LF AI & Data

Ibrahim Haddad, PhD

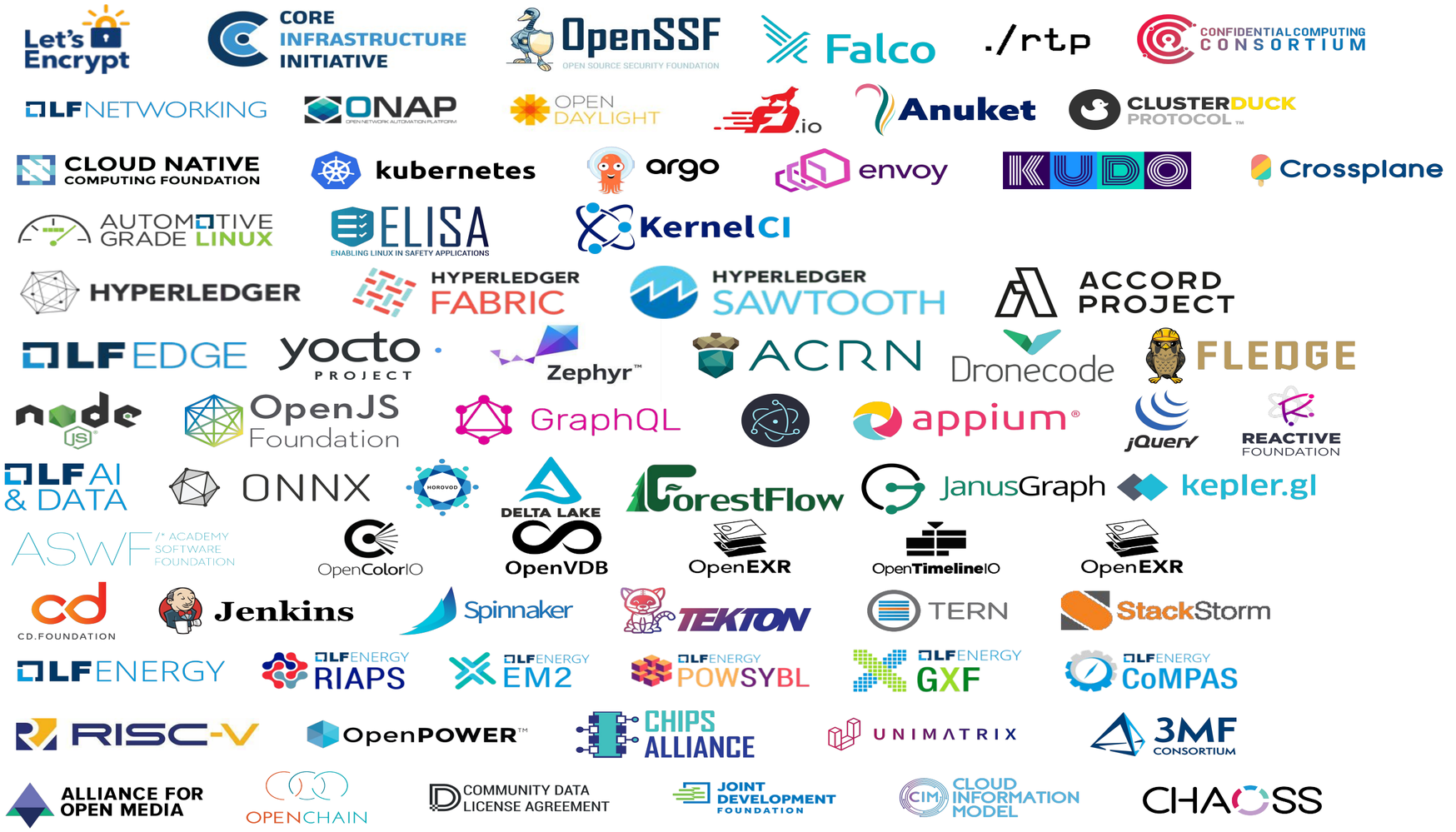
Introduction to the Linux Foundation

The Linux Foundation's goal is to create the greatest shared technology investment in history by enabling open collaboration across companies, developers and users.

We are the nonprofit organization of choice to build ecosystems that accelerate open source technology development and commercial adoption on a global scale.

We are behind some of the most critical projects in the world

- Security
- Networking
- Cloud
- Automotive
- Blockchain
- Edge/IoT
- Web
- AI
- Film
- CI/CD
- Energy
- Hardware
- Standards



Building large, sustainable ecosystems requires collective resources

3,000+
Members From 41 Countries

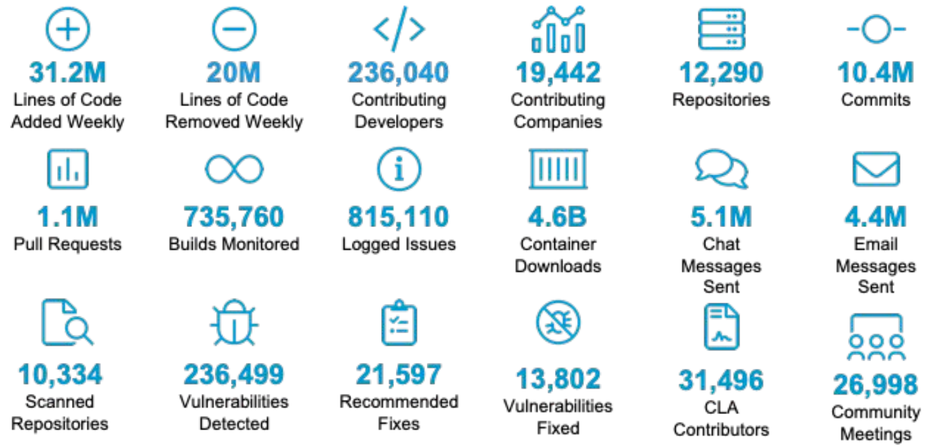
100%
of Fortune 100 Tech & Telecom

242,960+
Developers Contributing Code

980+
Critical Open Source Projects

$100B
Shared Value

The Linux Foundation is a critical part of modern technology



For an Enterprise, Open Source is about Business Strategy

Open Source Strategic Impact

- Accelerate the development of open solutions
- Provide an implementation to an open standard

- Commoditize a market
- Reduce prices of non-strategic software assets
- Share development costs

- Drive demand by building an ecosystem for products & services

- Partner with others
- Engage customers
- Strengthen relationships with common goals

Increase market opportunity

- If a certain part of our stack became a commodity and every player adopted it, where can we get first mover and industry leader advantages to expand our market opportunity?
- Are there open, industry adopted southbound or northbound APIs to allow for interoperability and/or data egress/ingress?
- Where can we increase customer value with offerings of additional data services, analytics, management, and security? What industry open source base stack can facilitate this?

Take market share through de facto standards

- Where in the stack can we disrupt a market leading proprietary implementation with a standardized open source implementation that creates a market opportunity for us?
- Which peers, partners and suppliers, do I need to involve to ensure an open source effort becomes adopted industry-wide?
- Where can a customer or competitor use open source to create a de facto standard that forces us to change our product development roadmap and R&D/IT investments?

Decrease costs

- Where is there opportunity to externalize commodity R&D through shared open source development?
- Where is the client shouldering duplicative costs in non differentiating vendor solutions and how we can we use open source to move our relationship up the stack while removing duplicative R&D?

Strategy is Built Upon **Category Creation** through EcoSystem Formation and Industry-Wide Standardization on Open Source Software

Open Source enables companies to reshape their industry value chain in ways that are both neutral (to create **Industry Platforms**) and that drive their own competitive goals (through active **Stakeholdership**)

Old World “Standards” = Specifications with Divergent Implementations

New World Standards = **Industry-wide adoption** of **shared technology development** and critical deployment alongside specifications to become a de facto platform

Stakeholders invest in active, multi-year **Membership** to see hundreds of millions, sometimes billions, in return by collecting value from **Industry Platforms**

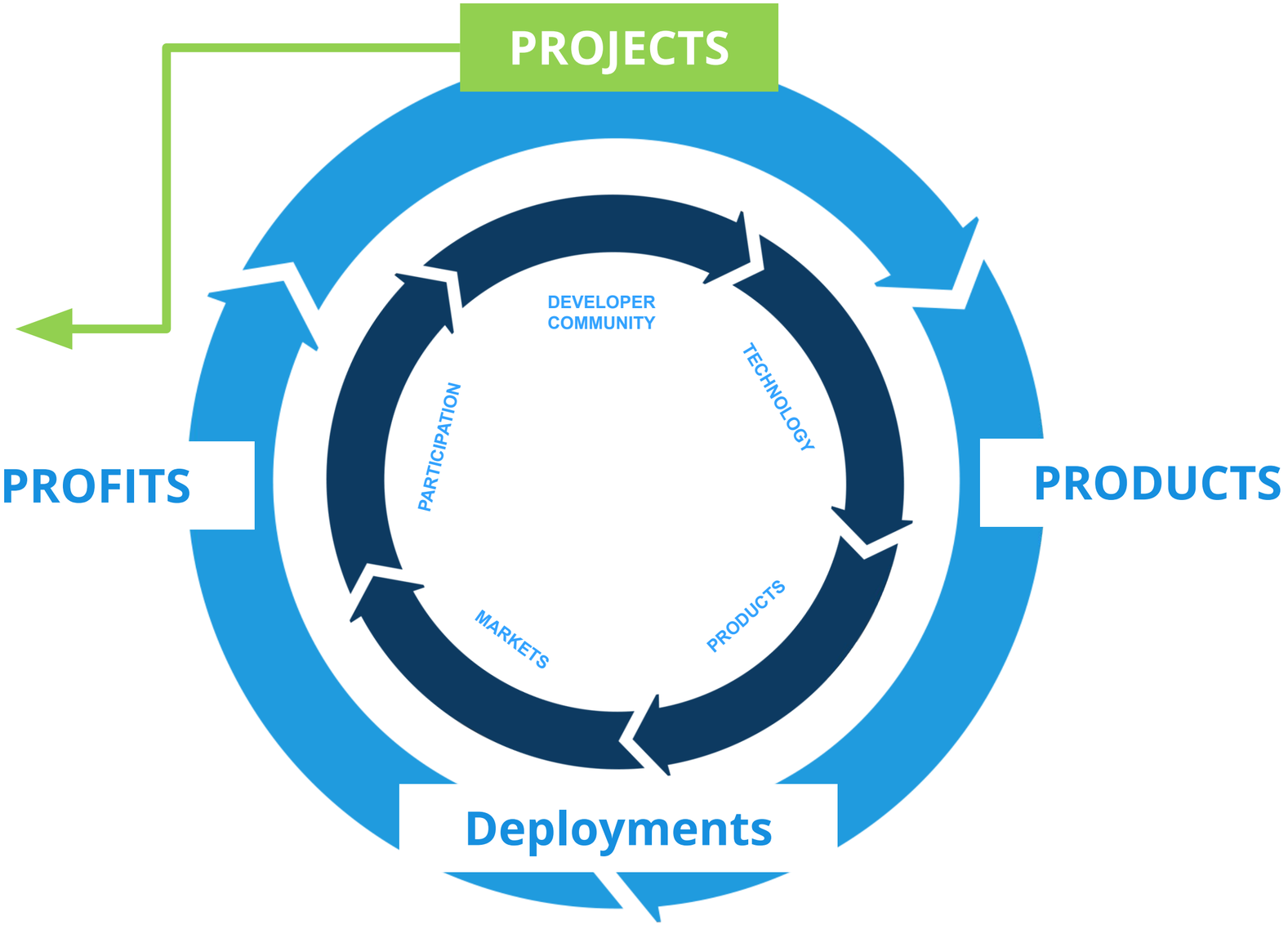
“You are either at the table or on the table. “

—Anonymous Executive

Projects

Open Source Journey – First Came Projects

Successful Open Source Development depends on the complete life cycle of projects, products that market will adopt and deploy



Our Role Has Been Recognized Alongside Tech Titans

- “This category looks at those companies, associations and projects that have inspired development and IT shops to build upon the work they have created, and recognizes them for their leadership as we begin to create a digital world we only could have dreamed about a generation ago.
- SD Times Influencers: Apple, Facebook, Google, IBM, Intel, Microsoft, GitHub, Netflix, Red Hat, Slack, *The Linux Foundation*

In 2019, The Linux Foundation is the Only Non-Profit Recognized For Three Consecutive Years For Its Critical Role in the Industry

The Linux Foundation Platform assists projects in 5 key areas

Governance and Membership	• Project technical governance, policies, etc. • Ongoing business development and membership recruitment • Membership management
Development Methodology and Processes	• Technical decision making • Project life cycle • Release processes
Infrastructure	• CI/CD infrastructure using open source best practices • Release engineering, DevOps • Security and reliability
Ecosystem Development	• Evangelism, marketing and outreach initiatives • Events bringing developers, users and solution providers together • Training for developers and administrators, establishing professional certifications
IP and Assets Management	• Code provenance • Trademark management • IP Policy, license scanning, IP defenseas

Building large, sustainable ecosystems require collective resources

Events

We gathered over **35,000 attendees** from over **11,000 organizations** across **113 countries** in 2018

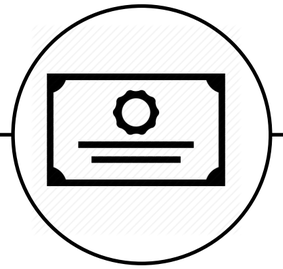
Legal

We **manage IP for the world's most important tech** and have some of the world's top source legal team in house

Training

We have **trained millions of students** through free and paid online training, online skills certification, and on site e-learning

Certification

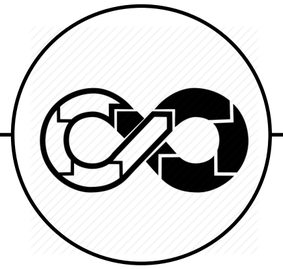


We have designed and implemented both software and hardware **testing and certification programs**

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 mgmts., IT ops and support

Application Security

In addition to massive peer review, **our projects our regularly audited and pen tested.** We offer bug bounties, dependency analysis, and code scanning.

We enable organizations to become open source leaders

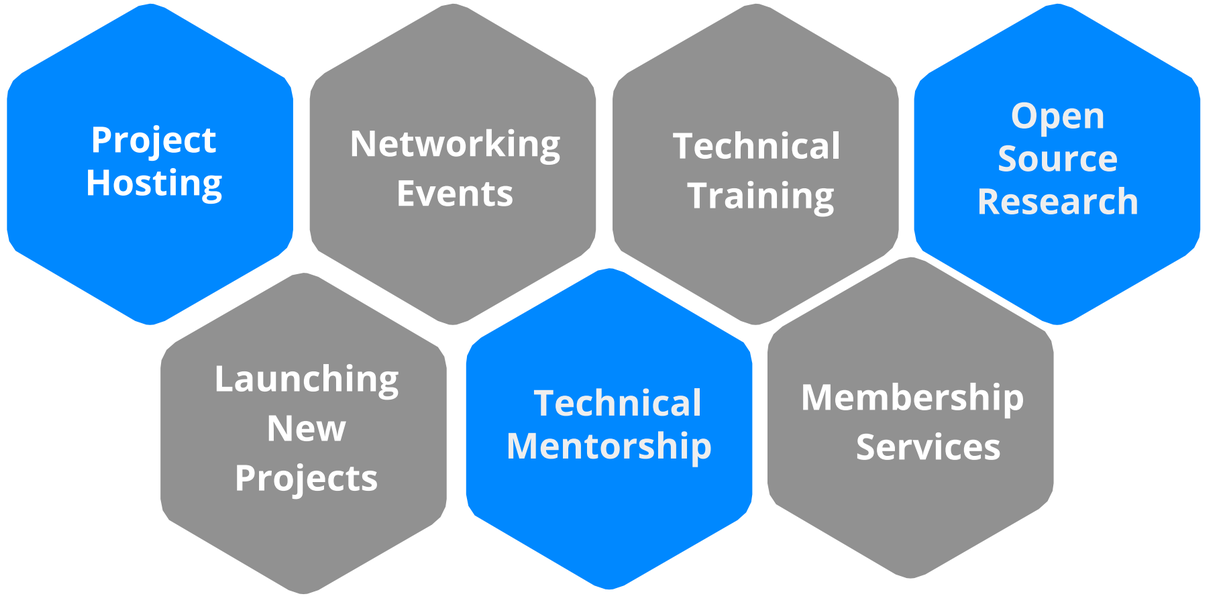
We help organizations navigate the open source ecosystem

Our services help companies manage open source strategically.

We host the world's most important projects.

Our events are the place where top developers and users of open source meet.

Our training programs meet the demand for OSS skills from the source.



