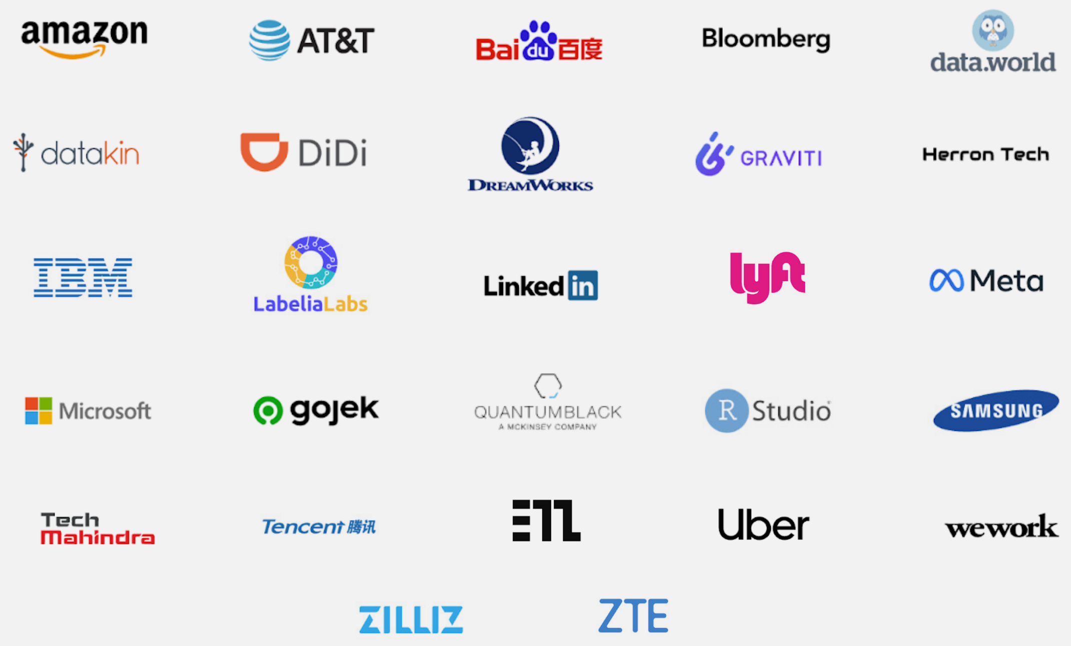


FIGURE 9

Companies and organizations hosting technical projects in LF AI & Data



At LF AI & Data, projects move through four stages, the last of which is archival, when two thirds of a community vote to archive a project.[4] **FIGURE 10** presents the currently hosted projects in the first three stages:

- ***Sandbox***, for projects that extend other LF AI & Data projects with functionality or interoperability or fit the LF AI & Data mission and offer potentially novel approaches to existing solutions
- ***Incubation***, for projects that have achieved and maintain a core infrastructure initiative and have documented best practices and transparent communications
- ***Graduation***, for projects that have a steady, substantial flow of commits and code contributions from a diversity of organizations and integration with other hosted projects

LF AI & Data provides a neutral, trusted home for developers to collaborate on open AI and data technology projects.

[4] Please review the LF AI & Data Project Life Cycle Document to learn about each stage, how to move from one stage to another and associated benefits.

