

# Meeting of the LF AI & Data Technical Advisory Council (TAC)

December 1, 2022

 LF AI & DATA

# Antitrust Policy

- › Linux Foundation meetings involve participation by industry competitors, and it is the intention of the Linux Foundation to conduct all of its activities in accordance with applicable antitrust and competition laws. It is therefore extremely important that attendees adhere to meeting agendas, and be aware of, and not participate in, any activities that are prohibited under applicable US state, federal or foreign antitrust and competition laws.
- › Examples of types of actions that are prohibited at Linux Foundation meetings and in connection with Linux Foundation activities are described in the Linux Foundation Antitrust Policy available at <http://www.linuxfoundation.org/antitrust-policy>. If you have questions about these matters, please contact your company counsel, or if you are a member of the Linux Foundation, feel free to contact Andrew Updegrove of the firm of Gesmer Undergone LLP, which provides legal counsel to the Linux Foundation.

# Recording of Calls

## Reminder:

TAC calls are recorded and available for viewing on the [TAC Wiki](#)

# Reminder: LF AI & Data Useful Links

- › Web site: [lfaidata.foundation](https://lfaidata.foundation)
- › Wiki: [wiki.lfaidata.foundation](https://wiki.lfaidata.foundation)
- › GitHub: [github.com/lfaidata](https://github.com/lfaidata)
- › Landscape: <https://landscape.lfaidata.foundation> or <https://l.lfaidata.foundation>
- › Mail Lists: <https://lists.lfaidata.foundation>
- › Slack: <https://slack.lfaidata.foundation>
- › Youtube: <https://www.youtube.com/channel/UCfasaeqXJBCAJMNO9HcHfbA>
- › LF AI Logos: <https://github.com/lfaidata/artwork/tree/master/lfaidata>
- › LF AI Presentation Template: [https://drive.google.com/file/d/1eiDNJvXCqSZHT4Zk\\_-czASlz2GTBRZk2/view?usp=sharing](https://drive.google.com/file/d/1eiDNJvXCqSZHT4Zk_-czASlz2GTBRZk2/view?usp=sharing)
  
- › Events Page on LF AI Website: <https://lfaidata.foundation/events/>
- › Events Calendar on LF AI Wiki (subscribe available): <https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=12091544>
- › Event Wiki Pages: <https://wiki.lfaidata.foundation/display/DL/LF+AI+Data+Foundation+Events>

# Agenda

- › Roll Call (2 mins)
- › Approval of Minutes from previous meeting (2 mins)
- › Xtreme1 new Sandbox Proposal (40 min)
- › LF AI General Updates (2 min)
- › Open Discussion (2 min)

# TAC Voting Members - Please note

Please ensure that you do the following to facilitate smooth procedural quorum and voting processes:

- Change your Zoom display name to include your First/Last Name, Company/Project Represented
  - example: Nancy Rausch, SAS
- State your First/Last Name and Company/Project when submitting a motion
  - example: First motion, Nancy Rausch/SAS

# TAC Voting Members

Note: we still need a few designated backups specified on [wiki](#)

## Member Representatives (8 out of 16 required for quorum)

Member Company or Graduated Project	Membership Level or Project Level	Voting Eligibility	Country	TAC Representative	Designated TAC Representative Alternates
4paradigm	Premier	Voting Member	China	Zhongyi Tan	
Baidu	Premier	Voting Member	China	Ti Zhou	Daxiang Dong, Yanjun Ma
Ericsson	Premier	Voting Member	Sweden	Rani Yadav-Ranjan	
Huawei	Premier	Voting Member	China	Howard (Huang Zhipeng)	Charlotte (Xiaoman Hu) , Leon (Hui Wang)
Nokia	Premier	Voting Member	Finland	@ Michael Rooke	@ Jonne Soininen
OPPO	Premier	Voting Member	China	Jimmy (Hongmin Xu)	
SAS	Premier	Voting Member	USA	*Nancy Rausch	JP Trawinski
ZTE	Premier	Voting Member	China	Wei Meng	Liya Yuan
Adversarial Robustness Toolbox Project	Graduated Technical Project	Voting Member	USA	Beat Buesser	
Angel Project	Graduated Technical Project	Voting Member	China	Bruce Tao	Huaming Rao
Egeria Project	Graduated Technical Project	Voting Member	UK	Mandy Chessell	Nigel Jones, David Radley, Maryna Strelchuk, Ljupcho Palashevski, Chris Grote
Flyte Project	Graduated Technical Project	Voting Member	USA	Ketan Umare	
Horovod Project	Graduated Technical Project	Voting Member	USA	Travis Addair	
Milvus Project	Graduated Technical Project	Voting Member	China	Xiaofan Luan	Jun Gu
ONNX Project	Graduated Technical Project	Voting Member	USA	Alexandre Eichenberger	Prasanth Pulavarthi, Jim Spohrer
Pyro Project	Graduated Technical Project	Voting Member	USA	Fritz Obermeyer	

\*Current TAC Chairperson

# Minutes approval



# Approval of November 17, 2022 Minutes

Draft minutes from the November 17 TAC call were previously distributed to the TAC members via the mailing list

## **Proposed Resolution:**

- › That the minutes of the November 17 meeting of the Technical Advisory Council of the LF AI & Data Foundation are hereby approved.

# Proposal to Host Xtreme1 in LF AI & Data

The Next GEN Open-source Platform for Multisensory Training Data.

---

Dr. Alex S. Liu

[alex@basic.ai](mailto:alex@basic.ai)

# Agenda

- 1** Problem statement
- 2** Introduction to Xtreme1 and key features
- 3** Xtreme1 technology deep dive
- 4** Incubation request



# Problem statement

---

As training data becomes the new bottleneck of AI model development, Data-Centric MLOps is trending now.

# What we learned from 6 years labeling business

UBS Global research report: AI engineers now spend 70%-90% of their time on training data.



## The Repetition in D&M Fine-tune

Fine-tuning data annotation schema and modeling is highly repetitive and costly



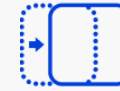
## Data Complexity

Training data annotation is much complex than before



## Data Quality Issue

Data quality is hard to control comprehensively from accuracy, consistency, diversity, quantity and distribution



## Data Drift

Data keeps changing from experimental lab to physical world's application



## Data Security

Public cloud SaaS is becoming unacceptable



## High Cost

Data Labeling is expensive



## Lack of Transparency

Outsourcing labeling process lacks transparency and trade-off between cost accuracy is unmanageable