LF AI & DATA

LF AI & Data Foundation

Ibrahim Haddad, PhD
Executive Director

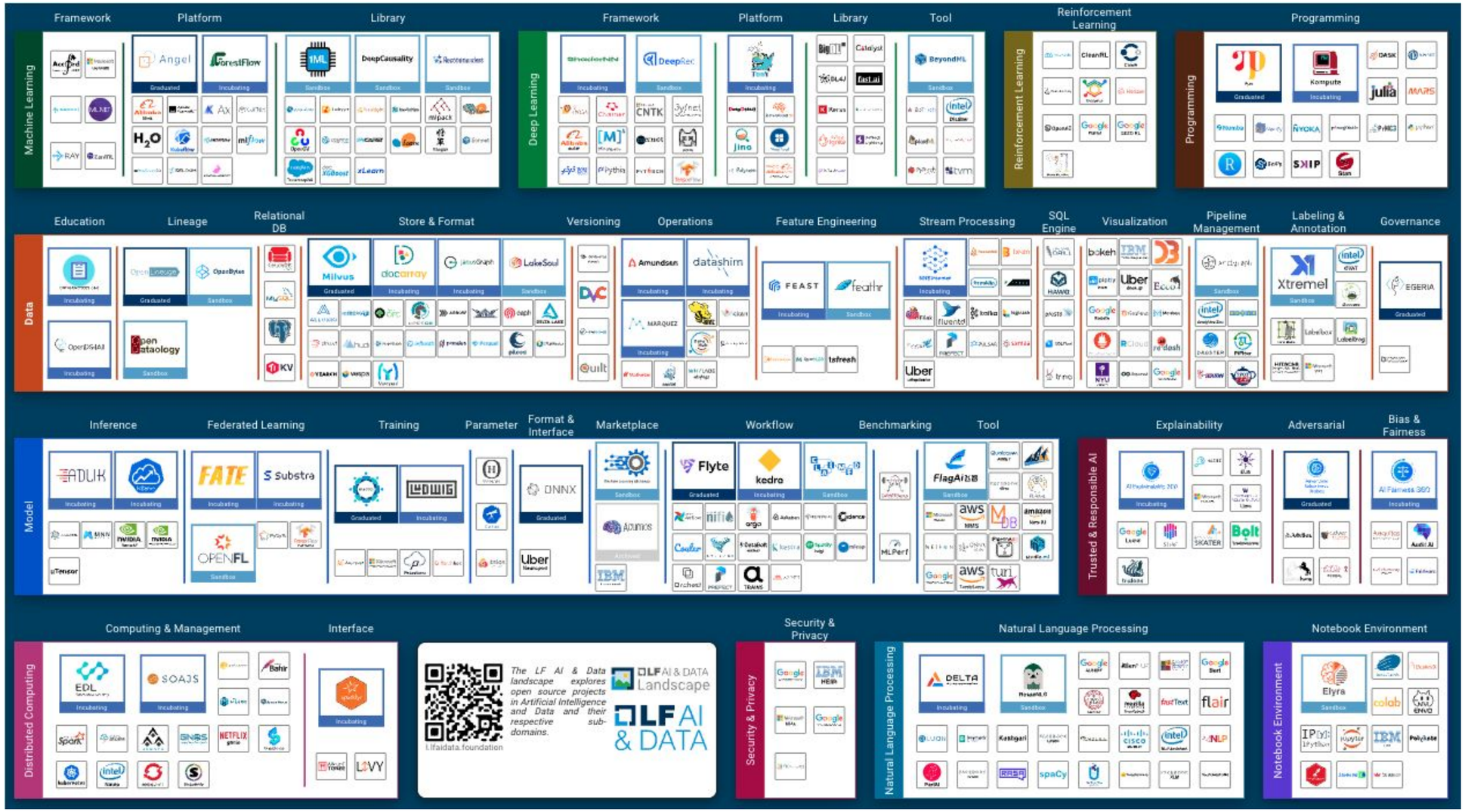
Who is LF AI & Data?

The LF AI & Data is a global not for profit foundation that hosts critical components of the global AI & Data technology infrastructure. It brings together the world's top developers, end users, and vendors to identify and contribute to the projects and initiatives that address industry challenges for the benefit of all participants.

Evolution & Mission

Our mission is to build and support an open AI community, and drive open source innovation in the AI, ML, DL and Data domains by enabling collaboration, sharing best practices, supporting development efforts, and the creation of new opportunities for all the members of the community.

A growing ecosystem: The barrier to entry in AI is lower than ever before, thanks to open source software



LF AI & DATA

LF AI & DATA Landscape

340+ Projects

2.8M+ GitHub Stars

95K+ Developers

200+ Founding Org

600M+ LoC

1M+ LoC / Week

1000s of Contributing Orgs

Open source libraries, frameworks, platforms, tools, are a two-way street: they make AI accessible to everyone, and companies benefit from a community of other contributors helping accelerate open AI applied research.

Cost	**TTM**	**Interoperability**
Value	**Access to talent**	**Integration**

Open source benefits for AI & 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

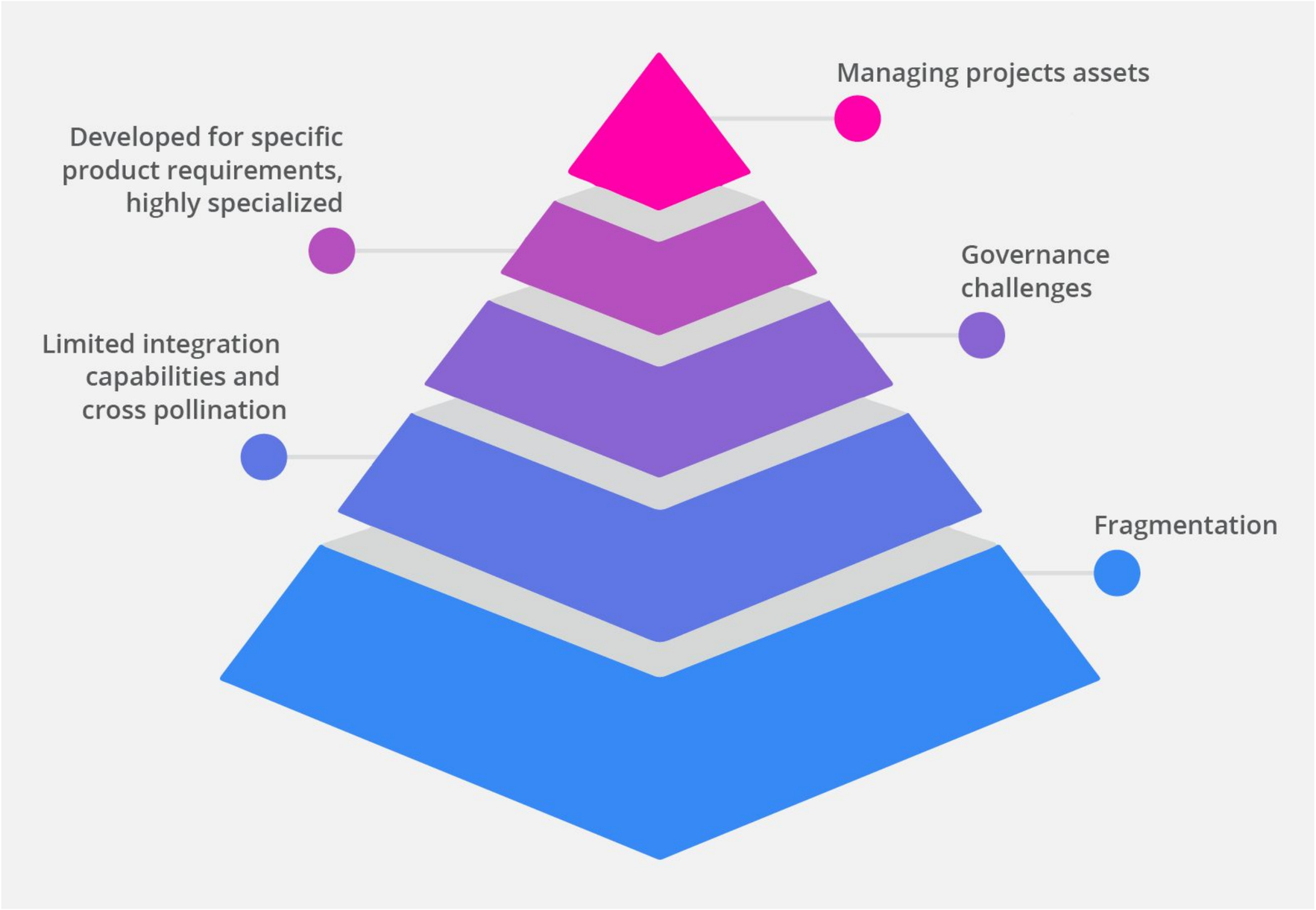
AVAILABILITY

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

But with challenges…



Members

July 12, 2023

Foundation Structure and Governance

Foundation Governance

(Funding effort)

Technical Coordination

Hosted Projects

Governing Board

Outreach

Legal

Strategy

Budget

Technical Advisory Council

ML Workflow & Interop

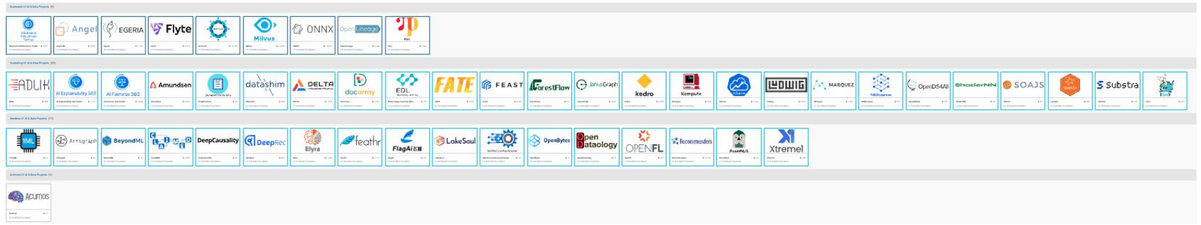
Trusted AI

BI & AI

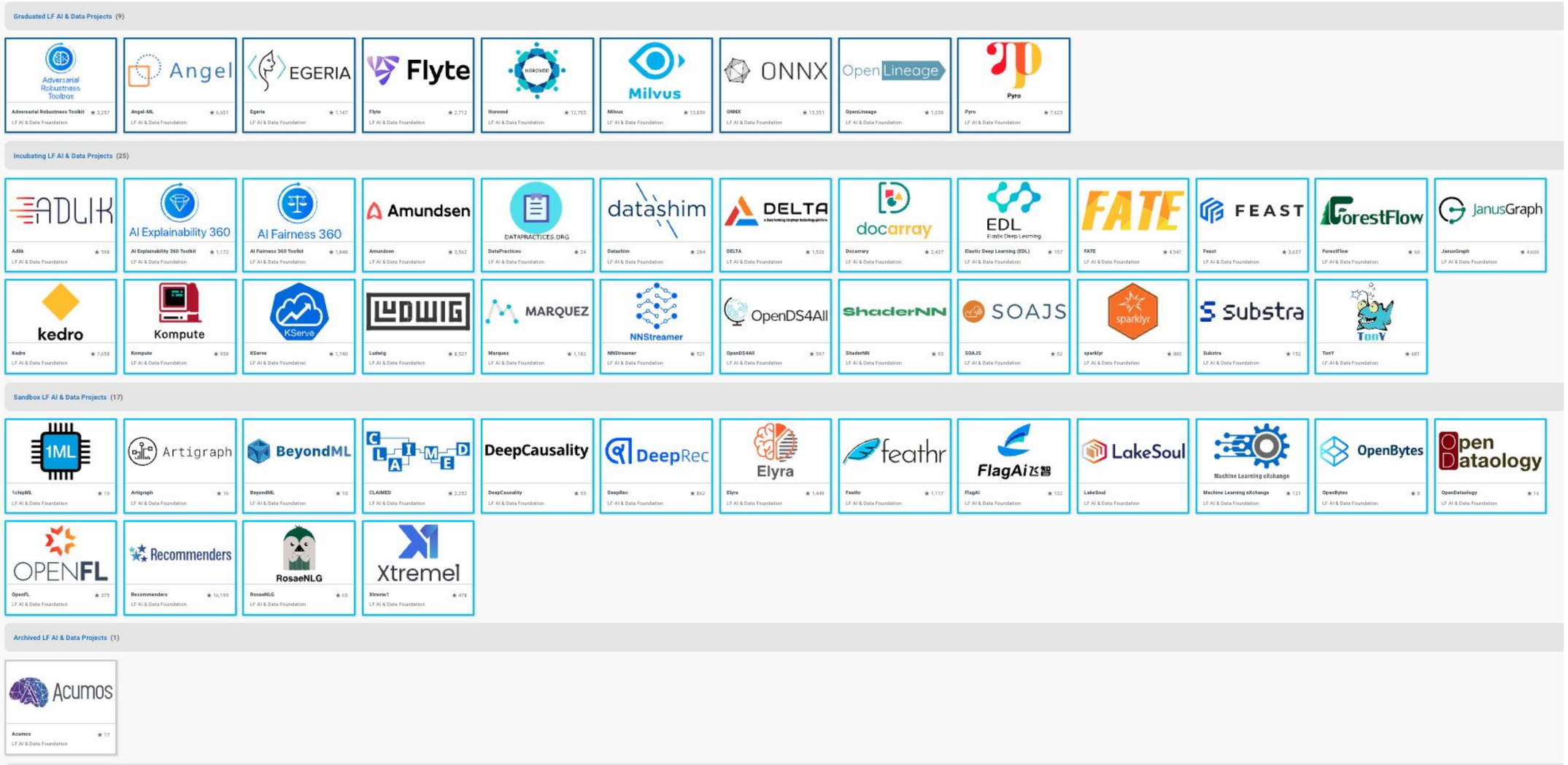
DataOps

ML Ops

MLSecOps



Projects Hosted in LF AI & Data



Key Stats

57
Member Organizations

50
Active Technical Projects

6
Active Committees

650+
Contributing Organizations

206,000+
GitHub PRs

27,000+
Active Contributors
(>68K total contributors)