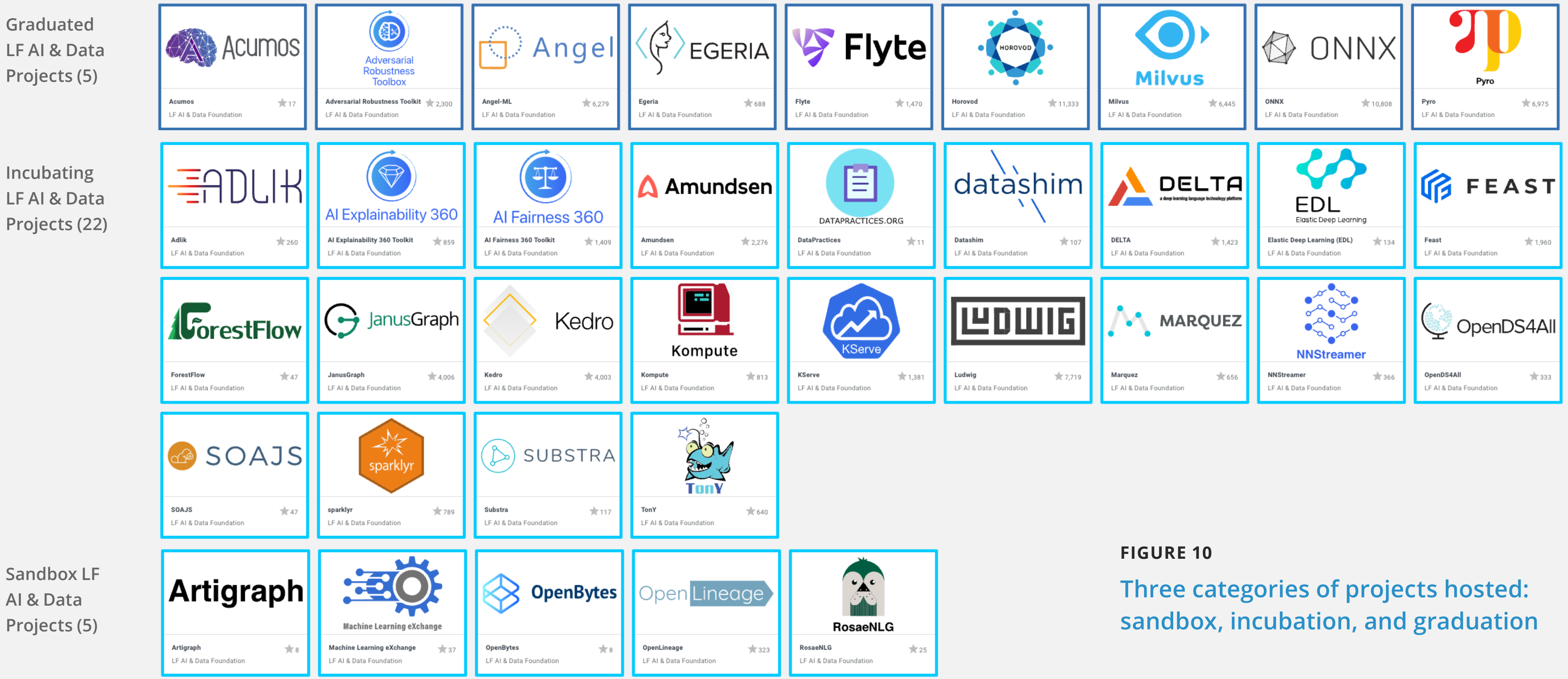


FIGURE 10

Three categories of projects hosted: sandbox, incubation, and graduation

The foundation offers a strong portfolio of services to open source projects—not just code management and technical decisions but training and certification, events management, and marketing and legal services. **FIGURE 11** overviews the services LF AI & Data offers its hosted projects[5]. Through these shared services, we want to provide new projects a sandbox in which to incubate and to support projects in incubation and graduation.

[5] To discuss hosting opportunities, please email info@lfaidata.foundation.

FIGURE 11

Comprehensive list of services to LF AI & Data hosted projects

NEUTRAL HOSTING

A neutral home for an open source project increases the willingness of developers from software companies, startups, academia, and elsewhere to collaborate, contribute, and become committers.

DEDICATED STAFF

Projects have access to full-time staff (executive director, program manager, project coordinator) who cultivate the maturity and adoption of open source AI and data projects

TRAINING AND CERTIFICATION

We develop training classes and, through the Linux Foundation, can execute and launch certification programs in support of hosted projects.

EVENTS MANAGEMENT

Events are part of LF AI & Data's core strategy to help projects build a community and accelerate knowledge-sharing and integration. Many LF AI & Data projects have their own events.

DEV-FOCUSED OPERATION

Services include IT infrastructure, release management, IT ops, support, security audits, and a host of tools (FOSSA, LastPass, Slack, Synk, Zoom, etc.).

MENTORSHIP

Members of the LF AI & Data technical advisory committee and leaders of graduated projects are available to support and mentor new projects.

LEGAL SERVICES

We help projects navigate licensing requirements, IP regimes, trademark management, compliance scans, export control filings, and developer certificate of origin or contributor license agreement integration with GitHub, etc.

MARKET SERVICES

We offer a wide range of marketing services to increase project awareness, project adoption, and the number of contributors.

DESIGN AND AESTHETICS

Our in-house team provides graphic design resources for new logos, websites, and website refreshes or enhancements.

PROGRAM MANAGEMENT

We have decades of experience in program management of open source projects. We bring best practices to all LF AI & Data hosted projects.

LFX PLATFORM EXPERIENCE

This Linux Foundation product offers a set of integrated tools for project insights, security, easy contributor license agreements, crowdfunding, member engagement, and more.