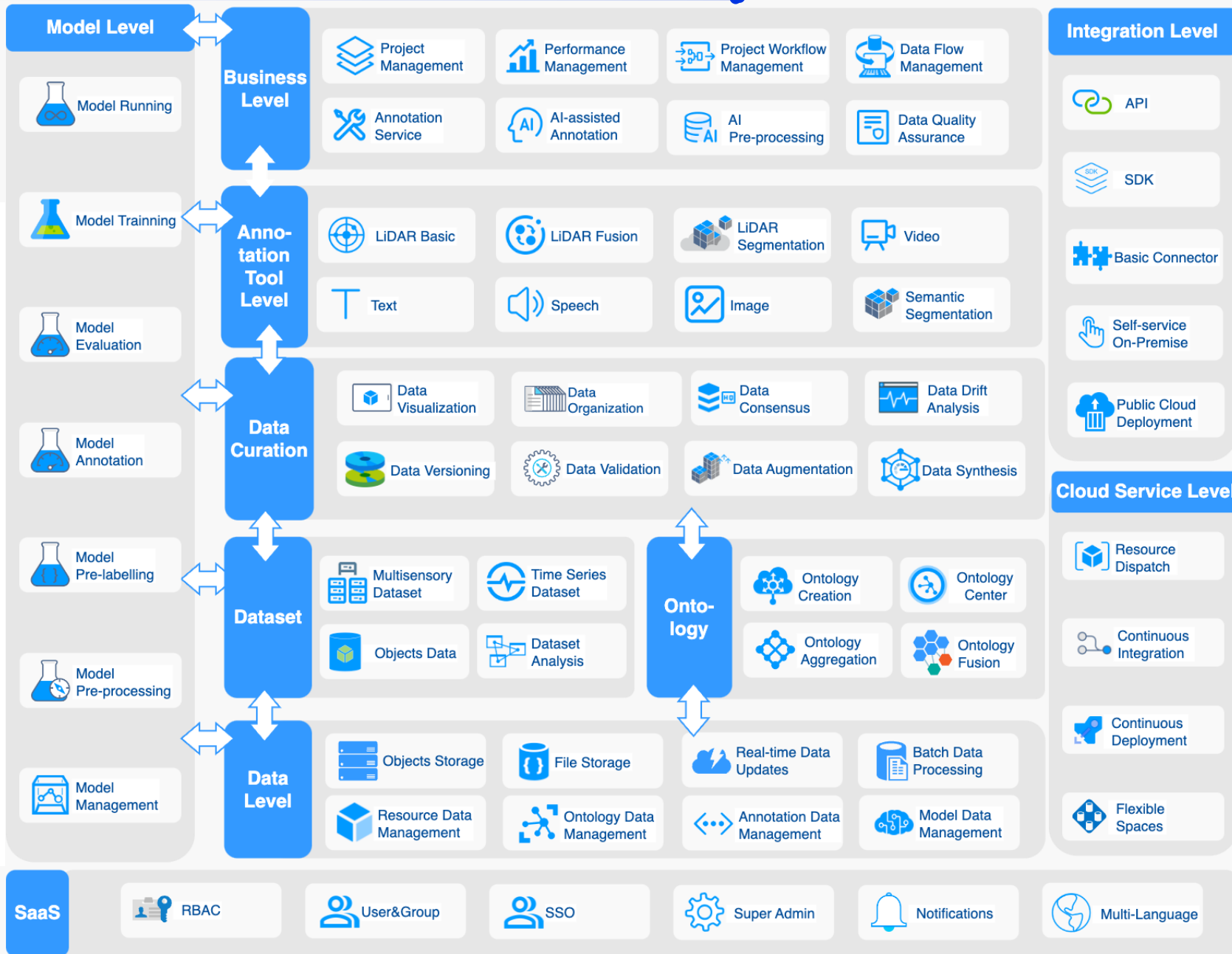
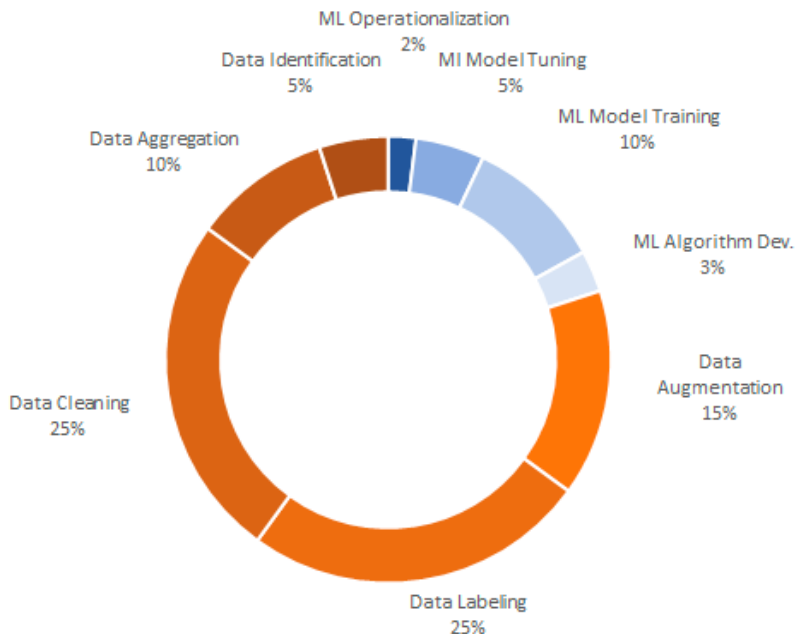
## Fragmented ML tools

Switching between tools can be costly and limiting

# Our solution: Data-centric MLOps

Goal: To save the every minute that you spend on processing training data.

80% of time spent for Machine Learning Projects is allocated to Data related tasks



Source: Cognylitica; Factordaily



# Introduction to Xtreme1

---

Xtreme1 is the next generation open-source platform for multisensory training data.

# Xtreme1 Open Source



<https://github.com/basicai/xtreme1>

## Motivations

1. AI is born for open source
2. MLOps infras requires global user feedbacks
3. Open source helps proof of compliance with data security and privacy
4. Open-source multisensory training data software is missing in the market

basicai / xtreme1 Public

Edit Pins Unwatch 6 Fork 11 Starred 119

Code Issues 6 Pull requests Discussions Actions Projects Wiki Security 29 Insights Settings

main 2 branches 1 tag Go to file Add file Code

About

Xtreme1 - The Next GEN Platform for Multisensory Training Data. #3D annotation, lidar-camera annotation and image annotation tools are supported!

[www.basic.ai](http://www.basic.ai)

computer-vision annotation-tool 3d-annotation data-curation labeling-tool 3d-bounding-boxes point-cloud-labeling cuboid-annotations point-cloud-annotation

Readme Apache-2.0 license Code of conduct 119 stars 6 watching 11 forks

Releases 1

Release v0.5 Latest 8 days ago

Packages

Xtreme1



```
# wget https://github.com/basicai/xtreme1/releases/download/v0.5/xtreme1-v0.5.zip
```

```
# unzip -d xtreme1-v0.5 xtreme1-v0.5.zip
```

```
# cd xtreme1-v0.5 && docker compose up
```

- Advantages
  - 6 years of training data tool polishing
  - Rich data processing experience, involving various industries and data modalities
- Actions
  - 5000+ commits, 10+ full-time contributors
  - Keep contributing the code to improve the feature set
  - Respect all valid issues and PRs and incorporate them into

# Open Source Comparisons

	Xtreme1	LabelStudio	Labelme	CVAT	doccano	VoTT
Major Features						
Text & Speech	Y	Y	N	N	Y	N
Image & Video	Y	Y	Y	Y	N	Y
Lidar and 2/3D Fusion	Y	N	N	N	N	N
Ontology	Y	N	N	N	N	N
AI-assisted Labeling	Y	Y	N	Y	N	N
Dataset Curation	Y	Y	N	N	N	N
Multi-people Collaboration	Y	Y	N	Y	Y	Y
Open API & Models	Y	Y	N	Y	N	N



# Key Features

---

- **K1: Ontology Center**
- **K2: Annotation Suite**
- **K3: Dataset Curation**
- **K4: Model Integration**

Ontology is a perfect way of standardizing the problem definition and intra-/inter- dataset management.