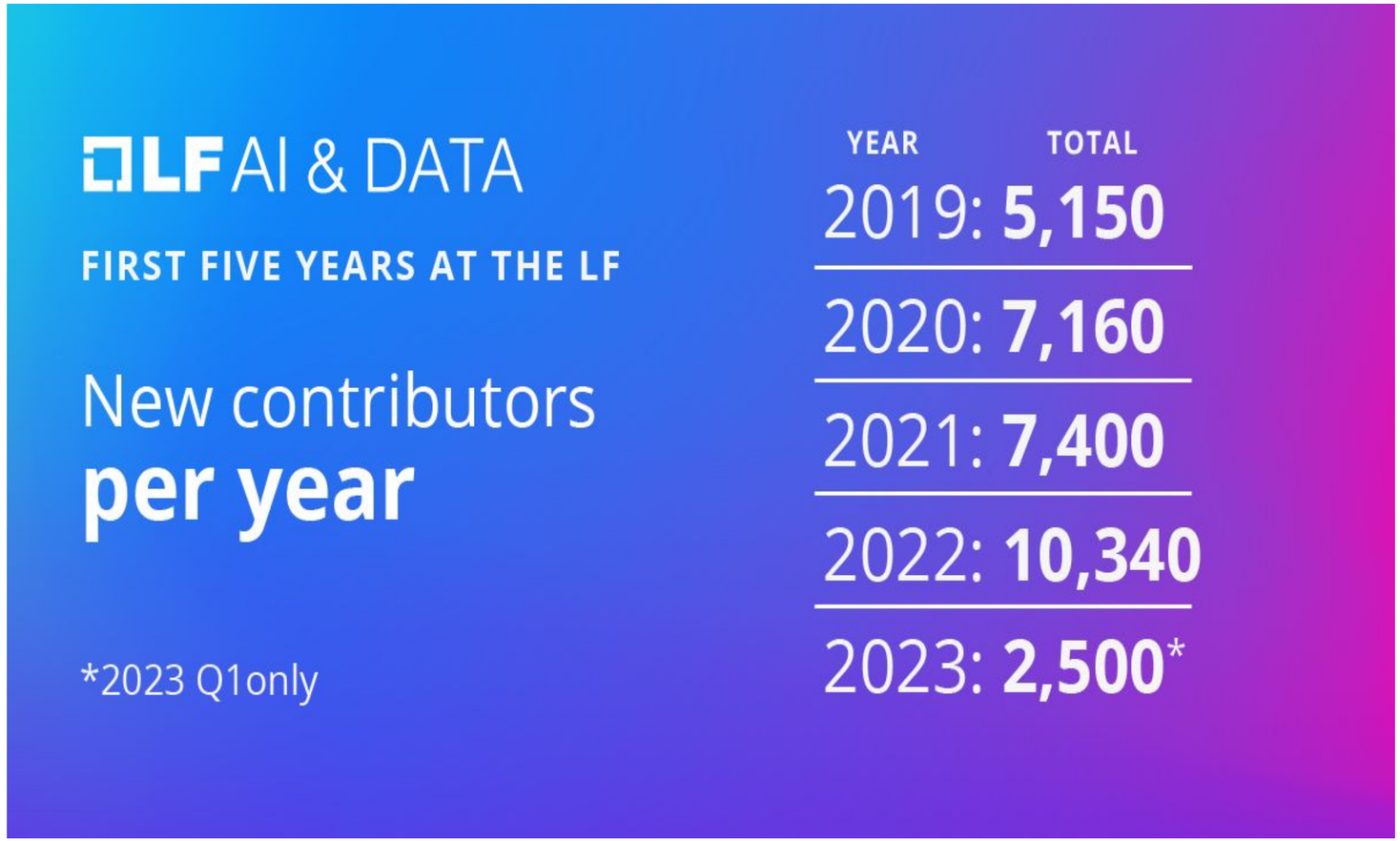
20+
Technical Integrations

Total Contributors

May 19, 2023, Source: LFX Insights

New Contributors Stats

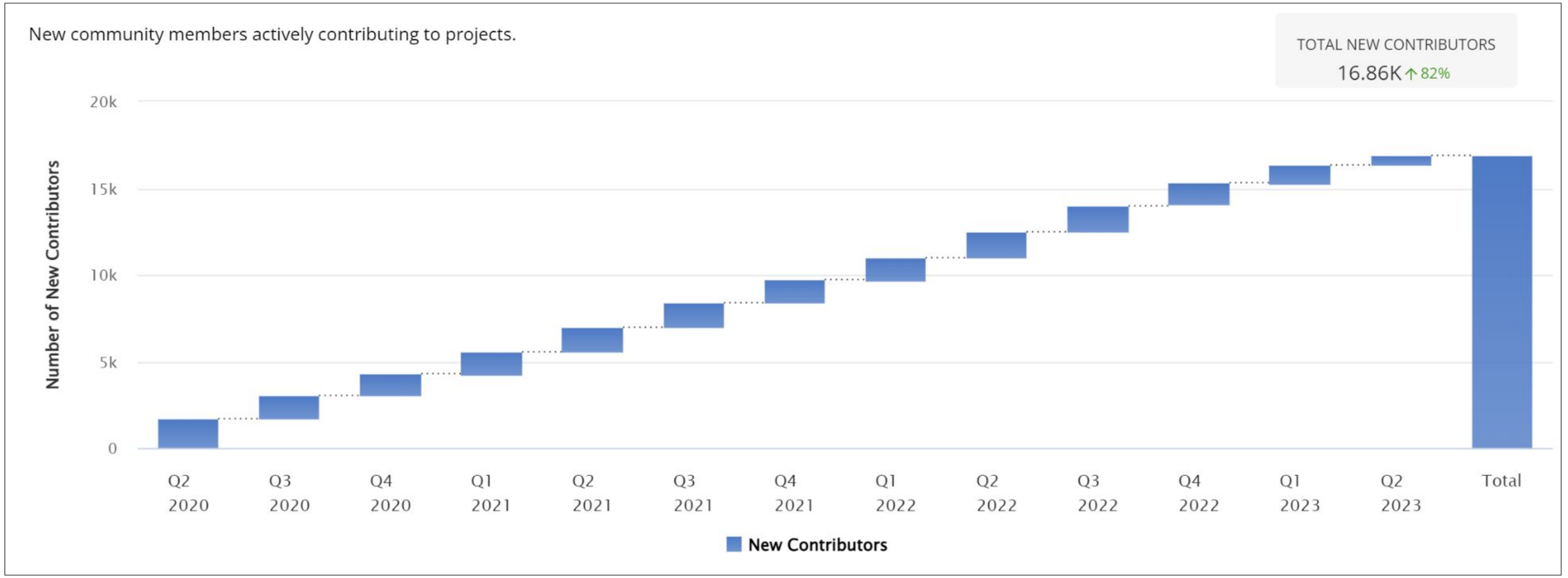
May 19, 2023, Source: LFX Insights



Total Number of Contributor (Q2 2020 - Q2 2023)

Contributor Growth (Q2 2020 - Q2 2023)

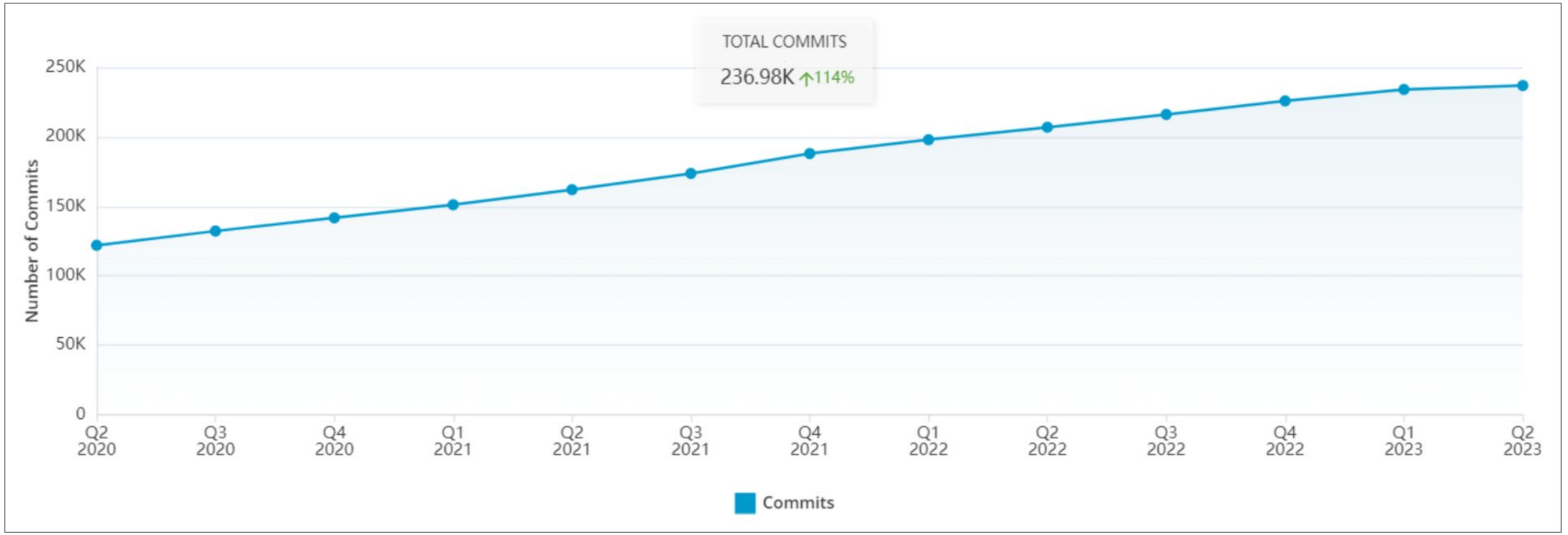
May 19, 2023, Source: LFX Insights



Our programs are enabling our projects to increase the contributions from existing developers and welcome new developers into projects at an unprecedented rate.

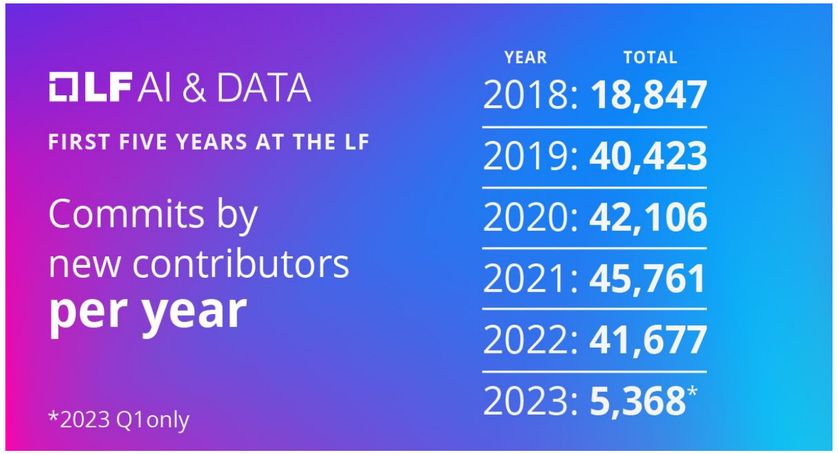
Commit Growth (Q2 2020 - Q2 2023)

May 19, 2023, Source: [LFX Insights]



Code Stats

May 19, 2023, Source: LFX Insights



Across 297 total repositories, an average of 35.64K LOC were added per repository on a weekly basis.

Development Velocity

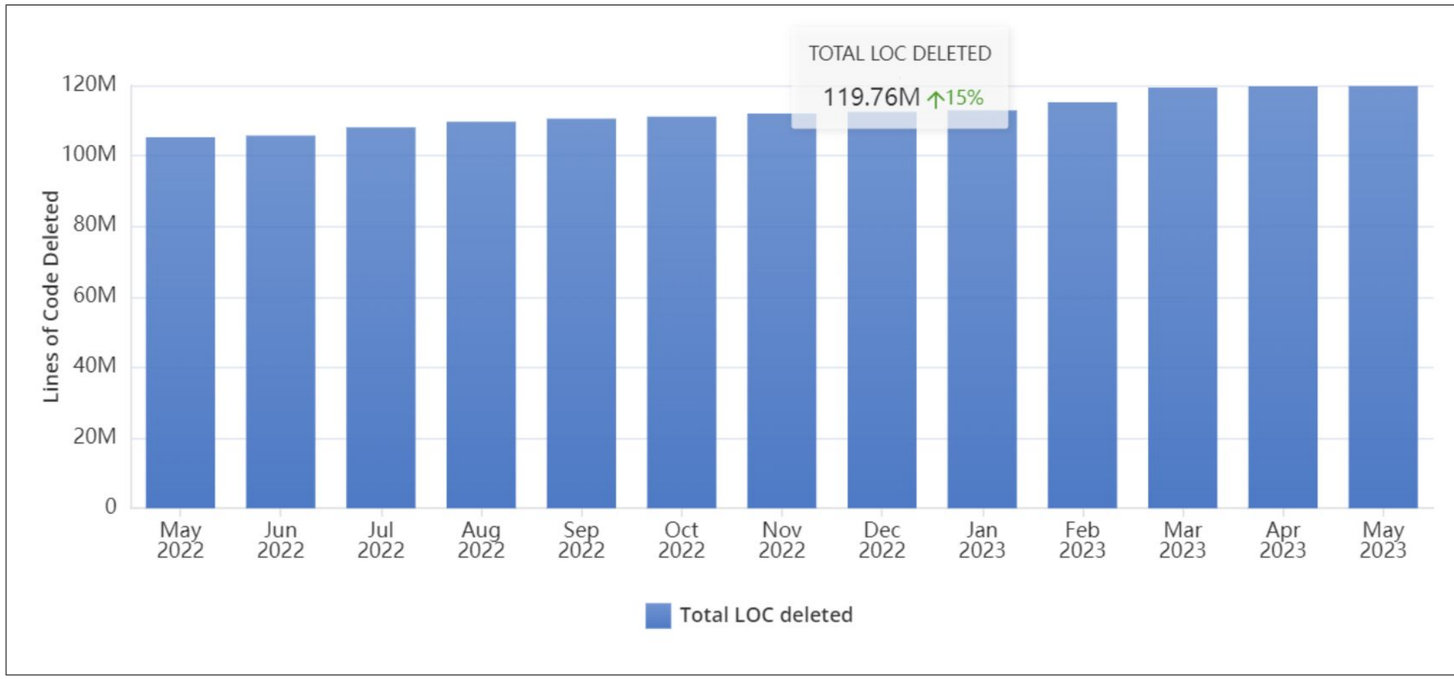
<u>May 2022 - May 2023:</u>

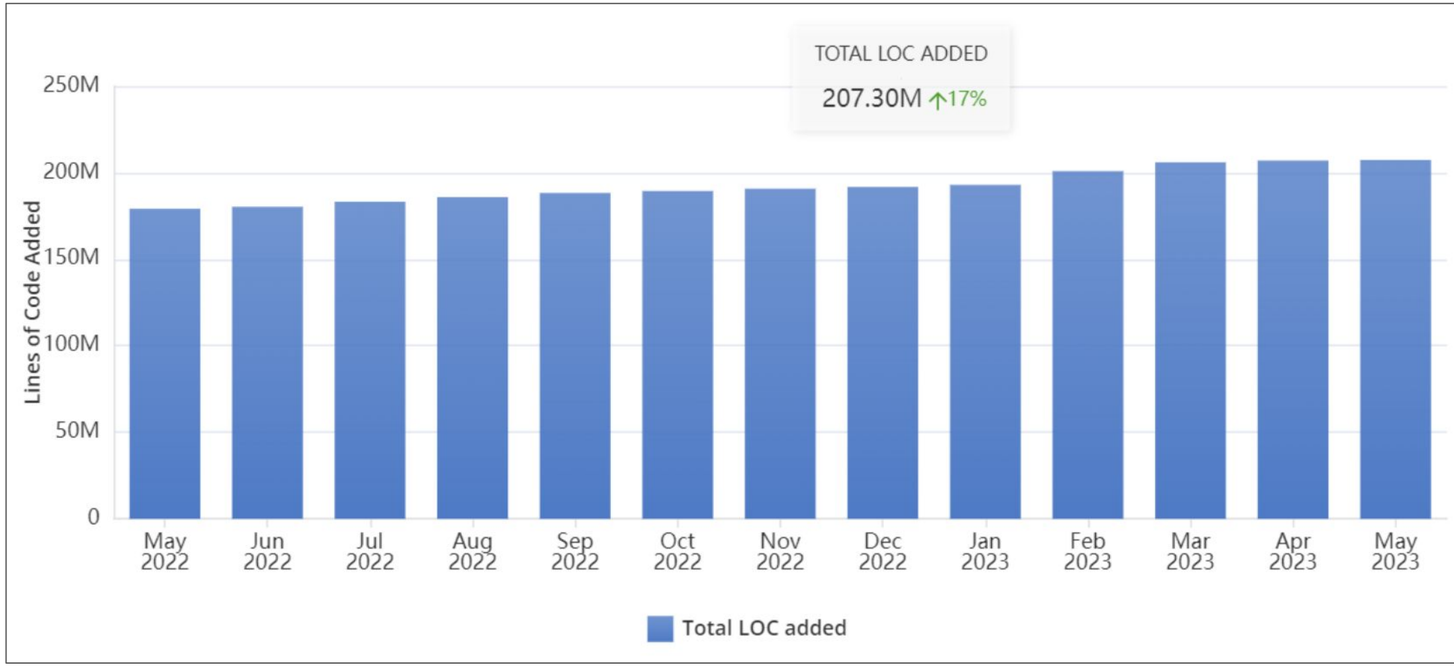
- 119.76 M LoC deleted
- 207.30 M LoC added

Average net LoC added daily:

239,000

Every single day of the week





Code Stats

May 19, 2023, Source: LFX Insights

Across 297 total repositories, an average of 35.64K LOC were added per repository on a weekly basis.

An average of 275.20M LOC were added in the last 5 years.

Organization Engagement

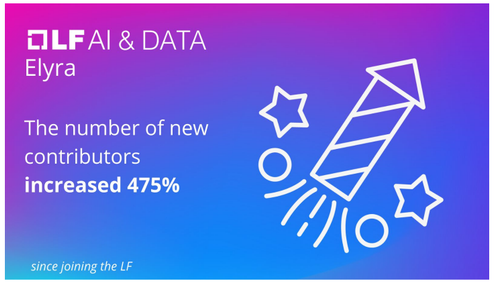
May 19, 2023, Source: LFX Insights

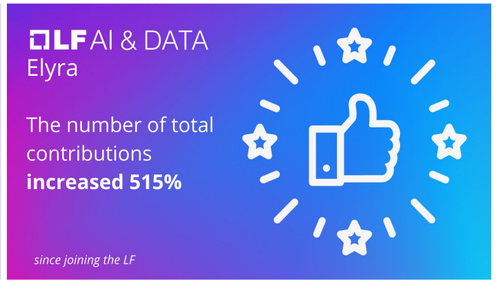
Key Insights:

1. A total of **658 organizations participated in the code commits** during the past 5 years.
2. An average of 234 commits were contributed by individual contributors during the last 5 years.
3. An average of 364 commits were contributed by unaffiliated contributors during the last 5 years.