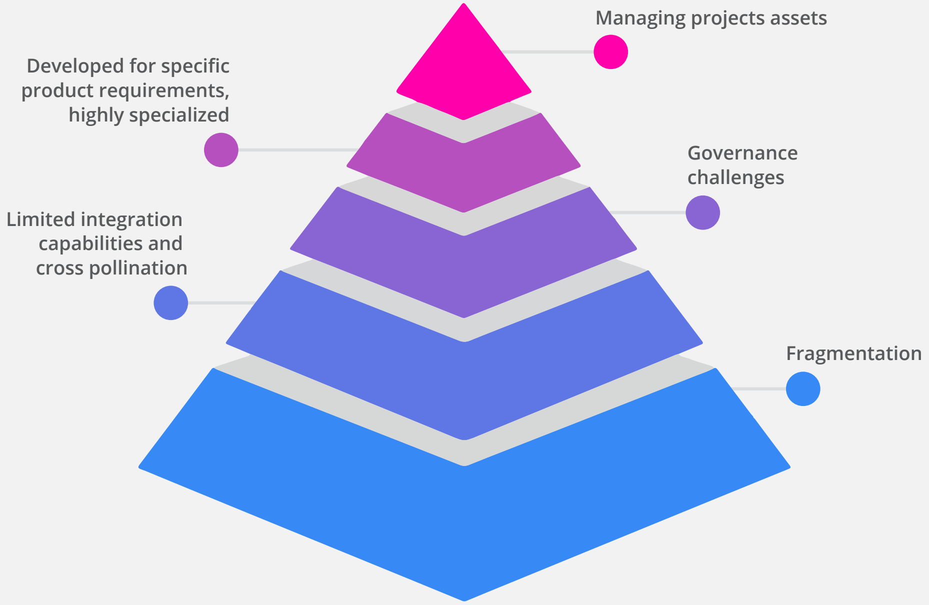
Challenges and Opportunities

At a high level, organizations and even governments face these ongoing challenges:

- Hiring qualified talent
- Perpetually needing additional computational power
- Nurturing trust in the AI-enabled products and services
- Overcoming bias
- Ensuring data privacy, security, and governance
- Complying with national policies and legislation

FIGURE 12

Common challenges of open source projects in the AI and data ecosystem


Managing projects assets
Developed for specific product requirements, highly specialized
Governance challenges
Limited integration capabilities and cross pollination
Fragmentation

These challenges also represent a set of new opportunities for new projects, new collaborations, and new potential achievements such as:

- Implementing and verifying trusted and responsible AI
- Deriving value and insights efficiently from the large sets of collected data
- Marrying AI with edge computing as the need for real-time decision-making pushes AI closer to the edge
- Specializing AI chip design to put AI directly on the silicon
- Demanding ever more efficient and more intelligent algorithms
- Adopting federated learning to protect privacy while training AI with sensitive user data

Common Ecosystem Challenges

Despite a very thriving ecosystem, several challenges span multiple areas (**FIGURE 12**). All these challenges inhibit project investment and adoption. Organizations avoid projects with governance challenges and integration limitations; instead, they participate in and adopt stable projects with an open and fair governance model. LF AI & Data addresses these ecosystem challenges and provides a neutral hosting environment that will help attract contributors and help projects grow their user base. In this section, we elaborate on these challenges.

FRAGMENTATION

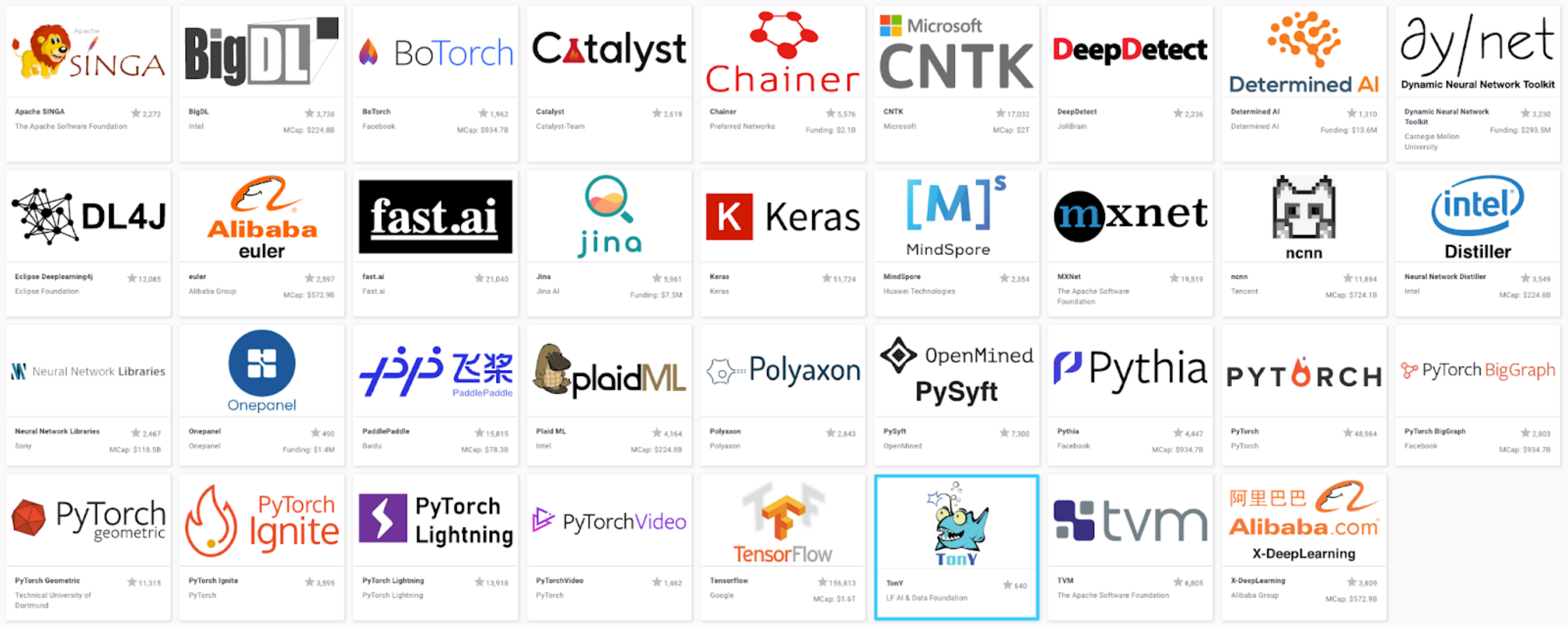
Many companies attempt to solve particular software problems internally before they decide to open source their efforts. Once they open these projects to an outside community of developers, we can see that these companies are trying to solve the same problems with similar sets of functionalities. For example, in the subdomain of deep learning (**FIGURE 13**), we see much overlap, with projects competing for developer and contributor attention and enterprise adoption. We see this same competition for talent in all the other subdomains we track. Eventually, energy consolidates around a few winning projects, and talent flocks to new open source projects, startups, and new initiatives. Even so, this fragmentation of effort early on concerns members of the open source AI and data ecosystem. LF AI & Data provides the structure needed to help members winnow projects faster, concentrate talent earlier, and attract more adopters.