## Organization

Ontology can serve as the catalogue of massive data. New dataset can be created by simply searching and merging.



## Recommendation

Good practice ontologies may be suggested before/during setting up the workflow to reduce the trial-and-error cost.



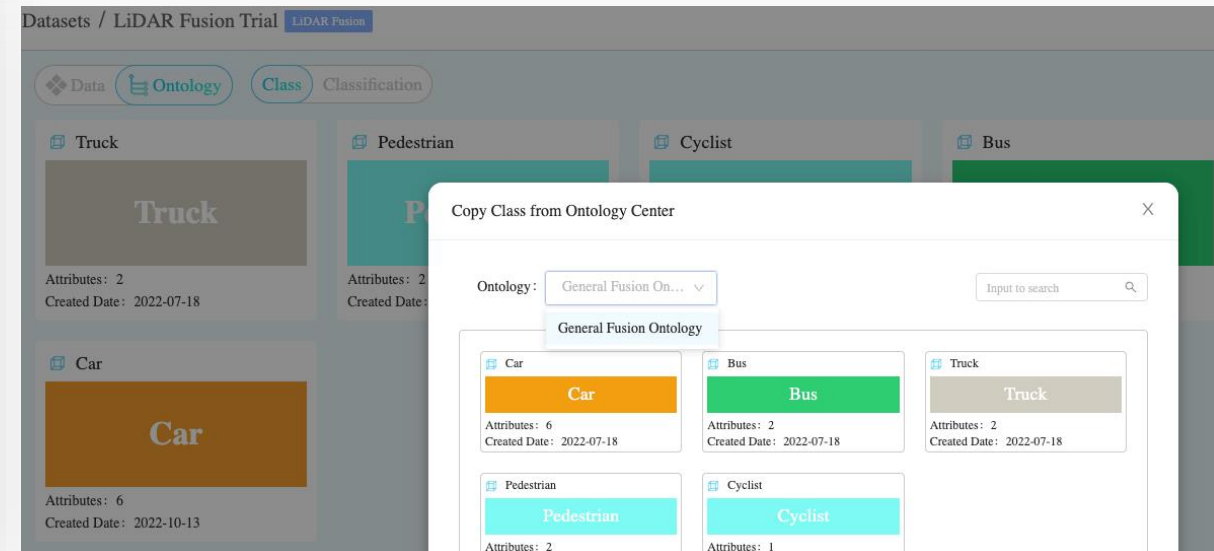
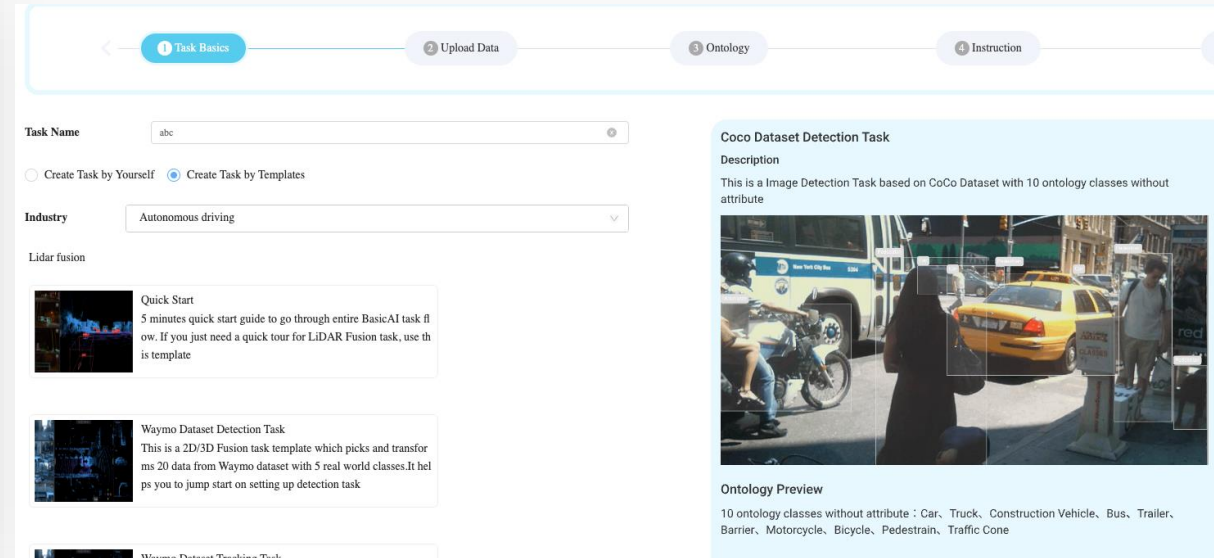
## Collaboration

Duplicate ontologies can be saved when dealing with multisensory fusion annotation tasks.



## Disambiguation

The same semantic ontology may vary its name in different projects or teams.



# K2: Annotation Suite

With the development of hardware, multiple/fused/HD sensors are widely used for various AI applications.



## LiDAR Basic

Support the 3D Lidar point cloud annotation by either single or multiple frames.



## LiDAR Fusion

Support the 2D&3D fusion annotation with automatic 3D-to-2D projection.



## LiDAR Segmentation

Support the 3D Lidar point cloud segmentation.



## Image

Support various image annotation tools, such as bbox, polygon, polyline, key points, curves, etc.



## Task Management

Flexible data distribution, member and role management.



## Performance Mgt

Multi-dimension metrics to evaluate and improve the efficiency of workforce.



## Workflow Management

Flexible workflow set up to accommodate different team size and quality requirements.