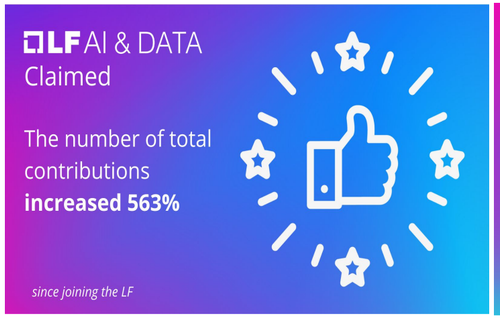
Sample Project Growth

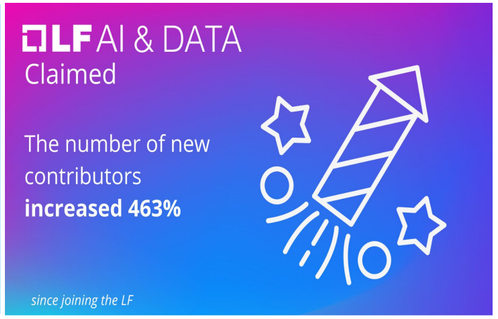
Elyra
Joined 01/2021



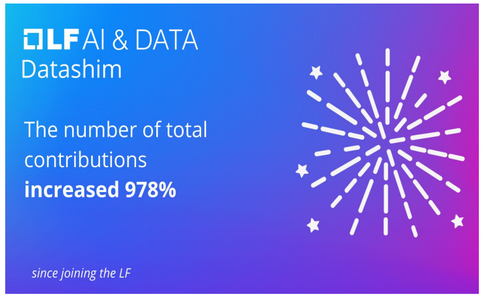


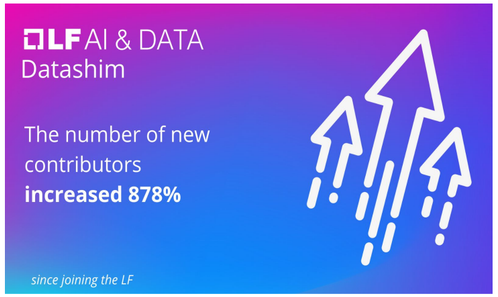
CLAIMED
Joined 11/2022



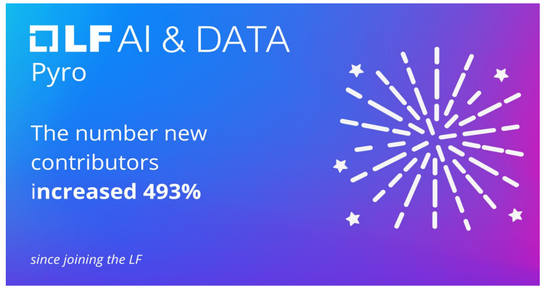


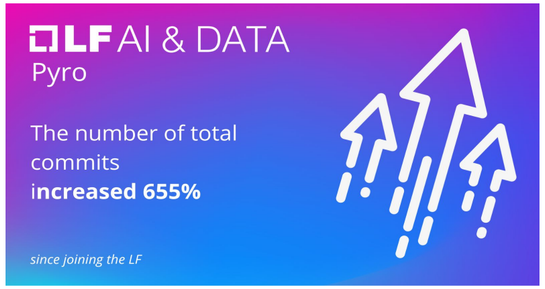
Datashim
Joined 01/2021



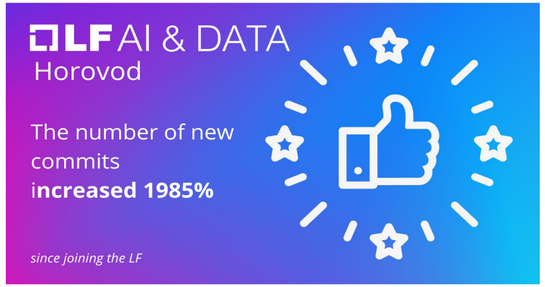


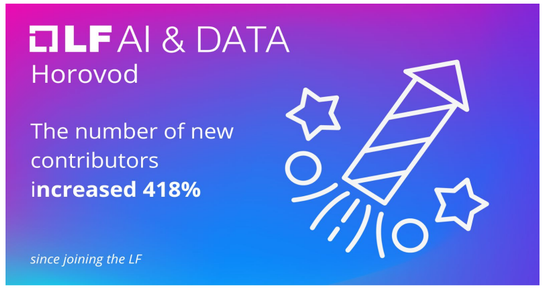
Pyro
Joined 01/2019



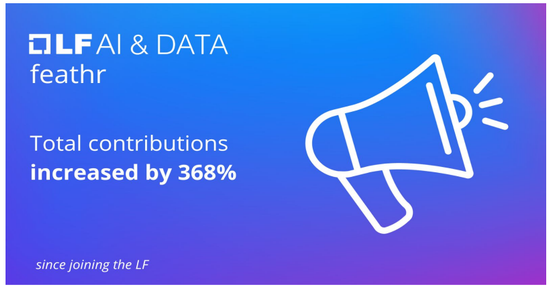


Horovod
Joined 12/2018



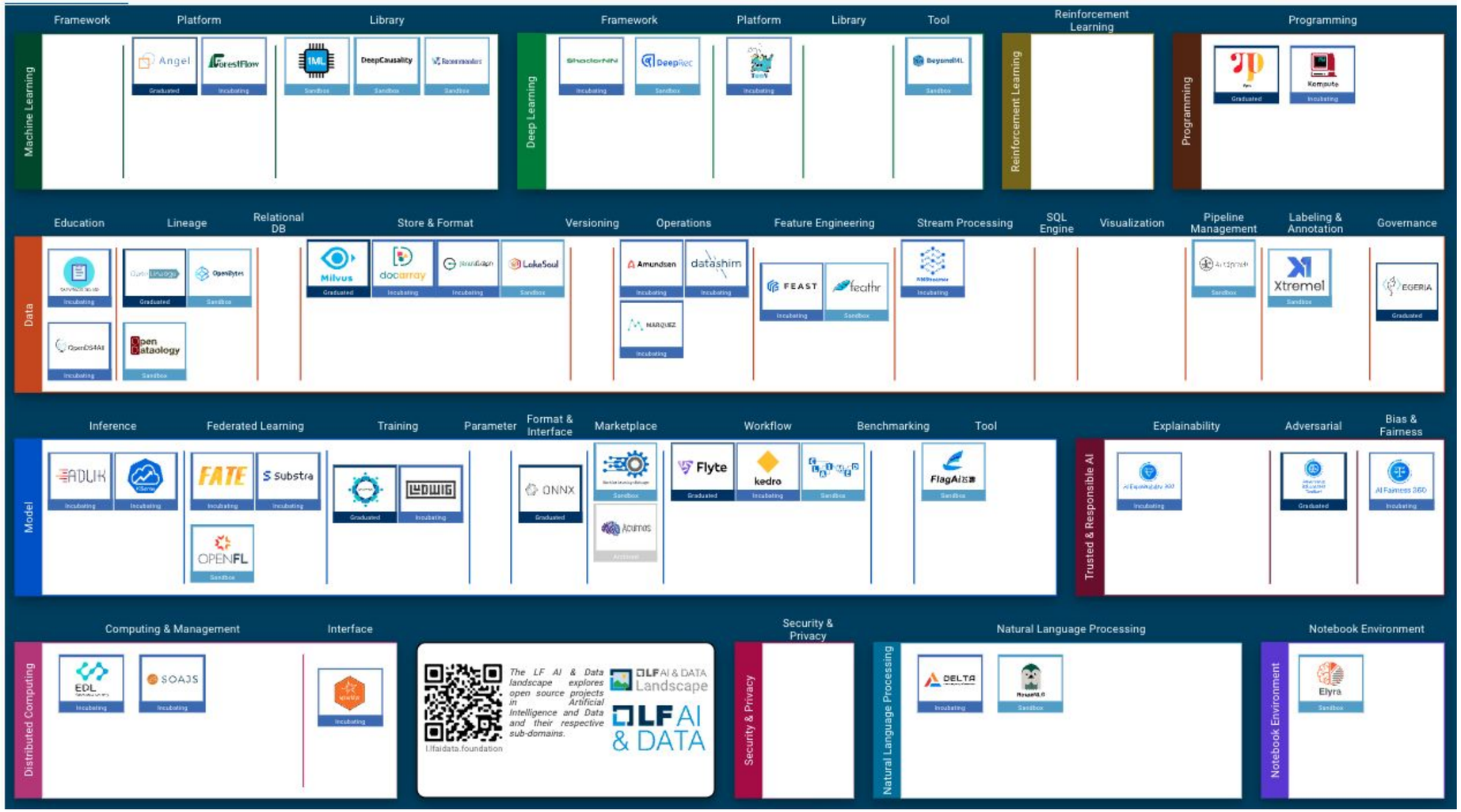


feathr
Joined 08/2022



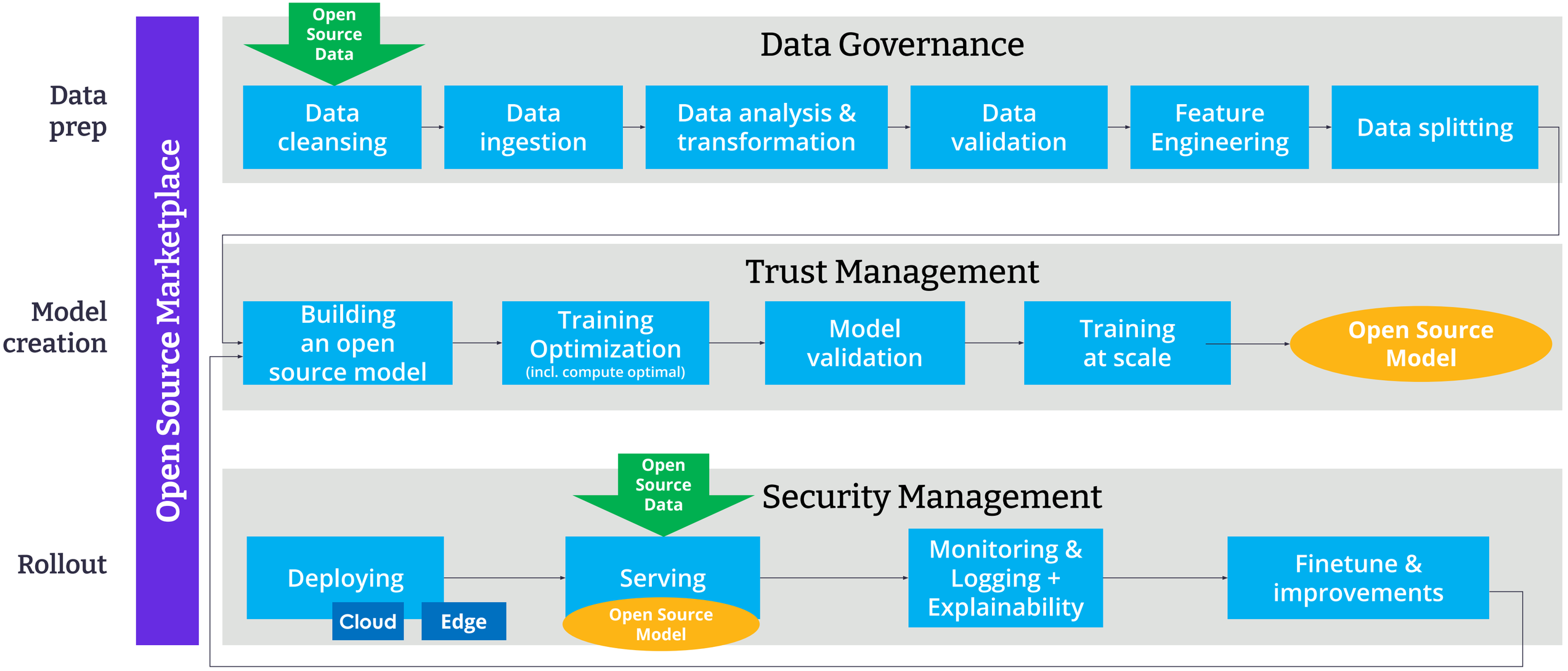
Over 13% of the ecosystem's key projects depend on LF AI & Data

for their open governance, safe haven for their assets, infrastructure, and enabling marketing, legal and event services, and staff that is eager to help and support the communities of these various projects

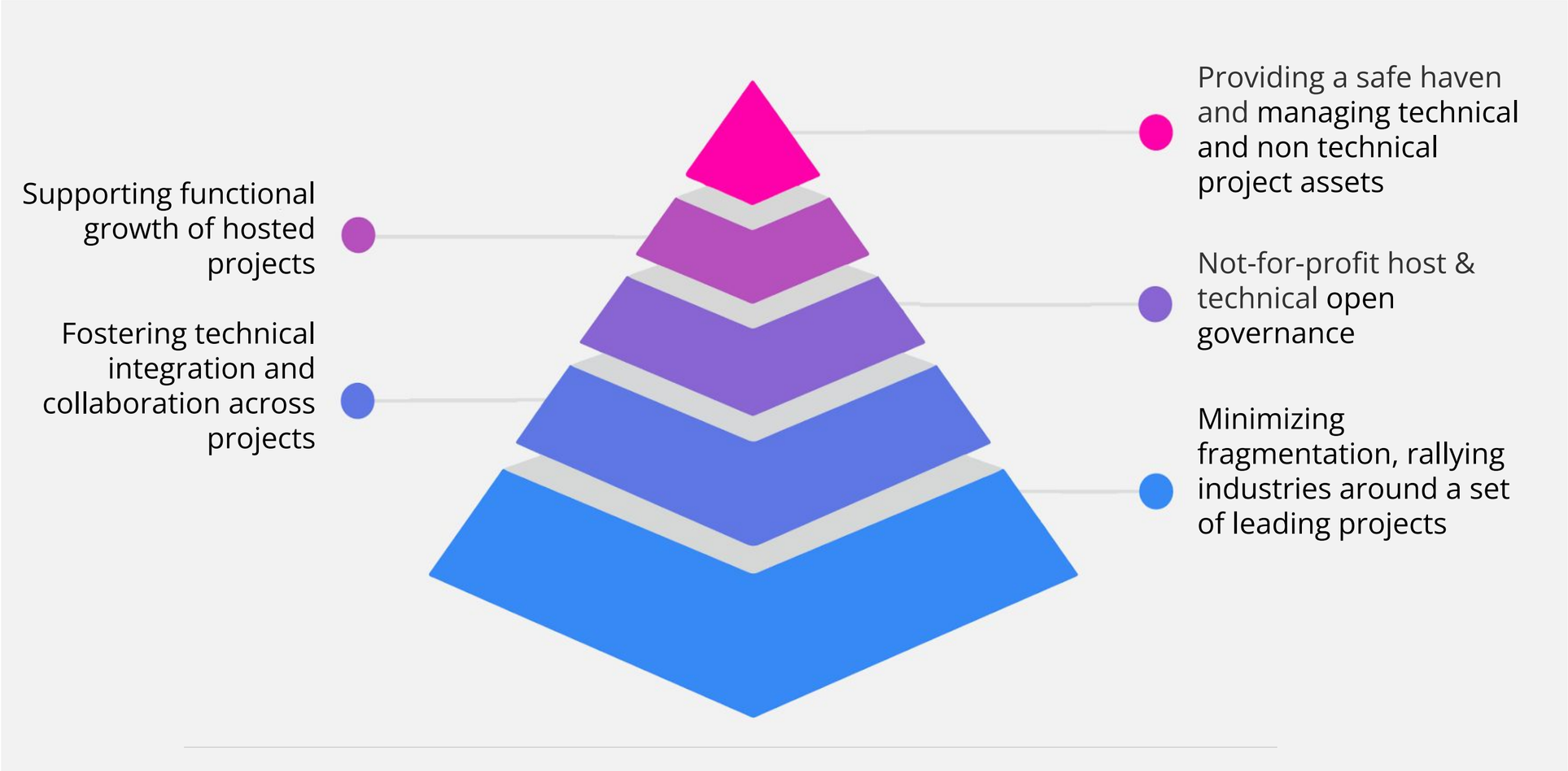


Industry Leaders Host Their Projects in LF AI & Data

With 50 Hosted Technical Projects, LF AI & Data offers all required open source software and elements required to build and manage an end-to-end ML Workflow



LF AI & Data is addressing key ecosystem challenges



Support Programs Available to 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.