INTEGRATION AND CROSS-POLLINATION

With so many competing projects in each category and subcategory, the available options for integration are vast. Therefore, any given project must prioritize the needs of its community and look deliberately to integrate with other projects. Parallel to that, most organizations engaged in open source AI follow an "AI first" approach to technology. They have open sourced some of their own AI and data-related projects, but they focus heads down on their projects and their collaborators. They contribute little if anything to other projects in the ecosystem. The result is limited cross-pollination of projects, with

FIGURE 13

Deep learning projects competing for developers and adoption



Apache SINGA ★ 2,272
The Apache Software Foundation

BigDL ★ 3,738
Intel
MCap: $224.8B

BoTorch ★ 1,962
Facebook
MCap: $934.7B

Catalyst ★ 2,619
Catalyst-Team

Chainer ★ 5,576
Preferred Networks
Funding: $2.1B

CNTK ★ 17,032
Microsoft
MCap: $2T

DeepDetect ★ 2,236
JoliBrain

Determined AI ★ 1,310
Determined AI
Funding: $13.6M

Dynamic Neural Network Toolkit ★ 3,230
Carnegie Mellon University
Funding: $293.5M

Eclipse Deeplearning4j ★ 12,085
Eclipse Foundation

euler ★ 2,597
Alibaba Group
MCap: $572.9B

fast.ai ★ 21,040
Fast.ai

Jina ★ 5,961
Jina AI
Funding: $7.5M

Keras ★ 51,724
Keras

MindSpore ★ 2,354
Huawei Technologies

MXNet ★ 19,519
The Apache Software Foundation

ncnn ★ 11,894
Tencent
MCap: $724.1B

Neural Network Distiller ★ 3,549
Intel
MCap: $224.8B

Neural Network Libraries ★ 2,467
Sony
MCap: $118.5B

Onepanel ★ 490
Onepanel
Funding: $1.4M

PaddlePaddle ★ 15,815
Baidu
MCap: $78.3B

Plaid ML ★ 4,164
Intel
MCap: $224.8B

Polyaxon ★ 2,843
Polyaxon

PySyft ★ 7,300
OpenMined

Pythia ★ 4,447
Facebook
MCap: $934.7B

PyTorch ★ 48,964
PyTorch

PyTorch BigGraph ★ 2,803
Facebook
MCap: $934.7B

PyTorch Geometric ★ 11,315
Technical University of Dortmund

PyTorch Ignite ★ 3,595
PyTorch

PyTorch Lightning ★ 13,918
PyTorch Lightning

PyTorchVideo ★ 1,462
PyTorch

Tensorflow ★ 156,813
Google
MCap: $1.6T

TonY ★ 640
LF AI & Data Foundation

TVM ★ 6,805
The Apache Software Foundation

X-DeepLearning ★ 3,809
Alibaba Group
MCap: $572.9B

limited integration capabilities within an extensive ecosystem of open source projects. LF AI & Data guides projects through their life cycles so that they seek out much-needed or even groundbreaking integrations.