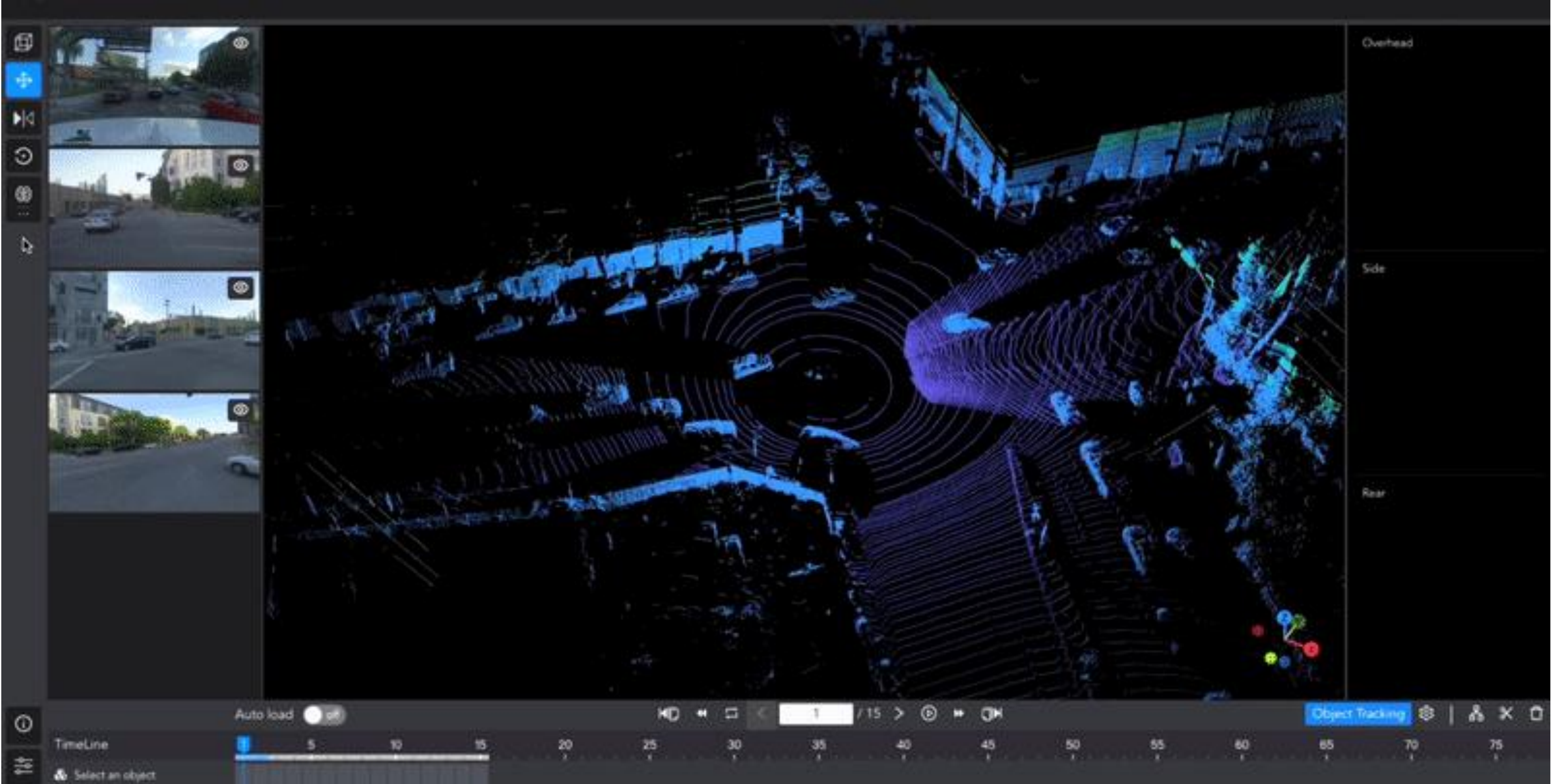


## AI-assisted Annotation

Various models are developed to improve the efficiency of semi-automatic annotations.

# K2: Annotation Suite

Xtreme1 is the world first open-source platform supporting multisensory training data with 3D model integration.



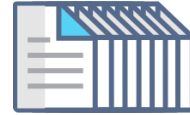
# K3: Dataset Curation

Accuracy is no longer the only metric of training data quality.



## Data Visualization

Support 2D&3D visualizations of frames, objects for fast data review and analysis.



## Data Organization

Support multiple data organization, such as clustering, sorting, filtering, tagging, batch operations.



## Data Consensus

Support human-human, human-machine comparison to improve the data consensus.



## Data Drift Alert

Track the performance of models and detect/alert the degradation in production.



## Data Versioning

Version control is essential for management of datasets varying with models, dates, and batches.



## Data Validation

Provide various comparison approaches between ground-truth and model predictions for fast model debugging.



## Data Augmentation

One stop of data augmentation, visualization and review for imbalance issues.



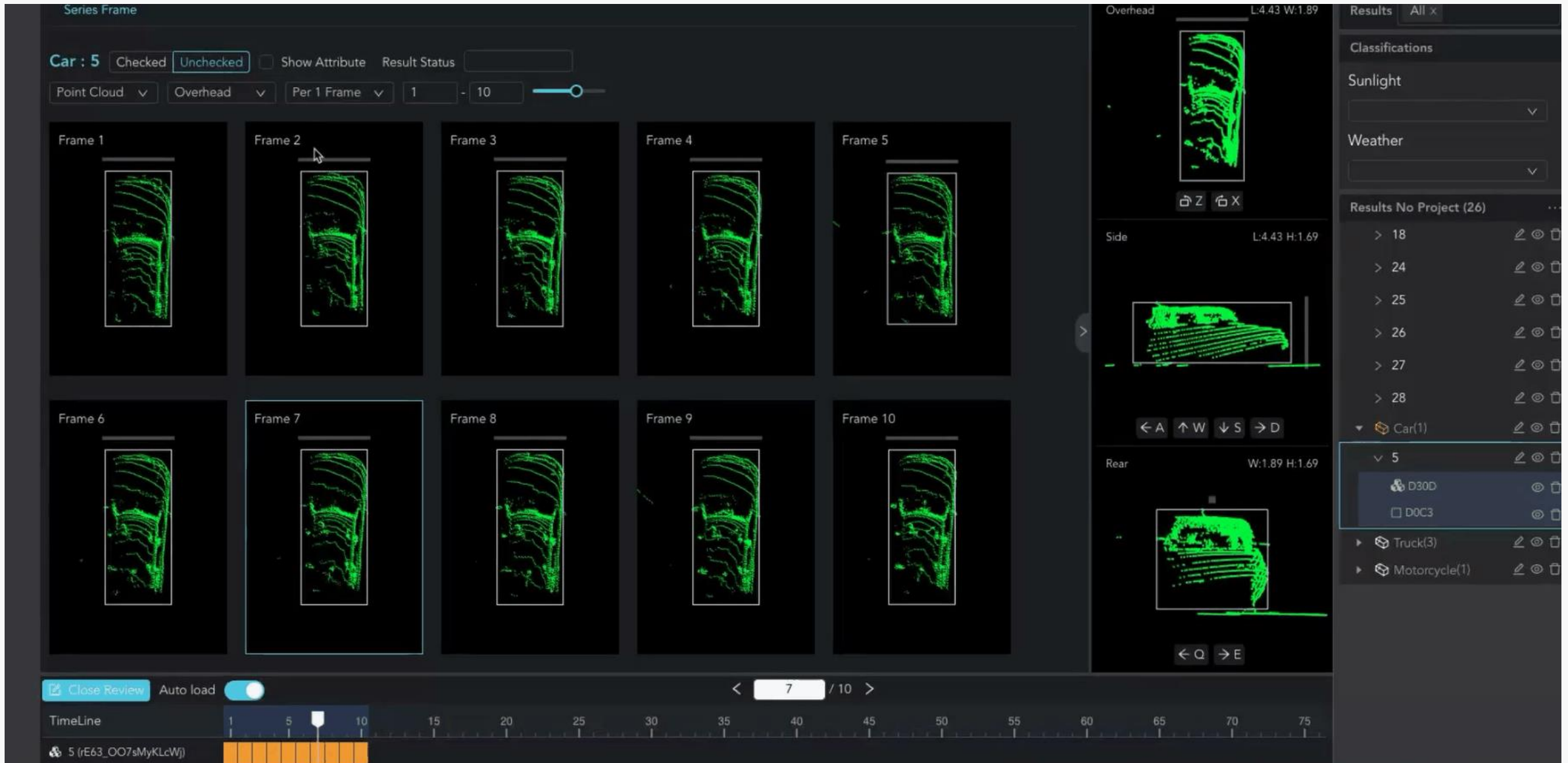
## Data Synthesis

Provide built-in and external models to generate artificial data, reduce the data cost and avoid privacy issues.



# K3: Dataset Curation

Visualization is one of the key features to evaluate the data quality or model performance.