Strong Project Participation from China

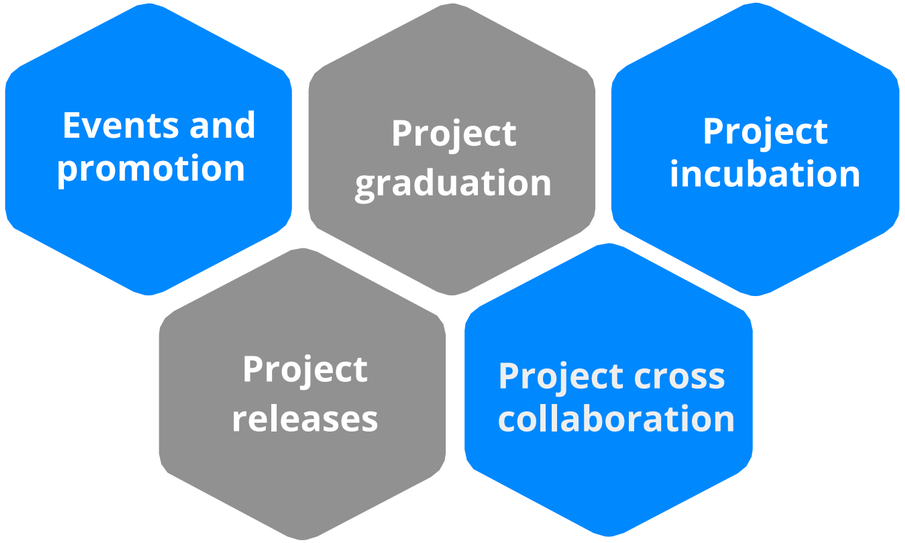
Hosted Projects

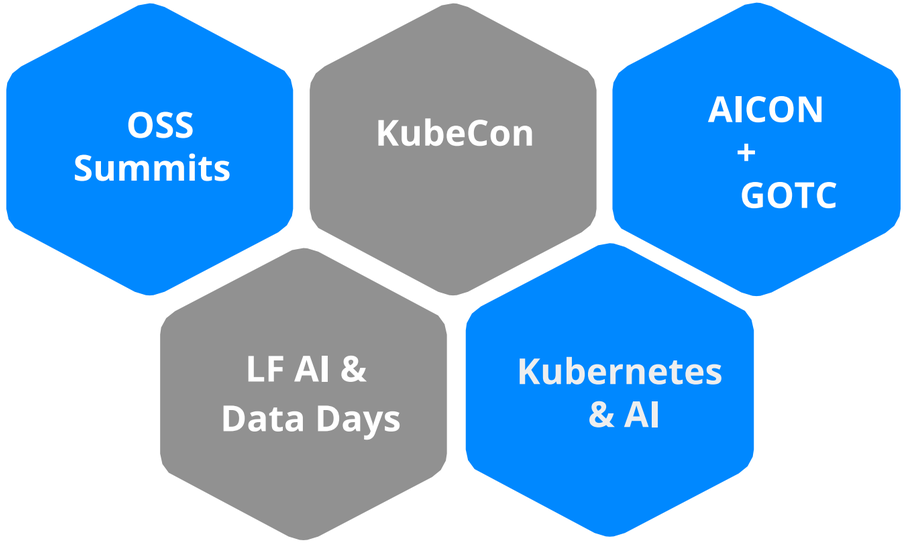
Organizations Hosting Projects

Focused Developer & Community Marketing Effort

MarComm services to support community and ecosystem engagement

Events and community ecosystem building services





Ongoing Major Activities

1. **Expanding Membership**
2. **Expanding Scope**
    a. Increased focus on Data tooling (data processing, pipeline management, validation, data storage, visualization)
    b. New focus on Gen AI tooling and Models
    c. New focus on hosting Data Sets
3. **Generative AI Commons**
    a. Collaborating with industry leaders and academia to establish, promote, and support the development and innovation of open science generative AI technologies
    b. Hosting and supporting a wide range of generative AI projects and tools that drive innovation and democratization of AI .
    c. Ensuring ethical, transparent, and accessible AI that benefit all of humanity.
4) **Model Licensing Framework and Self Certification**
    a) Support the open sourcing of model elements required for a true open source experience
    b) Create comprehensive licensing model for generative AI models and open science
    c) Create an ISO standard (similar to what the LF did with OpenChain) to help rate the openness level of an AI model by measure of how many of its elements are made available under an open source license
5) **Policy**
    a) Work to share knowledge and expertise to develop responsible and informed AI legislation
    b) Provide thought leadership and expertise as members embark upon their generative AI journeys

Re-cap

- Established itself as the foundation of choice for open source AI projects. We're limiting ourselves to 1 new project per month to ensure proper onboarding and integration of the project with our operation. That's even with tightened hosting requirements
- Grew a large developer community centered around AI and data and projects
- Enabled 20+ integrations across multiple key projects in the ecosystem (stopped tracking/counting over a year ago - in reality number is higher)
- Continue to address to trusted AI related concerns both via technical projects as well as the publication of recommended policies and practices

Call to action - Engage with us

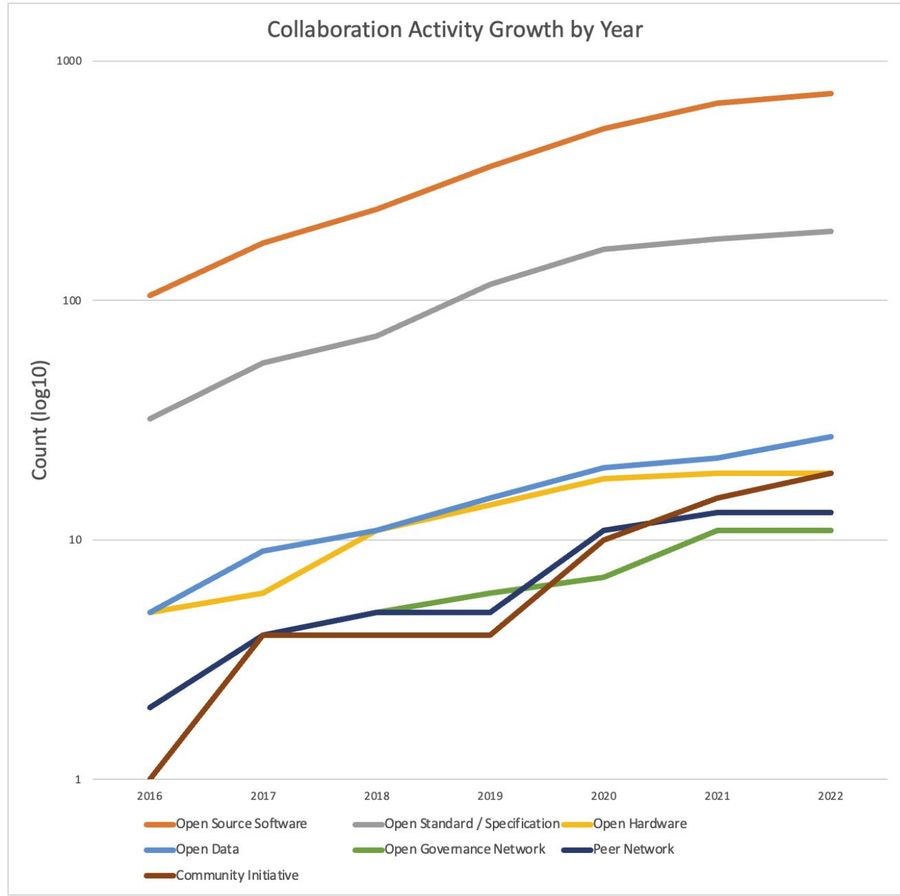
1. Attend our events
2. Participate in our open Technical Advisory Committee calls (bi-weekly)
3. Host projects with us and use our expertise and market position to grow your projects, drive adoption and become winning projects in your specific categories
4. Join us as a member and participate in shaping our strategy and drive thought leadership in a very complex and intertwined ecosystem

From Jim

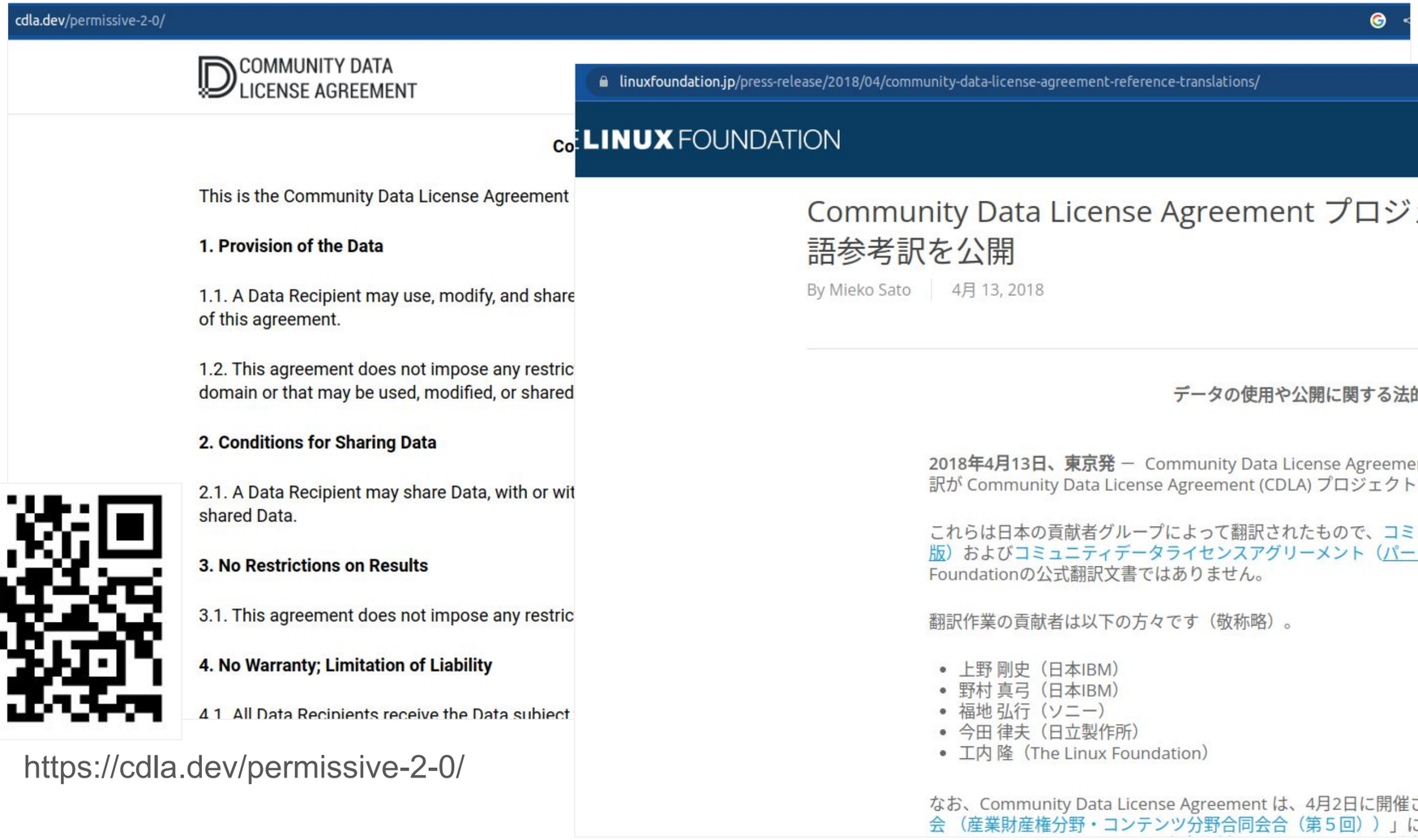
The growth in AI, deep learning, and ML led to a rise in Open Data proje

Activity Type	2022	2019	Change	CAGR
Open Source Software	733	362	371	27%
Open Standard / Specification	194	117	77	18%
Open Hardware	19	14	5	11%
Open Data	27	15	12	22%
Open Governance Network	11	6	5	22%
Peer Network	13	5	8	38%
Community Initiative	19	4	15	68%

Notes: 1) excludes any archived or formation stage projects, 2) 2022 Project Total = 858, 2019 Project = 416. a single project may include multiple collaboration activity types (e.g. Open Container Initiative includes both software and specification collaborations), 3) a Peer Network is a community of individuals working in common roles and exchanging best practices (e.g. TODO Group), 4) Community Initiative is a chartered non-technical working group or special interest group, 5) Collaboration Activity Growth By Year in the line chart is based on a log10 normalization of the same data by year.



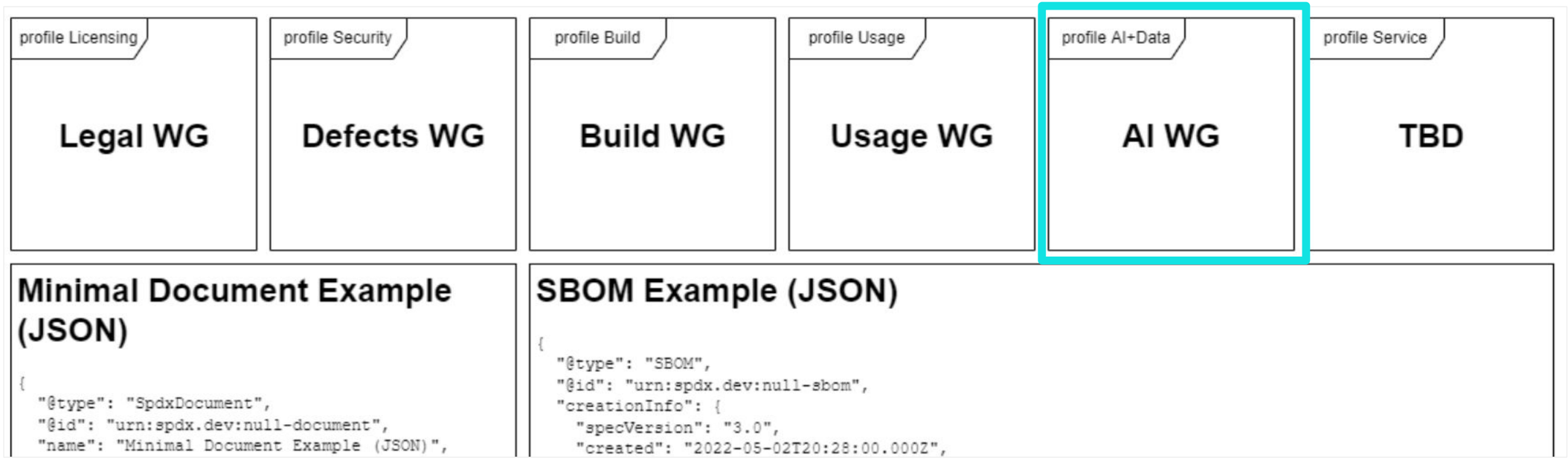
The growth in Open Data collaboration led to new demands for open data license agreements



https://cdla.dev/permissive-2-0/

There is a need to capture AI model information, and we can use familiar SBOM standards and tools to do this

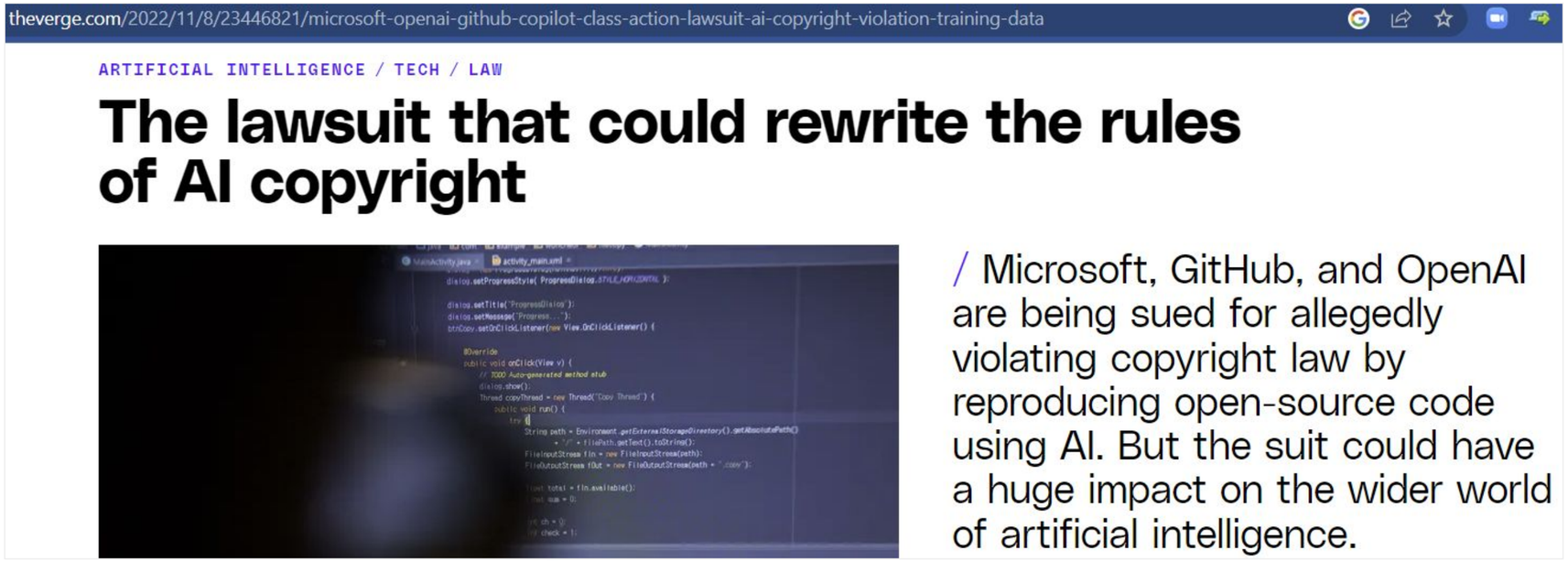
- SPDX 3.0 includes
  - AI profile to convey key properties about AI applications and models.
  - dataset profile to convey relevant metadata about training datasets for AI models and other applications.



https://github.com/spdx/spdx-3-model/blob/main/model.png

Future challenges will exist in building industry alignment around copyright and AI/ML models