GOVERNANCE

Governance is the stewarding of a project, its direction, and the conduct of its contributors. At the Linux Foundation, most open source communities have some governance guidelines, including procedures for addressing unacceptable behavior and for managing decisions, structures, and roadmap[6]. Without this basic structure, projects may devolve into mayhem.

This section excludes LF AI & Data hosted projects because they must document and publish their open governance on their website and GitHub. Many projects hosted outside LF AI & Data in the ecosystem wrestle with certain governance challenges of three types:

- ***No formal governance:*** Project founders dominate development activities, control and manage project assets, and own any decision making. A single large entity heavily influences some of the most popular projects, most likely because AI development requires a very narrow band of specialized knowledge. Most AI projects result from years of investment and talent acquisition. When a large entity wants to build an ecosystem and collaborate with others on a platform, it spins off these projects to the open source community.
- ***Some form of governance:*** This form typically favors the project's founders and guarantees them a majority vote. In many cases, the project's founders spin off the project to benefit from the network effect of open source in building an ecosystem, but they still want to maintain control of the project.
- ***Ad hoc governance:*** Project founders have poorly documented and loosely exercised their governance model if they have one at all.

Projects benefit greatly from the focused momentum of a single large host, but the lack of contributor diversity is an existential risk for most projects.

Governance challenges manifest in the unequal treatment of contributors, overly rigid control over a project's codebase, and conflict around successful projects' legal and administrative requirements, as we shall see. LF AI & Data provides an array of services, from mentorships to program management, that help project founders set up governance structures vital to each project's development.

Projects benefit greatly from the focused momentum of a single large host, but the lack of contributor diversity is an existential risk for most projects.

Developed internally to specific requirements

Because of the cost of human capital and the time required to develop such a complex technology, project founders often create AI systems (frameworks, platforms, libraries, tools, etc.) with a specific product or service in mind. During development, they may follow the source-available pattern of open source contribution. They develop plans and code internally and periodically publish them for the ecosystem to review and consume. External contributors may propose code, but project founders and hosts need not treat external contributors as equal to internal ones, regardless of the quality of their contributions.

[6] For more on governance, please see "Building Leadership in an Open Source Community." https://www.linuxfoundation.org/tools/building-leadership-in-an-open-source-community.

Few outsiders will continue contributing to the project if founders ignore their input, so the model's success depends on a fresh flow of talent.

Highly specialized, tightly controlled

In projects with highly specialized development, users of the code benefit greatly when their use cases overlap those of the project hosts because these projects provide them with consistent streams of product-ready code. However, if a user wants to adapt the host's code, but the host wants to maintain control over the project's direction and development road map, then the user has no means of broadening the project's scope. Such tight control over projects deters creative developers from becoming committers or maintainers needed to sustain the project over time. At LF AI & Data, we've found that deliberate open governance—that is, governance with a defined project structure, processes, and clear ways to broaden the project's scope and promote developers into committers and maintainers—usually mitigates this situation.

As with any technology where talent premiums are high, the network effects of open source are very strong.

Managing project value over time

As an open source project grows, its needs also grow. Indeed, it takes on a life of its own. Who will build and manage its website, pay for cloud-based testing, manage the trademark, file export controls, create and manage developer events and conferences, or perform license compliance scans? These tasks are time-consuming, overwhelming, stressful for project contributors, and sometimes disruptive to project participation. Talent may go elsewhere if no one steps up. Who should cover the costs of these legal and administrative tasks—which are essential to the project's health—and what should they get in return? How can we minimize such conflict and anxiety? Project founders, leaders, or key influencers can suggest hosting such a project in a neutral foundation such as LF AI & Data, which will absorb these costs and logistics so that project participants can focus on development, innovation, and integration.

> As with any technology where talent premiums are high, the network effects of open source are very strong.

HIRING TALENT

Before the strong coupling of AI research and academic R&D labs with enterprise, finding talent was a huge challenge. Today companies have more options than ever to find AI experience appropriate for their needs, but it's still tough. According to the Linux Foundation's "Open Source Jobs Report" of 2021, 92% of hiring managers reported difficulty finding sufficient talent with open source skills; 43% of hiring managers were looking for individuals with expertise architecting solutions based on OSS[7].

The current wave of AI hiring has pushed many AI researchers from academia to commercial and product R&D. In some cases, companies have sponsored academic AI labs or taken complete AI labs from various academic institutions into their employment, thereby infusing more money into academic R&D and nearly monopolizing these labs to benefit their own AI efforts. Neutral OSS foundations such as LF AI & Data with a rich array of AI and data projects and events naturally attract developer talent at diverse career stages.

[7] For details, see "The 2021 Open Source Jobs Report: 9th Annual Report on Critical Skills, Hiring Trends, and Education," Sept. 2021. https://www.linuxfoundation.org/tools/the-2021-open-source-jobs-report.

