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

October 21, 2021

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
- Wiki: wiki.lfaidata.foundation
- GitHub: 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

- 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)
- Introducing OpenBytes (new project for incubation)
- Approval of Minutes from previous meeting (2 mins)
- LF AI General Updates (2 min)
- Open Discussion (2 min)

TAC Voting Members

* = still need backup specified on wiki

Board Member	Contact Person	Email
AT&T	Anwar Atfab*	anwar@research.att.com
Baidu	Ti Zhou	zhouti@baidu.com
Ericsson	Rani Yadav-Ranjan*	rani.yadav-ranjan@ericsson.com
Huawei	Huang Zhipeng	huangzhipeng@huawei.com
IBM	Susan Malaika	malaika@us.ibm.com
Nokia	Jonne Soininen	jonne.soininen@nokia.com
OPPO	Jimin Jia*	jiajimin@oppo.com
SAS	Nancy Rausch	nancy.rausch@sas.com
Tech Mahindra	Amit Kumar	Kumar_Amit@techmahindra.com
Tencent	Bruce Tao	brucetao@tencent.com
Zilliz	Jun Gu	jun.gu@zilliz.com
ZTE	Wei Meng	meng.wei2@zte.com.cn
Graduate Project	**Contact Person**	**Email**
Acumos	Nat Subramanian	natarajan.subramanian@techmahindra.com
Angel	Bruce Tao	brucetao@tencent.com
Egeria	Mandy Chessell	mandy_chessell@uk.ibm.com
Horovod	Travis Addair*	taddair@uber.com
Milvus	Xiaofan Luan	xiaofan.luan@zilliz.com
ONNX	Jim Spohrer (Chair of TAC)	spohrer@us.ibm.com
Pyro	Fritz Obermeyer*	fritz.obermeyer@gmail.com

Introducing OpenBytes

Apply for Incubation at the Sandbox level

OpenBytes @ LF AI & Data

10/21/2021

GRAVITI

AGENDA

1 What is OpenBytes?

2 Why would we like to bring OpenBytes to LF AI & data?

3 What problems is OpenBytes trying to solve?

4 What’s coming next?

What is OpenBytes

Inspire AI Innovation with Open Datasets

The mission of the OpenBytes Project is to facilitate the wider sharing of and collaboration with data in the AI community. This is accomplished through the promotion of data standards and formats, as well as enabling contributions of data.

Our Supporters

GRAVITI ElectrifAi Motional

Predibase ZILLIZ jina

Why would we like to bring OpenBytes to LF AI & data?

LF AI & DATA

Community-driven

Open Source Solutions

Accelerates the Development of AI, ML, DL and Data

OpenBytes

Neutral and Open Governance

Promote Unified Data Sharing License and Standards

Community for Data Sharing

What problems are we trying to solve?

The value of open data has been shown in academic breakthroughs and business values...

ACADEMIC

Raise academic interest and fuel the revolution of CNN and machine learning in general

Inspire scholars write papers on sentimental analysis and recommender system by using Amazon Reviews

Apple published papers on arXiv and proposed VoxelNet: an end-to-end SOTA 3D object detection model, by applying the KITTI Dataset

BUSINESS

Open sourcing millions of image data with unified annotation standard fuels the development of ML