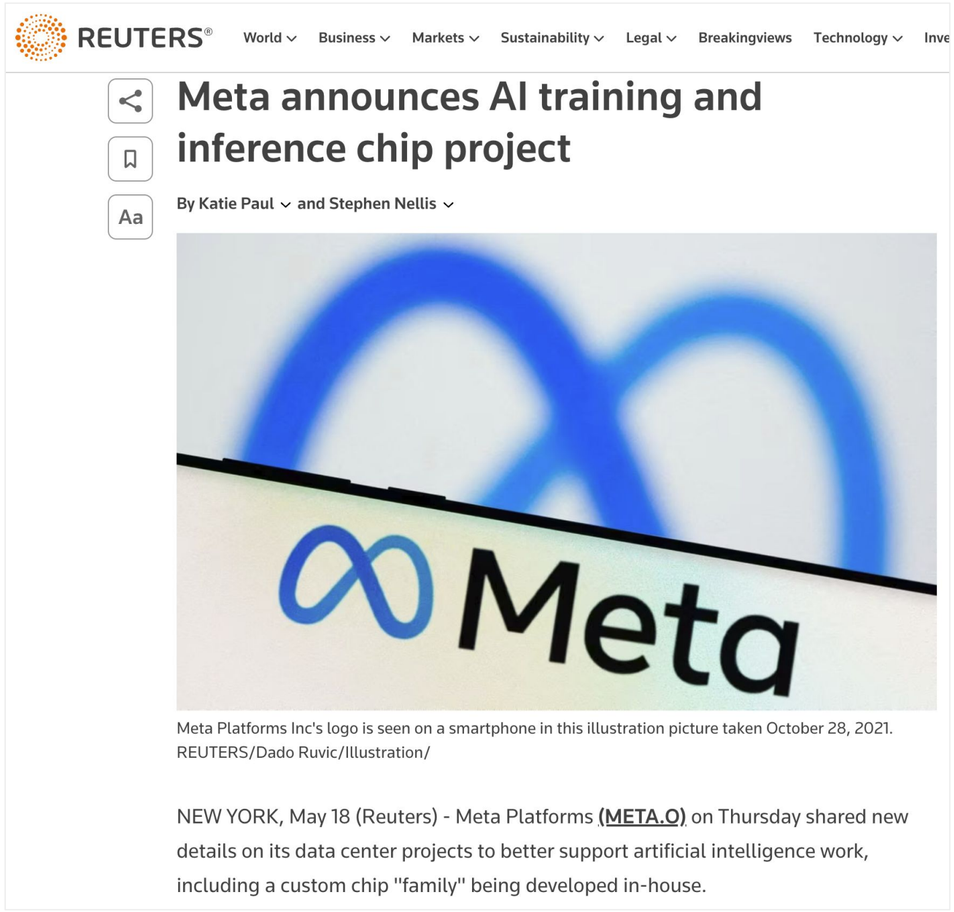
- Open source licensing for software source code is well understood, but…
- Less clear are the implications for AI/ML models built on data, source code, images, or text



Future challenges will exist in building industry alignment around copyright and AI/ML models

- Open source licensing for software source code is well understood, but…
- Less clear are the implications for AI/ML models built on data, source code, images, or text
- *Government regulation could change everything*

Open collaboration across technology opens new waves of innovation



REUTERS® World Business Markets Sustainability Legal Breakingviews Technology Inve

Meta announces AI training and inference chip project

By Katie Paul and Stephen Nellis

Meta Platforms Inc's logo is seen on a smartphone in this illustration picture taken October 28, 2021. REUTERS/Dado Ruvic/Illustration/

NEW YORK, May 18 (Reuters) - Meta Platforms (META.O) on Thursday shared new details on its data center projects to better support artificial intelligence work, including a custom chip "family" being developed in-house.

"Meta had initially turned to graphics processing units, or GPUs, for inference tasks, but found they were not well suited to inference work…

…executives realized it [Meta] lacked the hardware and software to support demand from product teams building AI-powered features…

As a result, the company scrapped plans for a large-scale rollout of an in-house inference chip and started work on a more ambitious chip capable of performing training and inference…

The MTIA chip also used only 25 watts of power" " [1]

"The processor cores are based on the RISC-V open instruction set architecture (ISA) and are heavily customized to perform necessary compute and control tasks….

"The servers that host these accelerators use the Yosemite V3 server specification from the Open Compute Project. "[2]

[1] https://www.reuters.com/technology/meta-announces-ai-training-inference-chip-project-2023-05-18/

[2] https://ai.facebook.com/blog/meta-training-inference-accelerator-AI-MTIA/

Why are companies working on AI/ML in open source?

1. Cost to recreate functionality is many times lower
   - It would take many years at a cost over ¥2.7 trillion to recreate the Linux kernel
2. Interoperability
   - Many customers want interoperability and working on a common implementation in open source delivers highe
3. Time to market
   - Use open source to build a solution faster than building it all in your one company
4. Necessary to build higher value solutions
   - The world could not do cloud computing or AI/ML without a common Linux OS
5. The best engineers in the world on a technology won't work for you
   - Engineers working in open source are often some of the most talented in the world
6. Incorporate development best practices
   - Open source developers have refined software development at scale and companies are trying to replicate open

Hosting Process

Technical Advisory Committee

- The TAC serves a coordination role:
  - Votes on new projects joining LF AI & Data, as well as on promoting projects across incubation stages
  - Facilitates communication and fosters collaboration across hosted technical projects
  - Communicates needs and requirements of the projects to the Governing Board
  - Onboards new projects, assists in progression of existing projects, and reviews projects annually
  - Defines and maintains the technical vision for the LF AI & Data Foundation
  - Creates a conceptual architecture for the projects, aligning projects, promoting, removing or archiving projects
  - Defines common practices to be implemented across LF AI & Data projects

- Meetings via conference calls take place every 2 weeks, are recorded and open to the general public
  - https://wiki.lfai.foundation/pages/viewpage.action?pageId=7733341

Project Incubation levels

Sandbox

- Any project that intends to join LF AI & Data Incubation in the future and wishes to lay the foundations for that.
- New projects that are designed to extend one or more LF AI & Data projects with functionality or interoperability libraries.
- Independent projects that fit the LF AI & Data mission and provide the potential for a novel approach to existing functional areas (or are an attempt to meet an unfulfilled need).

Incubation

Sandbox requirements plus:

- Have 2+ organizations actively contributing to the project.
- Have a defined Technical Steering Committee (TSC) with a chairperson identified, with open and transparent communication.
- Have a sponsor who is an existing LF AI & Data member.
- Have at least 300 stars on GitHub.
- Have achieved and maintained a Core Infrastructure Initiative Best Practices Silver Badge.
- Have the affirmative vote of the TAC.

Graduate

Incubation requirements plus:

- Have a healthy number of code contributions coming from at least five organizations.
- Have reached a minimum of 1000 stars on GH.
- Have achieved and maintained a Core Infrastructure Initiative Best Practices Gold Badge.
- Have demonstrated a substantial ongoing flow of commits and merged contributions for the past 12 months.
- Receive the affirmative vote of two-thirds of the TAC and the affirmative vote of the Governing Board.
- Have completed at least one collaboration with another LF AI & Data hosted project
- Have a technical lead appointed for representation of the project on the LF AI & Data TAC

Incubation Requirements

Sandbox

- Fit the scope and mission of LF AI & Data
- Have an OSI-approved license.
- Have a sponsor who is an existing LF AI & Data member. Alternatively, a new organization would join LF AI & Data and sponsor the project's incubation application.
- Have an open and documented technical governance. The LF team can help set this up.
- The project's founders adopt an open governance model documented in a Technical Charter for the project, and execute the Project Contribution Agreement transferring the project's assets to the LF.

Incubation

- Have at least three organizations actively contributing to the project.
- Have a defined Technical Steering Committee (TSC) with a chairperson identified, with open and transparent communication.
- Have reached a minimum of 500 stars on GitHub.
- Have achieved and maintained an OpenSSF Best Practices Badge Program (Silver).

Graduate

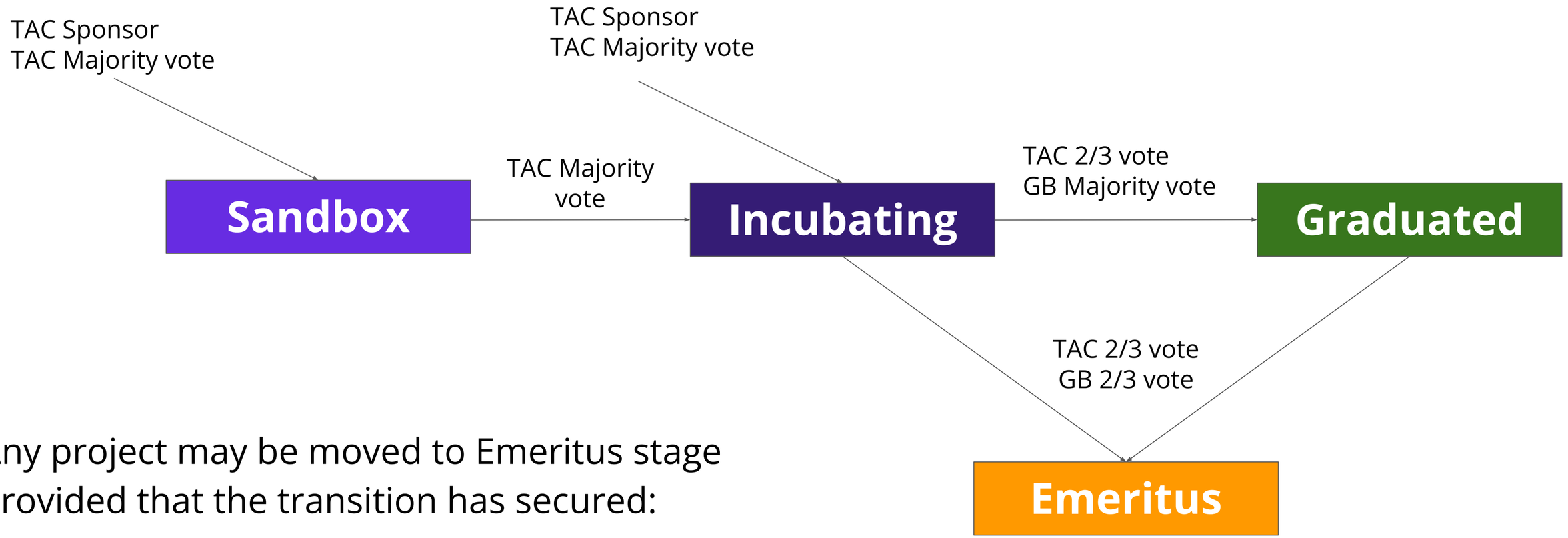
- Have a healthy number of code contributions from at least five organizations.
- Have reached a minimum of 1000 stars on GitHub.
- Have achieved and maintained an OpenSSF Best Practices Badge Program (Gold).
- Have demonstrated a substantial ongoing flow of commits and merged contributions for the past 12 months*.
- Have completed at least one collaboration with another LF AI & Data hosted project

Transitioning from Incubation to Graduation

- The TAC will undertake an annual review of all projects
- Projects in the Sandbox Stage are generally expected to move to Incubation within 12 months from joining the Foundation following an evaluation by the TAC committee
- Projects in the Incubation Stage are generally expected to move to Graduation within 12-18 months from joining the Foundation following an evaluation by the TAC committee

Projects can be provided with an extension of time in their stage (up to the discretion of the TAC)

Project Lifecycle



Any project may be moved to Emeritus stage provided that the transition has secured:

- TAC affirmative ⅔ vote
- GB affirmative ⅔ vote

Process for proposing a project for hosting in LF AI & Data

1. Decide on a date to present to the TAC and request incubation
2. Ensure that your project implements these recommendations
3. Submit a formal request to incubate the project via a GH PR
4. Prepare deck and share with ED about 10 days prior to the presentation
5. Present to the TAC and get approval
6. Onboard the project with the LF AI & Data team and integrate the project with our service
7. Announce the project becoming hosted in LF AI & Data

Incubation Benefits by Levels

Sandbox

- Neutral hosting of the project's trademark and assets by LF AI & Data.
- Appointment of a TAC member as a project sponsor
- LF AI & Data blog post or similar announcing the project's hosting in the Foundation
- Right to refer to the project as an "LF AI & Data Sandbox Project," and to display the LF AI & Data logo on the project's code repository and web properties
- An initial and regularly scheduled license scan of the project's codebase
- Ongoing source code security scans and reports
- Infrastructure support includes mailing lists, wiki space, slack channel, etc.
- Marketing, communication, and PR support are limited to significant announcements.
- Access to the LFX platform.
- Support of the Foundation staff who are eager to help with the project.

Incubation

Sandbox benefits plus:

- Right to refer to the project as an "LF AI & Data Incubation Project," and to display the LF AI & Data logo on the project's code repository.
- Creative and artwork support covering website, logo, and other required creative work.
- Marketing, communication, and PR support, including project promotion via blog posts, social media, and LF AI & Data website.
- Access to the Bevy platform for community-hosted events.

Graduate

Incubation benefits plus:

- LF AI & Data blog announcement or similar announcing the project graduation, including promotion activities.
- Graduation stage projects may receive support as determined by the Governing Board.
- Right to refer to the project as an "LF AI & Data Graduation Project," and to display the LF AI & Data logo on the project's code repository.
- Voting seat on the TAC.
- Advanced IT infrastructure support (pending board approval).
- Additional ecosystem development opportunities include training courses, certification development, and conformance programs (pending board approval).

LF AI & Data Membership Benefits

LF AI & Data is positioned to bolster cross-company collaboration and interoperability with our neutral IP zone, rich contributor programs, and most importantly, the trust end users place in us.

Organizations join LF AI & Data to take an active role in supporting the growth and evolution of the Open Source AI & Data ecosystem

Support our mission, programs, and the community by helping fund services that hosted projects rely on

Demonstrate thought leadership by participating in a wide reaching networking and marketing programs

Be part of **defining and maintaining technologies** that are at the forefront of the industry

Host projects with us and benefit from the services & support we offer our hosted projects to increase their adoption and footprint in the ecosystem

Provide technical leadership to the community via the TAC, its various sub-committees and efforts

Receive greater insight into the foundation's strategy, efforts, projects and initiatives through increased engagement with the ED, staff and committees' leads

LF AI & Data Membership Benefits

Marketing Amplification & Brand Awareness

Community Engagement

Thought & Tech Leadership Across Key Technologies

Marketing amplification and brand awareness

Broaden your reach and awareness in the community with LF AI & Data marketing programs. As a member you can participate in:

LF AI & Data Outreach Committee

Participate in the marketing committee monthly to engage with your peers in the cloud native space.

LF AI & Data Summit

LF AI & Data organizes its own community's open source developer summits focused on open source AI and Data technologies.

LF AI & Data Online Programs

Showcase your organization's open source technology by educating new and existing community members about best practices, trends, and new technologies

Public Relations & Analyst Relations Support

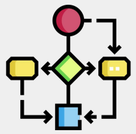
LF AI & Data supports members with analyst reports and research highlights.

LF AI & Data Blog

Showcase your thought-leadership and industry commentary, as well as share technical walkthroughs for LF AI & Data projects here.

Community Engagement

LF AI & Data is a constellation of open source projects. Our members leverage many efforts to engage with our project's ecosystems and share their stories. As a member you can participate in:

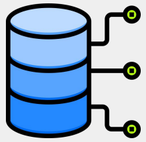


ML Workflow and Interoperability

Design a reference architecture for end-to-end ML Workflow, provide a reference implementation using hosted projects, and facilitate integration and interop across projects.

Trusted AI

A global group working on policies, guidelines, tools and use cases by industry to ensure the development of trustworthy AI systems and processes to develop them continue to improve over time.



DataOps

Share best practices, and bring technology awareness around DataOps across industries and enable collaboration across LF AI & Data member companies, projects, and external participants.

BI & AI

Integrate the power of AI and BI to make it CI (Cognitive Intelligence) by combing the speed machines accelerate (AI) with the direction intuited by human insight (BI).

LF AI & Data Day

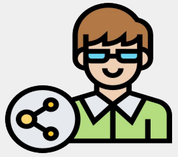
A 1-day event organized multiple times per year focused on specific technical projects and collaboration ongoing in LF AI & Data

Thought leadership

Members of the LF AI & Data can network and help shape the open source AI and Data market. As a member you can participate in:

Governing Board Participation

Participate in elections to serve on the Governing Board to oversee the vision of LF AI & Data and work with the TAC.



Technical Advisory Council Leadership

LF AI & Data TAC provides technical leadership to the open source AI and Data community, accepts projects into incubation, graduates projects and launches new efforts and collaborations.

Engagement with other Foundations and Industry Consortia

Participate in establishing collaborations with other umbrella foundations under the LF and broadly with other industry consortia organizations.

Interactive Landscape Placement

Our landscape is a comprehensive view of all open source AI and Data critical projects in the ecosystem. Anyone looking to adopt open source AI and Data projects comes here to review what technologies they should assess and adopt.