Opportunities

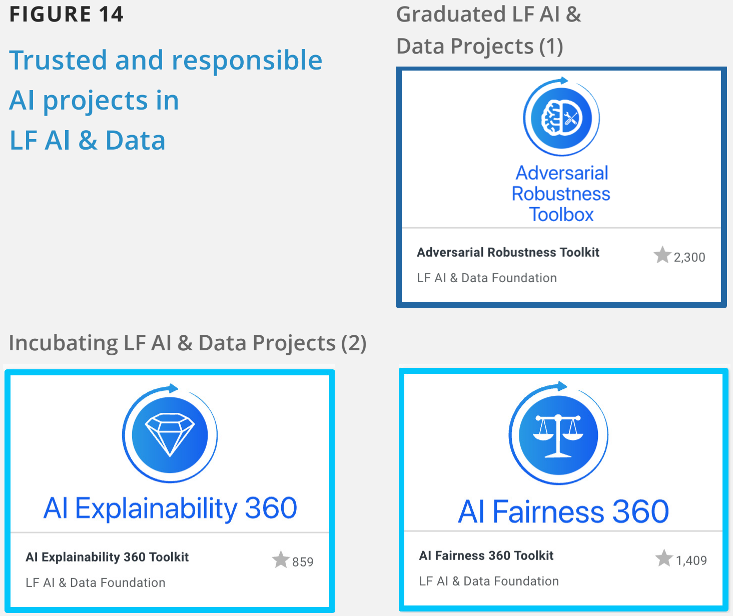
The open source AI and data ecosystem presents several opportunities for new R&D, new startups, and innovations.

LEVERAGING TRUSTED AND RESPONSIBLE AI

The infusion of AI in products and services has created opportunities to improve people's lives around the world. It also has raised concerns about the fairness, explainability, and security of these applications and systems. Various national and global initiatives are working to address these concerns. LF AI & Data and its member organizations consider trusted and responsible AI as a critical domain and as a global group working on policies, guidelines, and use cases to ensure the development of trustworthy AI systems and processes[8]. In addition, we are providing three software toolkits (**FIGURE 14**) to help achieve our goals from a technical perspective:

FIGURE 14

Trusted and responsible AI projects in LF AI & Data


Graduated LF AI & Data Projects (1)
Adversarial Robustness Toolbox
Adversarial Robustness Toolkit
2,300
LF AI & Data Foundation
Incubating LF AI & Data Projects (2)
AI Explainability 360
AI Explainability 360 Toolkit
859
LF AI & Data Foundation
AI Fairness 360
AI Fairness 360 Toolkit
1,409
LF AI & Data Foundation

The infusion of AI in products and services has created opportunities to improve people's lives around the world. It also has raised concerns about the fairness, explainability, and security of these applications and systems.

- *AI Fairness 360:* This extensible open source tool kit helps users to examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle[9].
- *AI Explainability 360:* This open-source library supports the interpretability and explainability of datasets and machine learning models[10].
- *Adversarial Robustness Toolbox (ART):* This open source tool helps developers and researchers to evaluate, defend, and verify machine learning models and applications against adversarial threats[11].

DERIVING INSIGHT AND VALUE FROM COLLECTED DATA

We're all familiar with the expression, "garbage in, garbage out," referring to the importance of inputting good data to derive valuable insights. With the global digitalization and transformation of industries and economies, data has become quite abundant; the challenge has shifted from finding data to selecting quality data, efficiently mining the data for actionable insights, and effectively converting those insights into business value.

[8] See online via the landscape interactive mode.
[9] For details, see https://ai-fairness-360.org/.
[10] For more information, see https://ai-explainability-360.org/
[11] For more on ART, see https://adversarial-robustness-toolbox.org/.