The screenshot displays a software interface for reviewing 3D point cloud data. The main area is a grid of 10 frames (Frame 1 to Frame 10) showing a car's point cloud in green. Frame 2 is currently selected. Above the grid, there are controls for 'Car: 5', 'Checked', 'Unchecked', 'Show Attribute', and 'Result Status'. Below these are dropdowns for 'Point Cloud', 'Overhead', 'Per 1 Frame', and a range selector from 1 to 10. To the right, three larger views are shown: 'Overhead' (L:4.43 W:1.89), 'Side' (L:4.43 H:1.69), and 'Rear' (W:1.89 H:1.69). A 'Results' panel on the far right shows 'Classifications' for 'Sunlight' and 'Weather', and a list of 'Results No Project (26)' including 'Car(1)', 'Truck(3)', and 'Motorcycle(1)'. At the bottom, there is a 'TimeLine' with a slider from 1 to 75, and a 'Close Review' button.

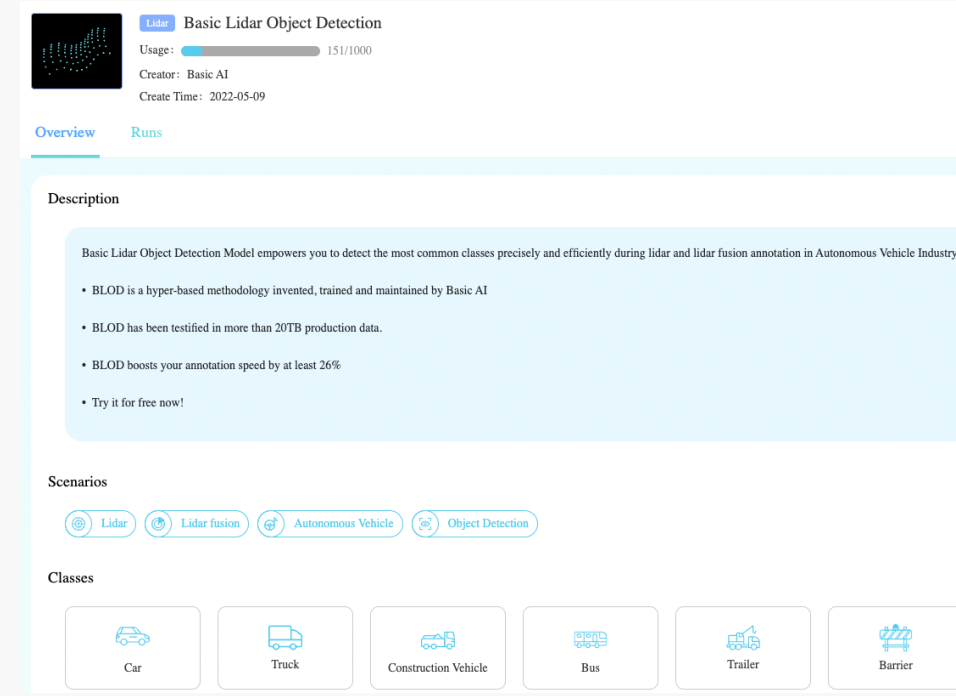
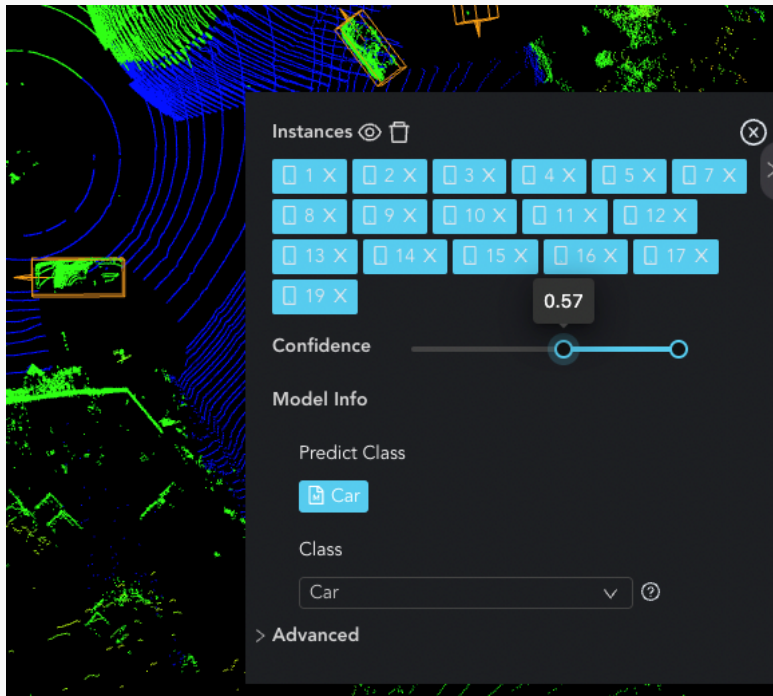


# K4: Model Integration

Xtreme1 is an AI-powered platform for maximizing the efficiency of data and model productions.

Data production	
Labeling	AI pre-labeling
Quality	AI data recommendation
2/3D Fusion	AI sensor calibration
Human Cost	Active Learning
Data Cost	Augmentation & Sythesis

Model production	
Transfer learning	Pre-trained models
Fast application	Off-the-shelf SOTA models
Fast modelling	Low code support
Life-long learning	Data pipeline and streaming





# Technologies

---

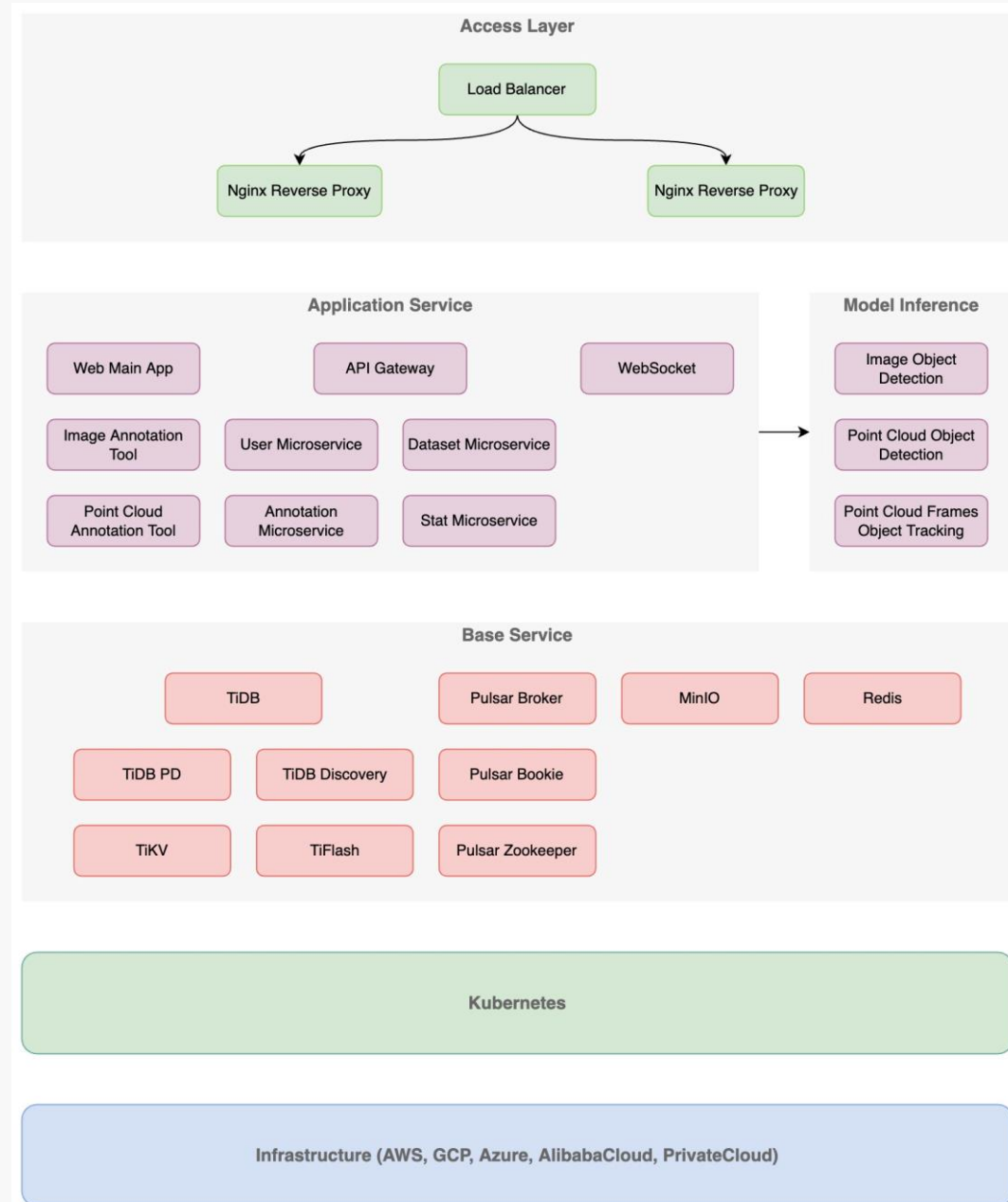
Xtreme1 compliances with the principles of cloud-native architecture to ensure the scalability, elasticity and stability of the platform services.