LF AI & Data Annual Dues

	Not Yet a Linux Foundation Member	Existing Linux Foundation Member
Premier Member	$120,000	$100,000
General Member	5,000 employees +: $45,000 2,000 - 4,999: $30,000 500 - 1,999: $25,000 Up to 499 employees: $10,000	5,000 employees +: $25,000 2,000 - 4,999: $15,000 500 - 1,999: $10,000 Up to 499 employees: $5,000
Associate Member	Free Limited to academic and nonprofit institutions respectively and requires approval by the Governing Board	

Premier Membership

LF AI & DATA
PREMIER MEMBER

Highest tier of membership – For organizations contributing heavily to open source AI and Data, and bringing their own projects to be hosted at the Foundation. They work in concert with the Foundation team members. These companies want to take the most active role in enabling the open source AI and Data ecosystem.

Premier members are eligible to:

(Enjoy all the benefits of General level, plus;)

- **Hold one (1) guaranteed seat** on the LF AI & Data Governing Board + one alternate (1) representative
- **Appoint one (1) voting representative in any subcommittees** or activities of the LF AI & Data Governing Board
- Receive greater **insight into LF AI & Data strategy and projects** through engagement with the LF AI & Data leadership team. Premier level members have the unique opportunity to **customize their experience** with LF AI & Data. The team will make themselves available to help achieve your strategic goals. We can help with guidance in open source contributions, new market creation, and/or open source project donation. **Have ideas? Just ask!**
- Enjoy most prominent placement in displays of membership including website, landscape and marketing materials.
- Create an individualized press release upon membership announcement with the LF AI PR team.

General Membership

LF AI & DATA
GENERAL MEMBER

Targeted for organizations that want to support of LF AI & Data and our mission. Organizations that join at the General level are deeply committed to using open source technology, helping LF AI & Data grow, voicing the opinions of their customers, and giving back to the community.

General members are eligible to:

- **Participate in elections** between **other General members** to appoint one (1) representative to the **LF AI & Data Governing Board**. Three (3) total General representatives will be elected to represent all General members. Voice your opinions amongst the leaders in the industry and help determine the strategic direction of LF AI & Data.
- Receive **greater insight into LF AI & Data strategy and project roadmaps** through increased engagement with the LF AI & Data Executive Director and staff.
- Create an announcement upon membership announcement with the LF AI & Data PR team.
- Participate in all Marketing, Community, Thought Leadership opportunities.
- Opportunity to **host "LF AI & Data Day"**, including on-demand webinars and livestream.
- Receive **discounts** on LF AI & Data event sponsorships.
- Demonstrate your support for LF AI & Data by displaying your logo on the LF AI & Data website, landscape and in marketing materials.

Associate Membership

LF AI & DATA
ASSOCIATE MEMBER

The Associate membership is a free complimentary membership limited to academic and non-profit institutions.

Associate members are eligible to:

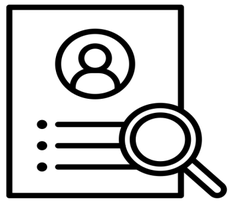
- Participate in all Marketing, Community, and Thought Leadership opportunities.
- Identify your organization as a member and display your logo on the LF AI & Data website, landscape and in marketing materials.
- Feature your organization in the quarterly new members announcement.
- Receive discounts on LF AI & Data events' sponsorship.

Specific opportunities for associate members:

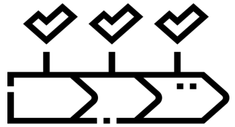
- Practical experience working with open source foundations
- Visibility into and within the ecosystem
- Internships opportunities with commercial members
- Collaboration opportunities
- Publication opportunities based on R&D conducted with projects
- Open source working experience for all participants
- Access to LF AI & Data events

Challenges and Trends

Global Open Source Trends



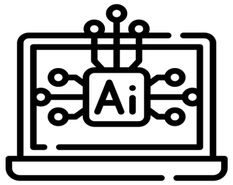
Training & Cert, Talent Shortage



Emphasis on SW Supply Chain

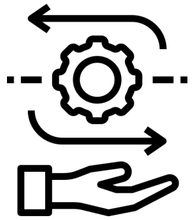
Rise of Adoption of OSPOs

Importance of Software Security



Infusing AI and ML in products and services

Focused Efforts on Ethical AI Practices

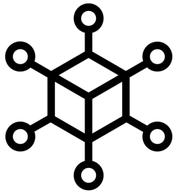


Criticality of OSS to Digital Transformation

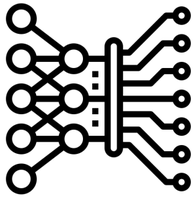
Accelerated OSS Adoption in Governments

AI Challenges and Opportunities

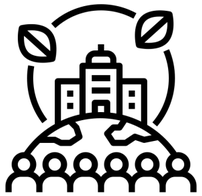
Trusted and Responsible AI



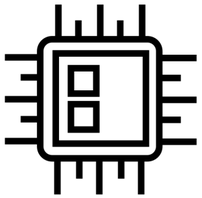
AI on edge devices



Federated learning



Social Impact



Specialized Hardware

Availability of Skilled Talent

Scalability, Efficiency

Policies, Regulations & Standards

User Acceptance, Building Trust

Data Challenges and Opportunities

Security

Integration

Labeling

Privacy

Storage

Quantity

Quality

Ethical

Diversity

Accessibility

Bias

Lineage

Licensing

Governance

Regulations

Standards

Legislations

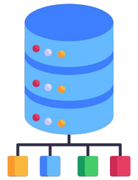
Compliance

Deriving value

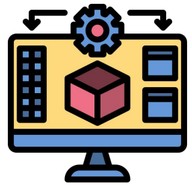
Next Frontier

Open Source Generative AI, Open Source Models and LLMs

Core LLM elements to contribute to the open source AI community's collective knowledge to foster collaboration, and accelerate progress



Datasets

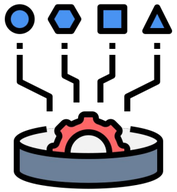


Model Architecture

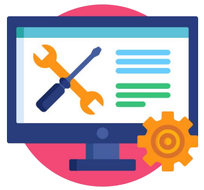
Preprocessing Code

Training Code

Evaluation Metrics & Benchmark



Model Weights and Parameters



Supporting Libraries & Tools

Documentation

Breaking it down

Open sourcing high-quality datasets allows others to reproduce and validate research, compare models, and develop new approaches.

Open sourcing the architecture of the model (structure, layers, and connections) will help others understand and build upon your work, leading to advancements and innovation in the field.

Open sourcing custom code or scripts for data preprocessing tasks (cleaning, transforming, or augmenting the dataset) can be valuable to the community to understand your data preparation techniques and reproduce your results.

Open sourcing custom code or scripts for data preprocessing tasks (cleaning, transforming, or augmenting the dataset) can be valuable to the community to understand your data preparation techniques and reproduce your results.

Open sourcing evaluation metrics and benchmarks (including the implementation of the optimization algorithms, hyperparameter tuning) will allow others to replicate your experiments, compare results, and further refine the model.

Open sourcing the trained weights and parameters of your AI model will enable others to use your pre-trained model for inference or fine-tuning on their own datasets, saving them the computational resources required for training from scratch.

Open sourcing custom libraries, frameworks, or tools that facilitate AI model creation, training, or deployment, can be immensely valuable to enable the community to benefit from your codebase, contribute improvements, and build upon your work.

Open sourcing documentation and user guides can help others understand and use your AI model effectively and will encourage adoption and facilitates collaboration.

Trends

1. **Model Size**
   LLMs have been steadily increasing in size and complexity. The shift from GPT-2 to GPT-3 saw a 100x increase in model parameters, and 6x increase from GPT-3 to GPT-4, enabling more nuanced language processing and higher-quality outputs.
2. **Pre-training and Fine-Tuning**
   Pre-training on large-scale datasets followed by fine-tuning on specific tasks has become a common approach in LLM development. This two-step process enhances model performance and allows for transfer learning across domains.
3. **Few-shot and Zero-shot Learning**
   LLMs have shown promising capabilities in few-shot and zero-shot learning, where they can perform tasks with minimal or no task-specific training examples. This trend has implications for rapid prototyping and reducing data requirements.
4. **Multimodal LLMs**
   There is an increasing focus on developing multimodal LLMs that can process and generate text in conjunction with other media types, such as images, audio and video. These models enable more comprehensive and interactive AI experiences.

Challenges

1. **Ethical Concerns**
   LLMs raise ethical concerns related to biases, misinformation amplification, and malicious use. Ensuring fairness, transparency, and responsible deployment of these models is crucial.
2. **Data Privacy**
   LLMs require large amounts of data for training, raising concerns about user privacy and data protection. Striking a balance between data access and privacy is an ongoing challenge.
3. **Dataset Limitations**
   Open source models rely on publicly available datasets, which may be limited in terms of diversity, quality, and potential biases. The availability of diverse and representative datasets is essential for training robust and unbiased models.
4. **Resource Intensiveness**
   Training and deploying large-scale LLMs demand significant computational resources and energy consumption, hindering accessibility and sustainability.
5. **Reproducibility and Benchmarking**
   Comparing and reproducing results across different open source LLMs can be challenging due to differences in architectures, training data, and evaluation metrics. Establishing standardized benchmarks and reproducibility guidelines are essential for transparent evaluation and advancement in the field.

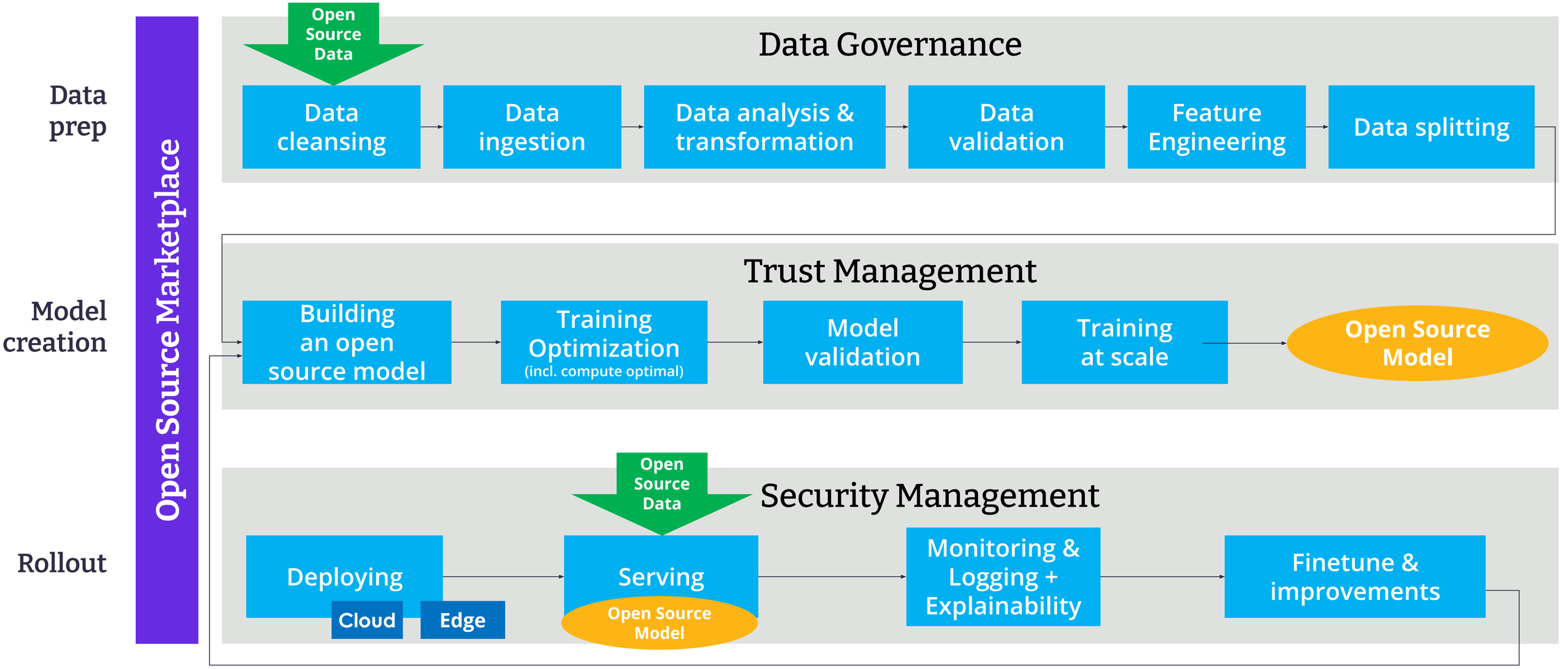
Closed LLMs

- **Lack of Accessibility**
  Closed source models are developed and owned by specific company, and access to these models may be restricted or available only through licensing agreements. This limited accessibility can restrict the ability of researchers and developers to utilize and modify the models.
- **Lack of Transparency**
  Closed source models often lack transparency in terms of their architecture, training data, and fine-tuning techniques. The inner workings and details of these models are not openly shared, making it challenging to understand how they make predictions or generate outputs. This lack of transparency can limit researchers' ability to analyze and improve the models.
- **Lack of Customizability**
  Closed source models may offer limited customization options. The ability to fine-tune or modify the models for specific tasks or domains may be restricted or entirely unavailable. This can limit the flexibility and adaptability of the models to specific use cases.
- **Intellectual Property Rights**
  Closed source models are typically protected by intellectual property rights, such as copyrights or patents. This means that using or distributing these models without proper authorization may infringe on these rights. This can limit the freedom to use and distribute the models for various purposes.

Opportunities – Benefits Open Source LLMs

- **Research and Innovation**
  Open source LLMs provide researchers with valuable resources for exploring new techniques, developing novel applications, and advancing the field of natural language processing.
- **Collaboration and knowledge sharing**
  Open source LLMs promote collaboration by allowing developers to build on top of existing models. Researchers and developers can build upon existing models, share improvements, and collectively advance the field.
- **Transparency**
  Open source LLMs encourage transparency and accountability by allowing researchers to inspect the models and data used for training.
- **Democratize access**
  Open source LLMs help to democratize access to state-of-the-art natural language processing technology enabling developers and organizations with limited resources to leverage state-of-the-art models for their applications.
- **Federated LLMs**
  Open source LLMs support the emergence of decentralized and federated LLMs, allowing for increased privacy and data control.

With 48 Hosted Technical Projects, LF AI & Data offers all required open source software and elements required to build and manage an end-to-end ML Workflow



LF AI & Data Role

1. Host open source LLMs
2. Harmonize the existing separated open source LLMs, build bridges, foster collaborations
3. Support the open sourcing of model elements required for a true open source experience
4. Present a framework to classify the openness of a given LLM based on the number of its elements that are made available under an open source license

The Linux Foundation and its umbrella LF AI & Data will be the vendor neutral, not-for-profit organization to host open source models and LLMs and all their associated building blocks.

We will continue supporting the AI and open source communities in new capacities, fostering collaboration, and accelerating development and innovation.

Join us, become a member!