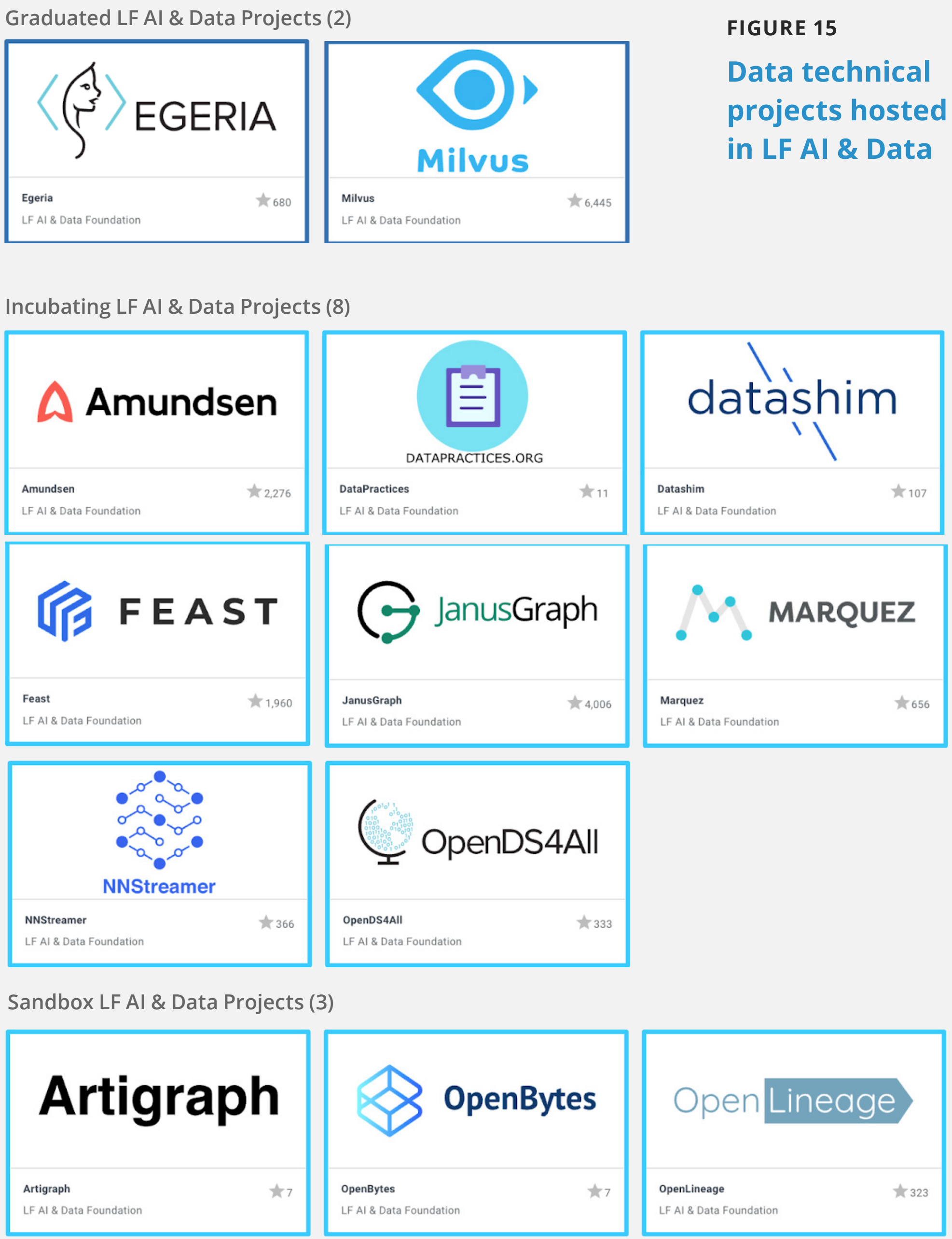


FIGURE 15

Data technical projects hosted in LF AI & Data

The LF AI & Data community recognizes the importance of data and has been keen on hosting and supporting key projects covering data lineage, format, store, operations, feature engineering, governance, stream processing, and pipeline management. **FIGURE 15** illustrates the technical data projects hosted in LF AI & Data[12].

DRIVING REAL-TIME DECISIONS

The need for real-time decision-making pushes AI closer to the edge. With edge AI, organizations have various untapped possibilities to enhance performance via decreased latency, improved real-time analytics, and increased scalability in terms of data processing.

INFUSING AI INTO HARDWARE

In the past few years, we have witnessed a surge in R&D in AI chip design, where innovators are looking to put AI directly on the silicon. Traditional chip designers and manufacturers and a slew of new entrants are racing to create a chip optimized to run machine learning workloads that will power the next generation of computing devices. Large organizations with deep pockets and Startups with promising ideas and appropriate venture capital see the AI chip as one of the most significant market opportunities in hardware today, along with chips to power the metaverse.

CREATING EVER-SMARTER ALGORITHMS

The need for more efficient and intelligent algorithms is an ongoing opportunity. Both academia and industry are innovating in the space, bringing new ideas to usher in a new era of smarter, faster, and most efficient algorithms.

[12] For the interactive version, visit https://landscape.lfaidata.foundation/card-mode?category=data&project=hosted

Closing Observations

The incubation model is very effective when appropriately executed. The Linux Foundation was a pioneer in establishing the incubation model and scaling it to a couple of dozen umbrella foundations currently hosting over 800 technical projects[13]. Many of the presently graduated LF AI & Data projects joined while in incubation, with only a couple of organizations contributing to the projects. With the support of the Foundation's portfolio of services, these projects grew to include hundreds of developers from dozens of contributing organizations and deployed commercially at a large scale.

Consolidation is bound to happen. We expect consolidation around multiple platforms, frameworks, and libraries that address the same challenges. This consolidation is already happening, with some projects slowing down as contributors join competing projects adopted by their sponsoring organizations. Unlike fragmentation scenarios, where there are winning and losing projects, we believe the net result will be a win-win as successful projects grab their share of contributors. Contributors to older projects will migrate to newer projects and repurpose their knowledge, experience, and skills to advance the development and drive the launch of startups.