## N-tier Architecture

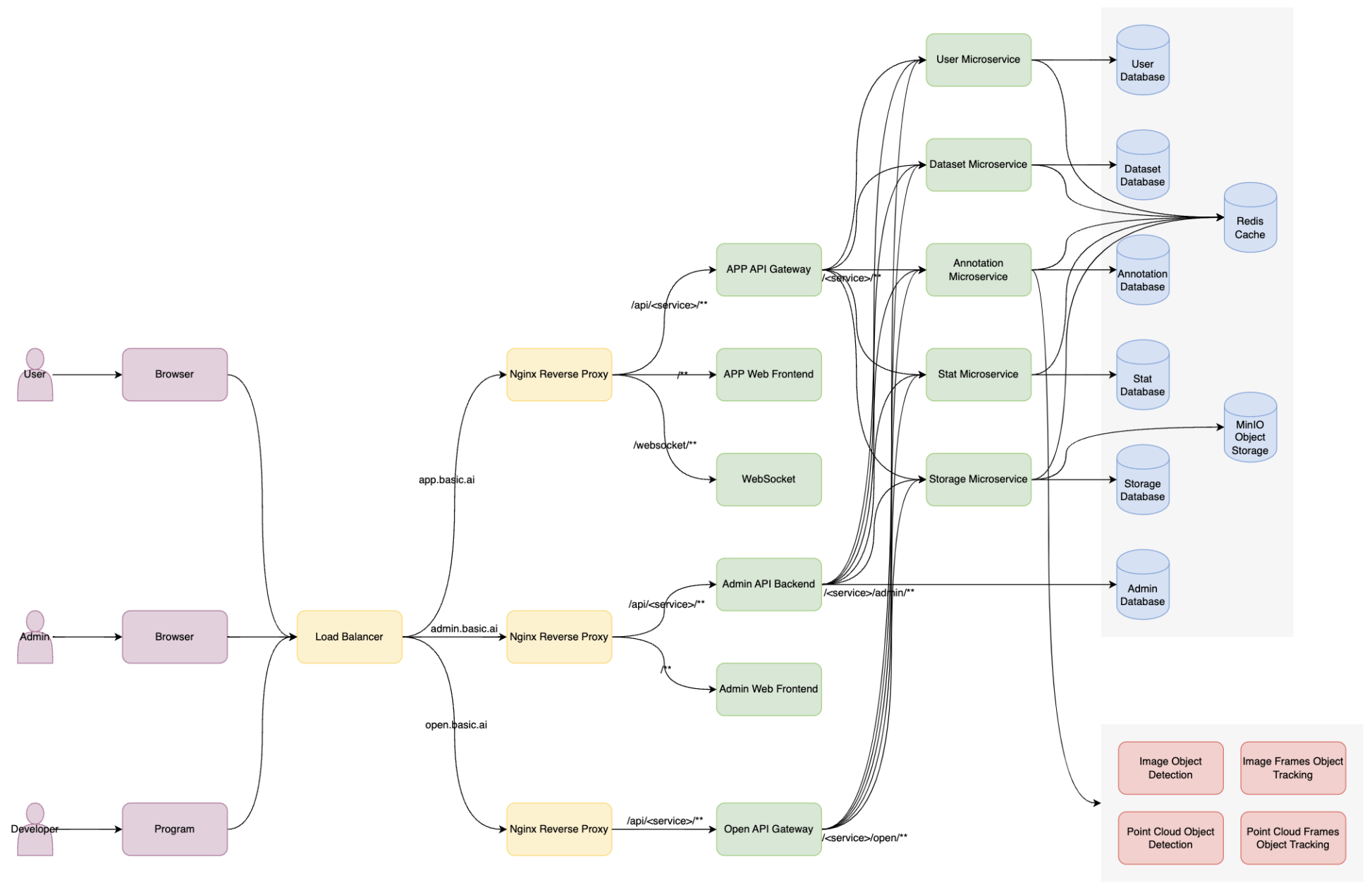
Xtreme1 is a platform with standardized workflow, but diversified tools and requirements of different data types and computing resources.

- Access Layer

Only necessary services are exposed for the data security.