Incubate your project

info@lfaidata.foundation

OLD FOR REFERENCE ONLY

Active and growing developer community

Jan 1- Dec 31, 2020

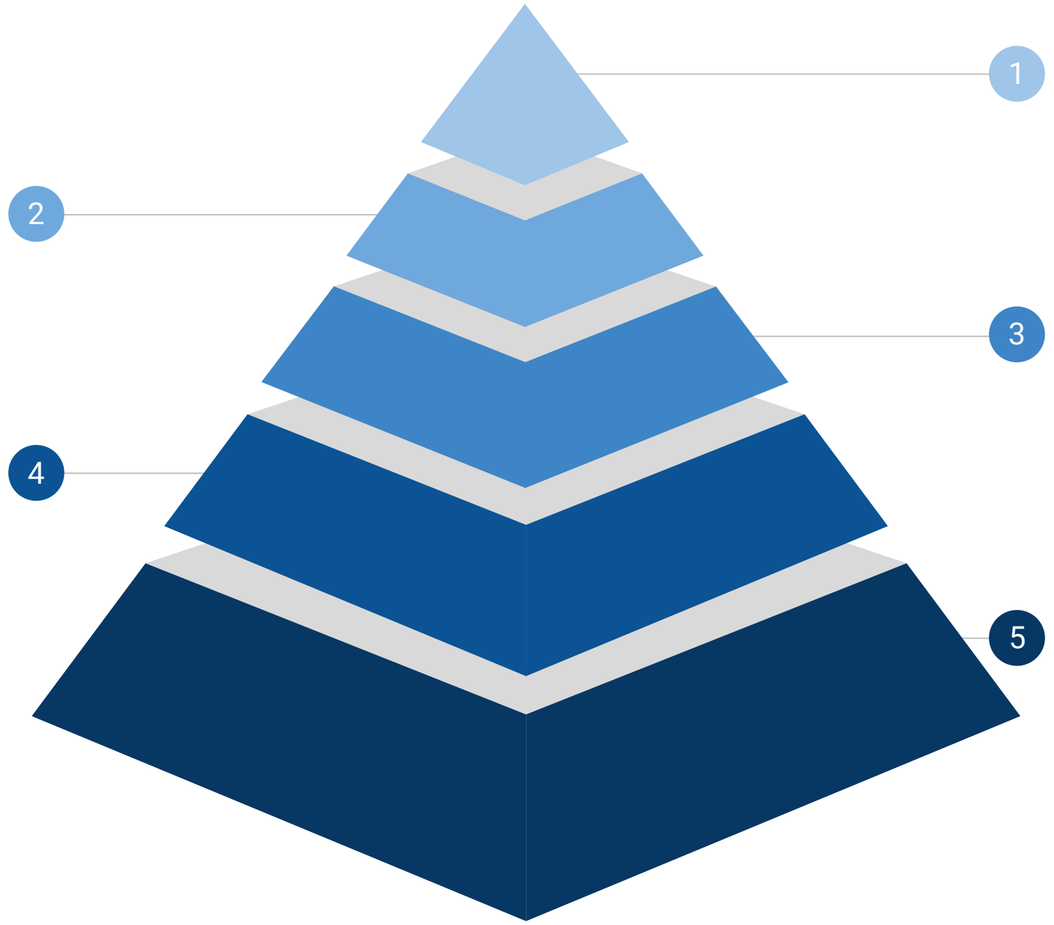
8.92K	100.03K	358	6.93K	34.65K	22.54K
Contributors	Commits	Repositories	Emails	PRs/Changesets	Total issues

Source: https://insights.lfx.linuxfoundation.org/projects/lfai

But with challenges

Managing projects assets
As project grows, it is unclear who should own legal and administrative tasks that are essential to the health of the project

Governance challenges
No formal governance, governance favors the creators of the project, project backer dominating dev and governance



Glass ceiling for project investment and adoption
in the absence of a fair governance and safe haven for the project's assets

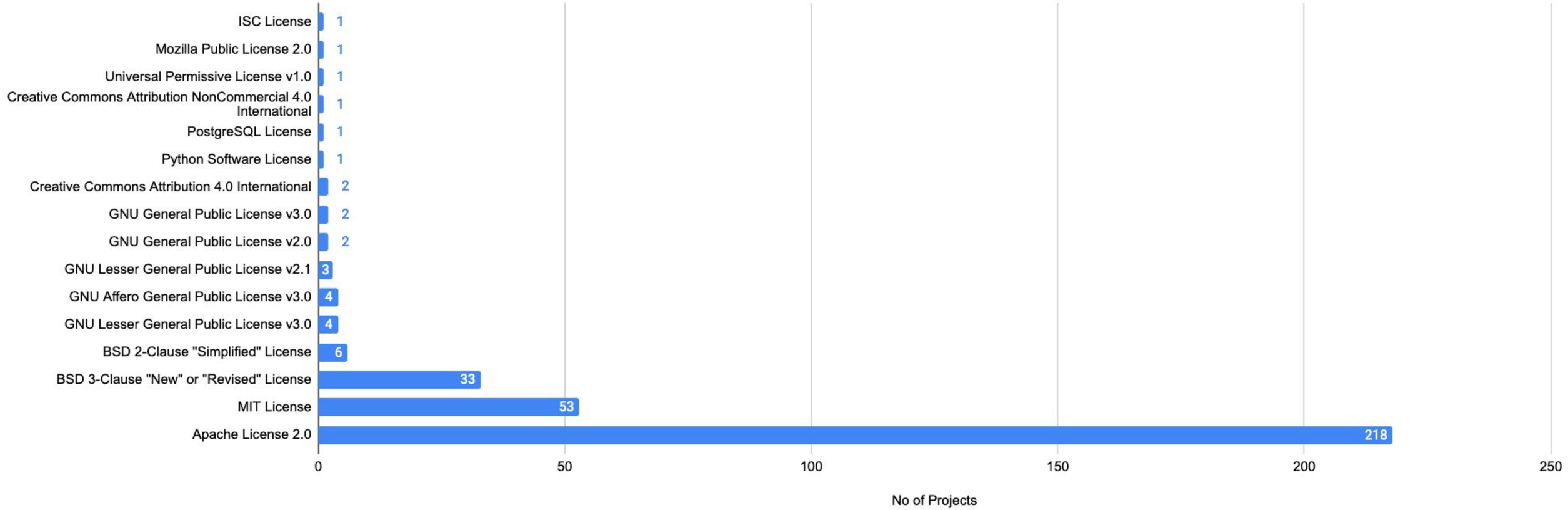
Developed internally for specific product requirements, highly specialized
The open source spinoff is a result of wanting to build an ecosystem and collaborate on building a platform

Fragmentation
Lack of integration
Lack of cross pollination

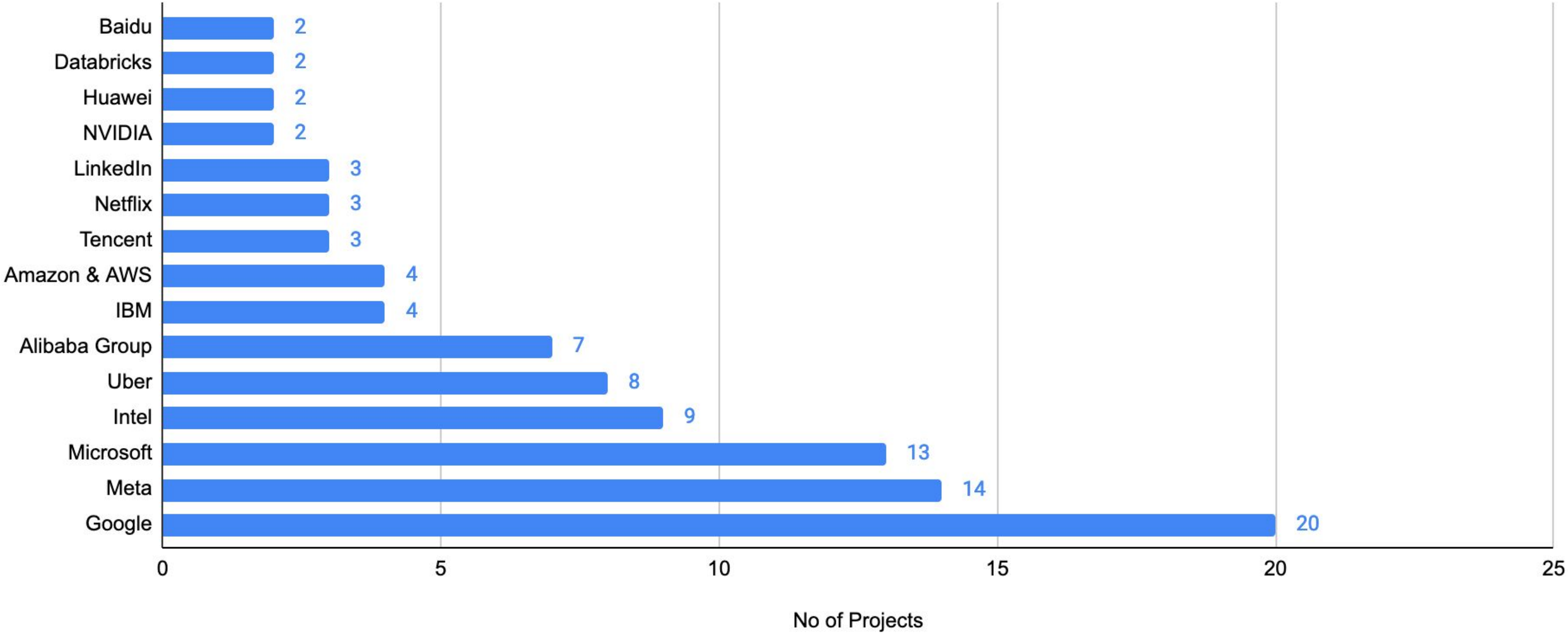
Why host your project in LF AI & Data?

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.	**Market services** We offer a wide range of marketing services to increase project awareness, project adoption, and number of contributors.	**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.
Design and aesthetics Our in-house team provides graphic design resources for new logos, websites, and website refreshes or enhancements.	**Program management** We have collectively 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 collaborator license agreements, crowdfunding, member engagement, and more.	

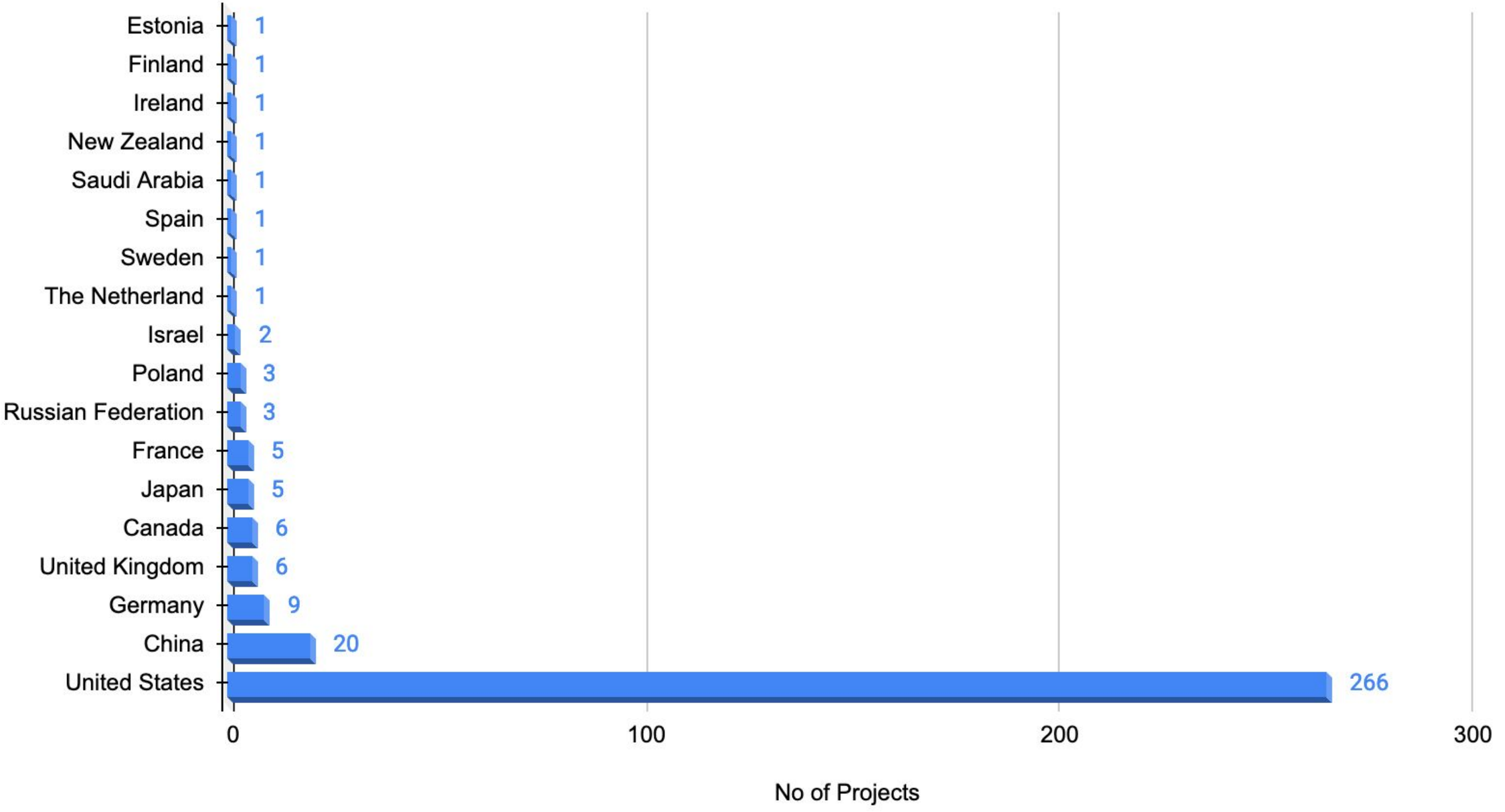
No of Projects per License



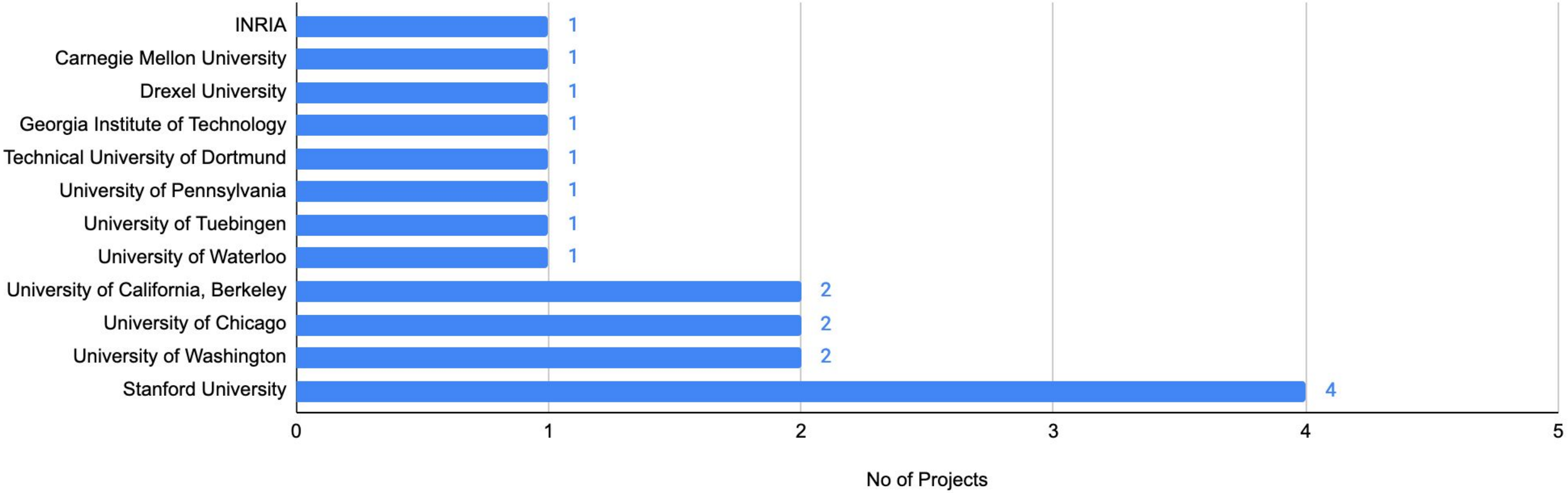
No of Projects per Company



No of Projects per Originating Country



Founding Universities



Getting Involved in LF AI & Data

Getting involved with technical projects

- Participate in the development: Review and submit patches, report bugs, request features, test, etc.
- Contribute to project documentation
- Join the projects' mailing lists and participate in the discussions
- Attend developer events for LF AI & Data projects
- Provide your testing and deployment feedback via appropriate project channel
- Start a local User Group Meetup

Getting involved with the Technical Advisory Council

- Support TAC leadership in inviting speakers to present for the LF AI & Data technical community
- Attend TAC Bi-weekly calls, participate in the discussions
- Share success stories, opportunities and challenges in relation to open source AI with the broader technical community
- Identify opportunities for collaboration on common interests and initiatives, seek input from peers
- Support technical leadership for integration and harmonization efforts with other open source projects
- Support TAC in hosting and sponsoring intra-project and inter-project developer events
- Support TAC in evaluating new projects for incubating; recommend new projects
- Support TAC Chair working with the GB to highlight opportunities and needed resources in support of hosted projects
- Participate in the ML Workflow effort aiming to provide a reference workflow and support integration across LF AI & Data projects
- Participate in the effort to enable collaboration with other LF umbrella foundations and external communities

TAC specific resources: mailing list, call calendar, and wiki.

Getting involved with the Outreach Committee

- Promotion of project updates, releases, and news via LF AI & Data social media accounts
- Attend Outreach Committee meetings and participate in ongoing activities
- Publish use cases, case studies, white papers, and deployment insights
- Marketing and PR support for demos at meetups and events
- Host vendor neutral content via LF AI blog site
- Get support for artwork, web site, content creation, etc., for LF AI & Data projects
- Contribute to the LF AI & Data landscape, promote in talks
- Coordination at events, speaking proposals, booth attendance, demos, etc.
- Volunteer to host an LF AI & Data Day and developer events
- Identify speaking opportunities and help secure speakers from the LF AI & Data community
- Help secure user stories about LF AI & Data based deployments
- Support LF AI & Data marketing and PR staff

Getting involved with the LF AI & Data Committees: BI & AI, DataOps, ML Workflow & Interop

- Join the committee as a representative of your company
- Join the conference calls and contribute to ongoing efforts
- Promote the work of the committee, invite your collaborators to participate
- Contribute to the landscape, promote in talks
- Coordination at events, speaking proposals, booth attendance, demos, etc.
- Subscribe to the mailing list and participate in discussions
- Host your related projects in the LF AI & Data Foundation
- Help secure user stories about LF AI & Data based deployments
- Invite prominent researchers and developers in the space to speak to the community