The choice of license affects the project's growth. All projects listed on the landscape have adopted an open source license approved by the Open Source Initiative (OSI)[14]. Developers are familiar with these licenses, and in general, the end-users and adopters of the software are aware of the licenses and their obligations. Put differently, no successful projects use custom open source licenses. Adopting OSI-approved open source licenses keeps with industry best practices that strongly discourage license customization.

Open data licenses are beginning to commoditize training data. Licenses such as Community Data License Agreement (CDLA) provide a structured set of guidelines to enable open sharing of datasets under defined terms[15]. The availability of training data under these terms will help democratize the overall AI marketplace by lowering the barriers to entry when offering an AI-backed service. Proprietary datasets will continue to exist, but data availability under the CDLA licenses (two versions exist) should allow everyone to build credible products, including smaller players.

We are faster and more innovative together. We live in a very exciting time! Open source has already won in AI and data. We must now focus on how we work with this new model of creating, licensing, distributing software, and collaborating with others. We are far more innovative in collaboration than in isolation. Evident by the data available to us today, open source as a methodology and practice has fueled our massive advances in AI. We're going now through the process of open source AI dominating the software world. This situation is the new normal. Let's celebrate it and continue our pursuit of technological advances in fair, transparent, and ethical ways.

[13] For details, see https://www.linuxfoundation.org/projects.

[14] To be listed on the landscape, a project must use an OSI-approved license. For more about the OSI, see https://opensource.org.

[15] For more on CDLA, see https://cdla.dev.

LF AI & Data References

Website	lfaidata.foundation	Mail Lists	lists.lfaidata.foundation
Wiki	wiki.lfaidata.foundation	Slack	slack.lfaidata.foundation
GitHub	github.com/lfaidata	Artwork	artwork.lfaidata.foundation
Landscape	landscape.lfaidata.foundation	Events	lfaidata.foundation/events/

About the Author

Dr. Ibrahim Haddad is Vice President of Strategic Programs at the Linux Foundation, where he's focused on facilitating a vendor-neutral environment for advancing the open source platform and empowering generations of open source innovators. Haddad leads the LF AI & Data Foundation providing a trusted hub for developers to code, manage, and scale open source AI and data projects. His work, and the work of the Foundation as a whole, support companies, developers, and the open source community in identifying and contributing to the technical projects that address industry challenges for the benefit of all participants. Before the Linux Foundation, he served as Vice President of R&D and Head of the Open Source Division at Samsung Electronics. Throughout his career, Haddad held technology and portfolio management roles at Ericsson Research, the Open Source Development Labs, Motorola, Palm, Hewlett-Packard, and the Linux Foundation. He graduated with Honors from Concordia University (Montréal, Canada) with a Ph.D. in Computer Science.

Disclaimer

This report is provided "as is." The Linux Foundation and its authors, contributors, and sponsors expressly disclaim any warranties (express, implied, or otherwise), including implied warranties of merchantability, non-infringement, fitness for a particular purpose, or title, related to this report. In no event will the Linux Foundation and its authors, contributors, and sponsors be liable to any other party for lost profits or any form of indirect, special, incidental, or consequential damages of any character from any causes of action of any kind with respect to this report, whether based on breach of contract, tort (including negligence), or otherwise, and whether or not they have been advised of the possibility of such damage. Sponsorship of the creation of this report does not constitute an endorsement of its findings by any of its sponsors.

twitter.com/linuxfoundation

facebook.com/TheLinuxFoundation

linkedin.com/company/the-linux-foundation

youtube.com/user/TheLinuxFoundation

LF AI & DATA

Part of the Linux Foundation, LF AI & Data supports open source innovation in artificial intelligence, machine learning, deep learning, and data. LF AI & Data was established to support a sustainable open source AI ecosystem that makes it easy to create AI and Data products and services using open source technologies. We foster collaboration under a neutral environment with an open governance model to support the harmonization and acceleration of open source technical projects.

THE LINUX FOUNDATION | Research

Founded in 2021, Linux Foundation Research explores the growing scale of open source collaboration, providing insight into emerging technology trends, best practices, and the global impact of open source projects. Through leveraging project databases and networks, and a commitment to best practices in quantitative and qualitative methodologies, Linux Foundation Research is creating the go-to library for open source insights for the benefit of organizations the world over.

To reference the work, please cite as follows: Ibrahim Haddad, “Artificial Intelligence and Data in Open Source: Challenges and Opportunities for Mass Collaboration at Scale,” foreword by Dr. Seth Dobrin, VP Data and AI, Chief Data Officer Cloud and Cognitive Software, IBM, March, 2022.