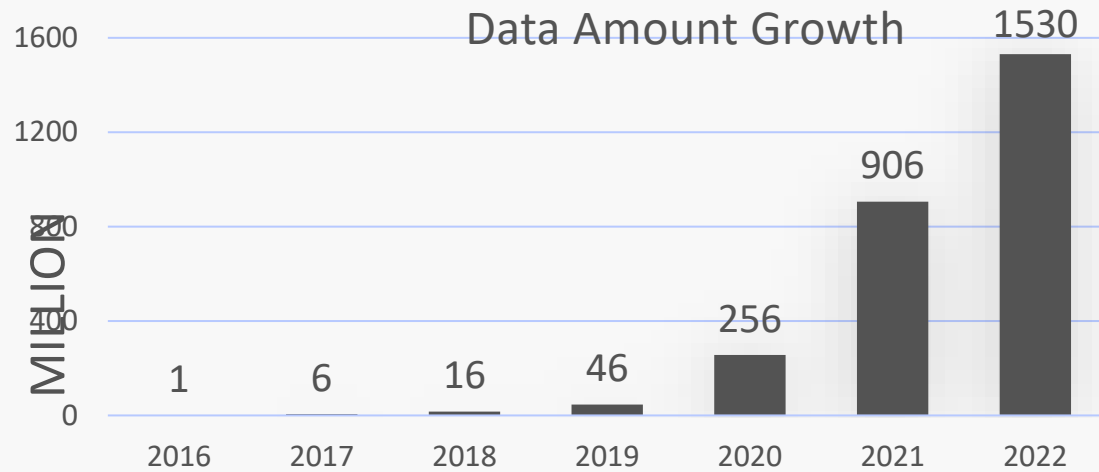


# Xtreme1 Routing Architecture

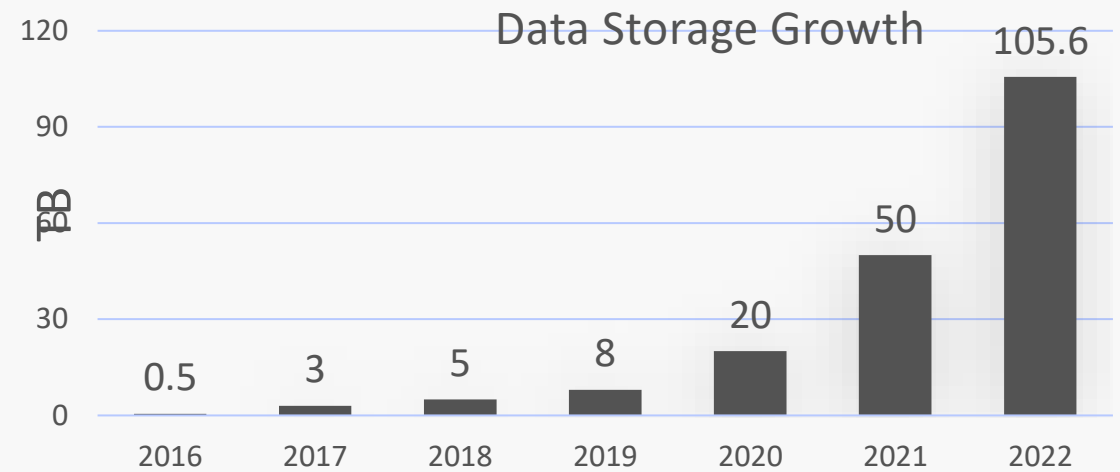


Xtreme1 adapts the distributed database to improve the data analysis in the full life-cycle of AI modelling.

## Structured Data



## Unstructured Data



### ● Challenges

- Large amount: ~40K datasets, ~3M working hours, ~1B data items
- Many types: meta-data, ground-truth, test data, data versions
- Many analyses: accuracy, distribution, progress, performance
- Multi-tenancy: computation/storage isolation, ~10K tenants

### ● TiDB

✓ Distributed

### ● Challenges

- Multi-model data: text, image, video, speech, point cloud
- Massive storage requirements: ~100TB
- High frequency data reading but low writing
- Global data regulation: data security and storage isolation

### ● MinIO

✓ S3 compatible

# Xtreme1 Computation



Xtreme1 has to orchestrate the compute nodes for different time-sensitive tasks and multiple heterogeneous clusters.

## Async Computing

The interface displays a task titled "Image COCO Object Detection" with a usage of 51/1000. It lists the creator as "Basic AI" and the creation time as "2022-06-07". Below this, there are tabs for "Overview" and "Runs". A "Run Model" button is visible. A table shows a single run with the following details:

Run Id	Dataset Name	Created At	Status	Progress	Actions
20221011071...	IMG128	2022-10-11 15:12:15	Success	<div style="width: 100%;"></div>	<a href="#">View Dataset</a> <a href="#">Delete</a>

## ● Challenges

- Data preparation: billions rendering, parsing, compression, chunking
- Data prelabeling: offline model prediction
- Data curation: auto tagging, sorting, clustering
- Modelling: training and evaluation

## ● Pulsar

- ✓ Cloud native, separating compute and storage
- ✓ Horizontally scalable, multi-tenancy support

## Scheduler

The scheduler interface displays a list of workloads under the namespace "basic-alldiv". The table shows the following workloads:

State	Name	Type	Image	Endpoints	Age	Health
Active	is-rtm	Deployment	registry.talos.basic.ai/basic/algorithm/images/service/is-rtm:444858f		8 days	Healthy
Active	mot3d-4f	Deployment	registry.talos.basic.ai/basic/algorithm/images/service/mot3d-4f:cb188c84		109 days	Healthy
Active	pod-tools	Deployment	registry.talos.basic.ai/basic/algorithm/images/service/pod-tools:cb593030		44 days	Healthy
Active	podet-open	Deployment	registry.talos.basic.ai/basic/algorithm/images/service/podet-open:bb43cc85		107 days	Healthy
Active	yolor-detection-coco80	Deployment	registry.talos.basic.ai/basic/algorithm/images/service/yolor-detection-coco80:52b4d748		67 days	Healthy
Active	yolor-detection-rem16	Deployment	registry.talos.basic.ai/basic/algorithm/images/service/yolor-detection-rem16:823c62ad		31 days	Healthy

Below this, there is another namespace "basic-backend" with several workloads:

State	Name	Type	Image	Endpoints	Age	Health
Active	admin	Deployment	registry.talos.basic.ai/basic/backend/admin:33e1abfb	443/HTTPS	38 days	Healthy
Active	annotation	Deployment	registry.talos.basic.ai/basic/backend/annotation:87cb57f		116 days	Healthy
Active	dataset	Deployment	registry.talos.basic.ai/basic/backend/dataset:994036a5		72 days	Healthy
Active	dataset-datasportjob	Deployment	registry.talos.basic.ai/basic/backend/dataset:994036a5		72 days	Healthy
Active	dataset-decompressionjob	Deployment	registry.talos.basic.ai/basic/backend/dataset:994036a5		72 days	Healthy
Active	dataset-imagecompressionjob	Deployment	registry.talos.basic.ai/basic/backend/dataset:994036a5		72 days	Healthy
Active	dataset-modeljob	Deployment	registry.talos.basic.ai/basic/backend/dataset:994036a5		72 days	Healthy

## ● Challenges

- Multiple Runtimes: dev/test/product environments
- Cross regions: US, Europe, SEA, China, India
- GPU computing: CPU/GPU clusters, high-end GPU sharing
- Elasticity: diverse container node duplicate needs

## ● Kubernetes + Rancher

- ✓ A unified interface for multiple clusters
- ✓ GPU sharing is officially supported by Nvidia



# Requesting Incubation at Sandbox Level

---

# Where does Xtreme1 fit on the landscape?

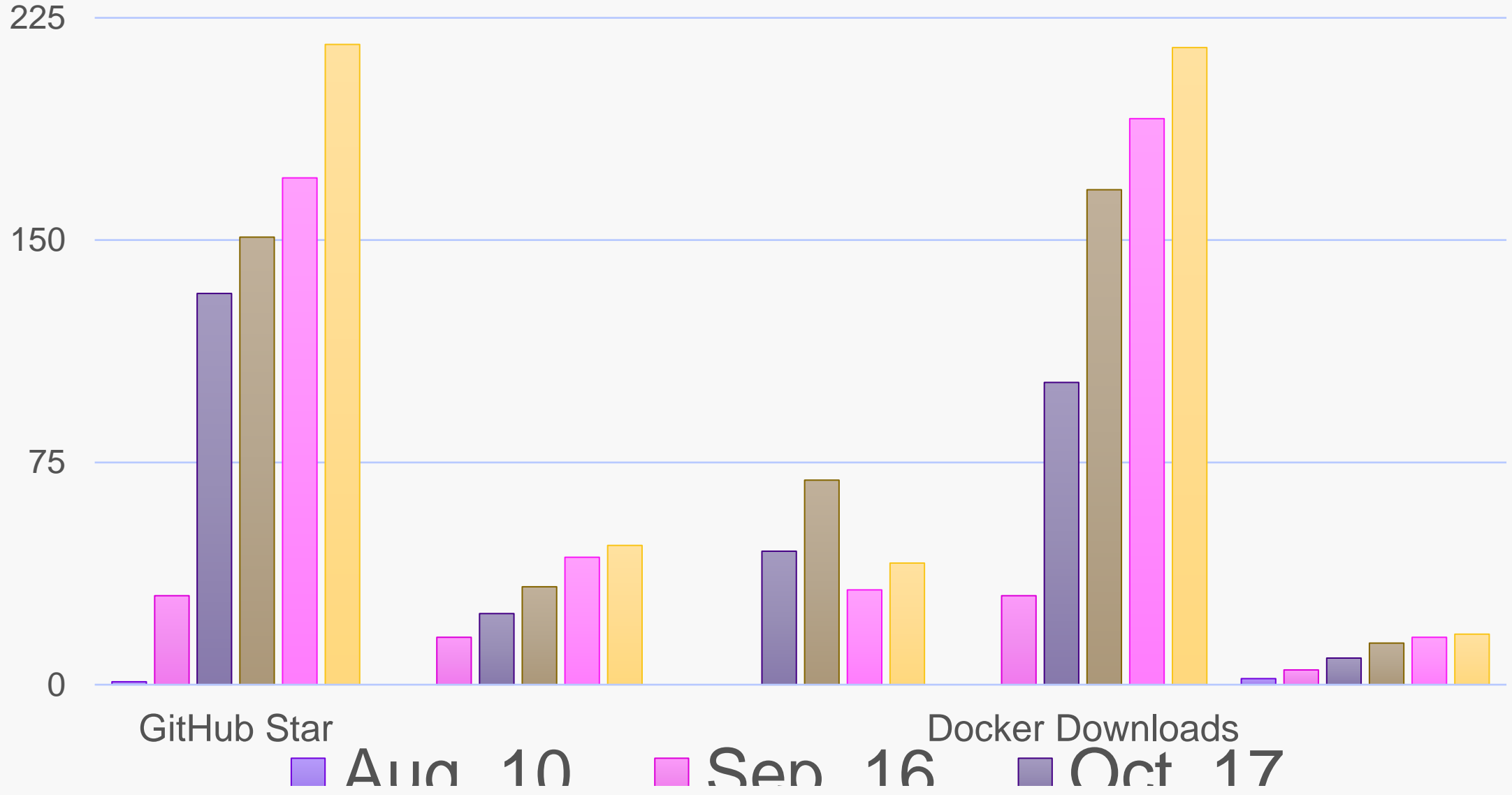


The LF AI & Data landscape explores open source projects in Artificial Intelligence and Data and their respective sub-domains.

1.lfai.foundation



# Xtreme1 Community Statistics as of Nov 19, 2022



## Neutral host of the project

- Vendor-neutral, Not for profit

## Growing community

- Increase contributors by converting new & existing users
- Opportunities to collaborate with other hosted projects
- Increase users by broader outreach through the foundation

## Open Governance model

- Transparent and open governance model
- Distills trust in the running & management of the project
- Neutral management of projects' assets by the foundation

# Roadmap

- 2022Q3, Xtreme1 was officially open-source on Sept. 15
- 2022Q4, Dataset curation for fast data quality control
- 2023Q1, Developer friendly tools, such as Python SDK, API, Command line
- 2023Q2, Support annotation tools for more data types
- 2023Q3, Support model training, testing, deployment

## **Unstructured data management, such as image searching, clustering, sorting**

- Milvus, DocArray

## **Structured data analysis, such as meta data, GT, logging, activities**

- Amundsen

## **Model management, train, test, export, deployment**

- Horovod, ONNX

Thank you for supporting our  
request to incubate in LF AI &  
Data.

Q&A

Don't forget to follow our GitHub



# Formal request

***The authors hereby formally request the incubation of Xtreme1 as a Sandbox Project in LF AI & Data***

# TAC Open Discussion

# LF AI General Updates

 LF AI & DATA



# Upcoming TAC Meetings

# Upcoming TAC Meetings

- › December 15, 2022 – OpenLineage moving from sandbox to incubation
- › December 29, 2022 – Holiday break, no meeting
- › January 12, 2023 – TonY annual review, Angel annual review (tentative)

Please note we are always open to special topics as well.

If you have a topic idea or agenda item, please send agenda topic requests to [tac-general@lists.lfaidata.foundation](mailto:tac-general@lists.lfaidata.foundation)

# Open Discussion

# TAC Meeting Details

- › To subscribe to the TAC Group Calendar, visit the wiki:  
<https://wiki.lfaidata.foundation/x/cQB2> \_\_\_\_\_
- › Join from PC, Mac, Linux, iOS or Android: <https://zoom.us/j/430697670>
- › Or iPhone one-tap:
  - › US: +16465588656,,430697670# or +16699006833,,430697670#
- › Or Telephone:
  - › Dial(for higher quality, dial a number based on your current location):
  - › US: +1 646 558 8656 or +1 669 900 6833 or +1 855 880 1246 (Toll Free) or +1 877 369 0926 (Toll Free)
- › Meeting ID: 430 697 670
- › International numbers available: <https://zoom.us/u/achYtcw7uN>

# Legal Notice

- › The Linux Foundation, The Linux Foundation logos, and other marks that may be used herein are owned by The Linux Foundation or its affiliated entities, and are subject to The Linux Foundation's Trademark Usage Policy at <https://www.linuxfoundation.org/trademark-usage>, as may be modified from time to time.
- › Linux is a registered trademark of Linus Torvalds. Please see the Linux Mark Institute's trademark usage page at <https://lmi.linuxfoundation.org> for details regarding use of this trademark.
- › Some marks that may be used herein are owned by projects operating as separately incorporated entities managed by The Linux Foundation, and have their own trademarks, policies and usage guidelines.
- › TWITTER, TWEET, RETWEET and the Twitter logo are trademarks of Twitter, Inc. or its affiliates.
- › Facebook and the "f" logo are trademarks of Facebook or its affiliates.
- › LinkedIn, the LinkedIn logo, the IN logo and InMail are registered trademarks or trademarks of LinkedIn Corporation and its affiliates in the United States and/or other countries.
- › YouTube and the YouTube icon are trademarks of YouTube or its affiliates.
- › All other trademarks are the property of their respective owners. Use of such marks herein does not represent affiliation with or authorization, sponsorship or approval by such owners unless otherwise expressly specified.
- › The Linux Foundation is subject to other policies, including without limitation its Privacy Policy at <https://www.linuxfoundation.org/privacy> and its Antitrust Policy at <https://www.linuxfoundation.org/antitrust-policy>. each as may be modified from time to time. More information about The Linux Foundation's policies is available at <https://www.linuxfoundation.org>.
- › Please email [legal@linuxfoundation.org](mailto:legal@linuxfoundation.org) with any questions about The Linux Foundation's policies or the notices set forth on this slide.