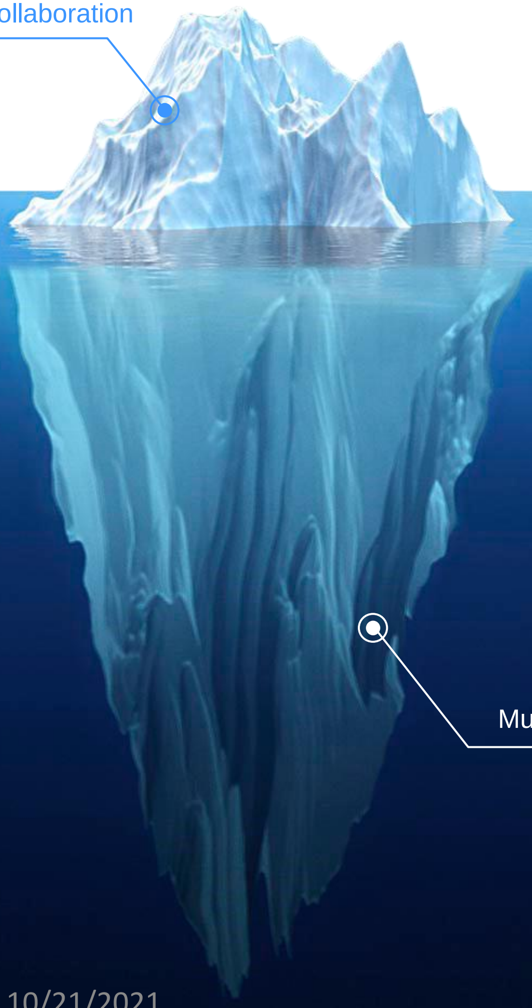
Collaborative innovation on autonomous driving by open sourcing valuable and rare road data

IBM releases the Finance Proposition Bank and Contracts Proposition Bank datasets, which are part of an active research program aimed at improving the natural language understanding technologies

Enable the Power of AI | Drive Science Innovation | Promote Openness and Fairness

What problems are we trying to solve?

...However, only a small tip of large data volume has been open to public due to standards, legal and resources limitations.



Only a tip of data is for open collaboration

Much hidden is unknown

What are the barriers behind it?

Standards

Dataset sharing policy, format and data quality

Legal

Licensing, privacy and security

Resources

Technical supports, funding and knowledge

BARRIER 1 - Standards

The lack of standards prevent publishers and users from publishing, distributing and reusing.

It is essential that open data fulfill certain standards and requirements in order to match the demand and expectations between potential data publishers and users.

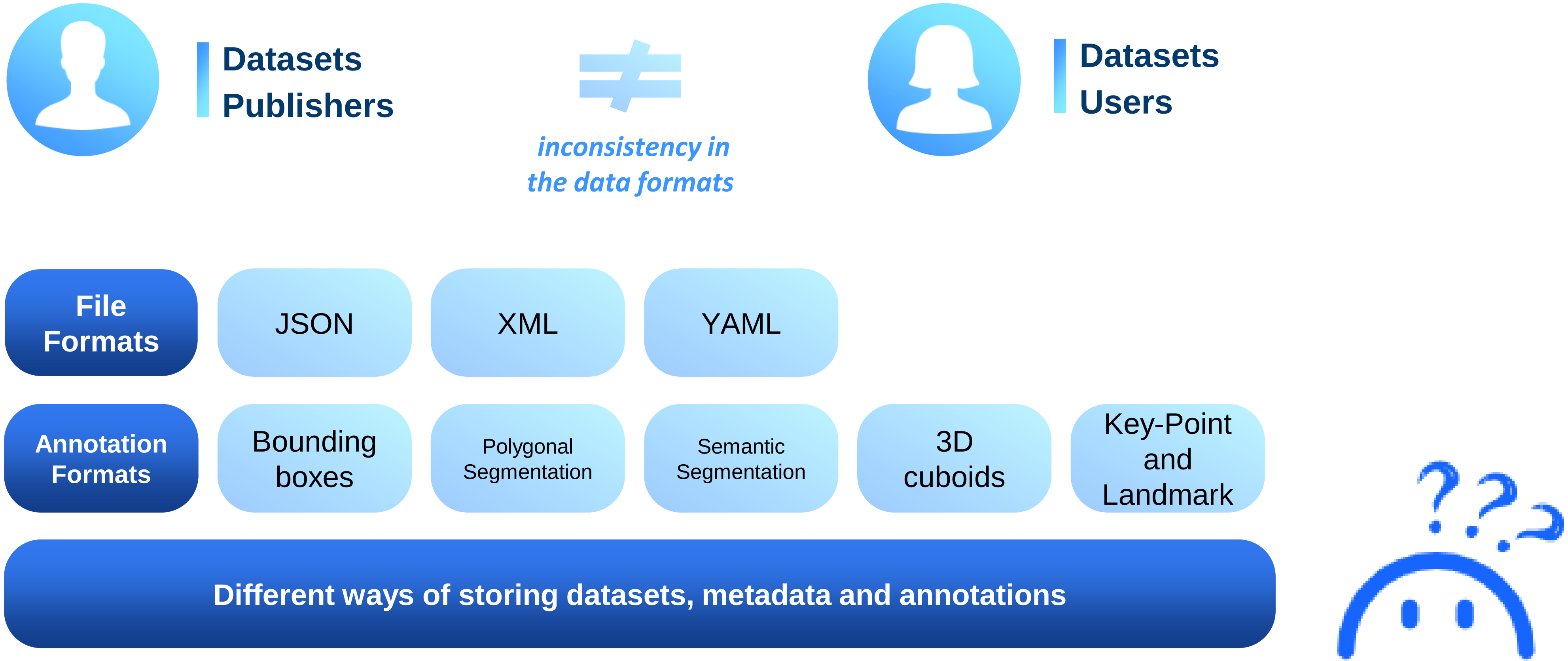
The guidelines should include:

1. Process to publish and maintain datasets
2. Policy for data desensitization
3. Guidelines of data quality assurance
4. Practices of defining, standardizing, and visualizing different types of datasets

BARRIER 1 - Standards

Distribution of dataset needs a standardized format of dataset.

Different formats and classifications of datasets make users harder to understand or reuse.



Datasets Publishers

≠

inconsistency in the data formats

Datasets Users

File Formats: JSON, XML, YAML

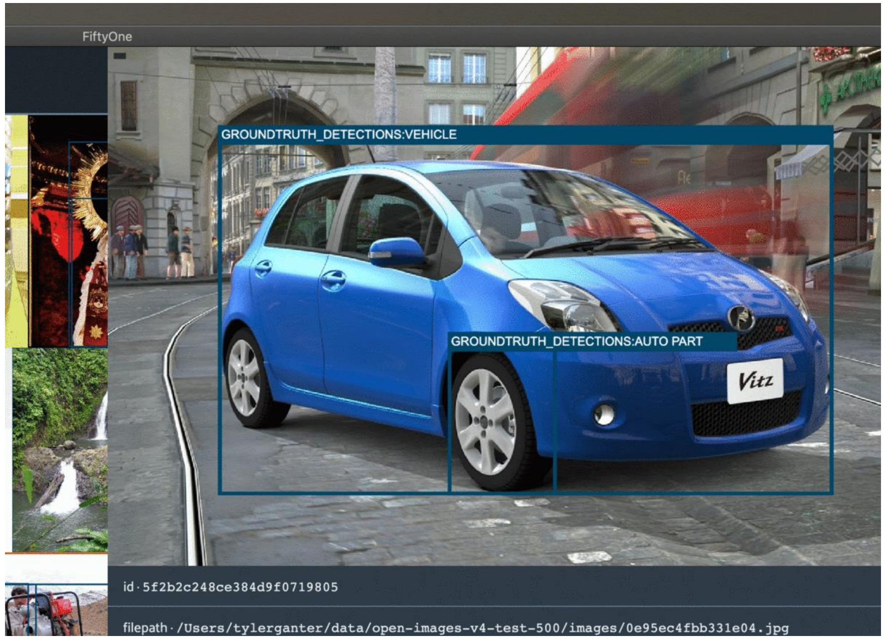
Annotation Formats: Bounding boxes, Polygonal Segmentation, Semantic Segmentation, 3D cuboids, Key-Point and Landmark

Different ways of storing datasets, metadata and annotations

BARRIER 1 - Standards

Distribution of dataset needs a standardized quality of dataset.

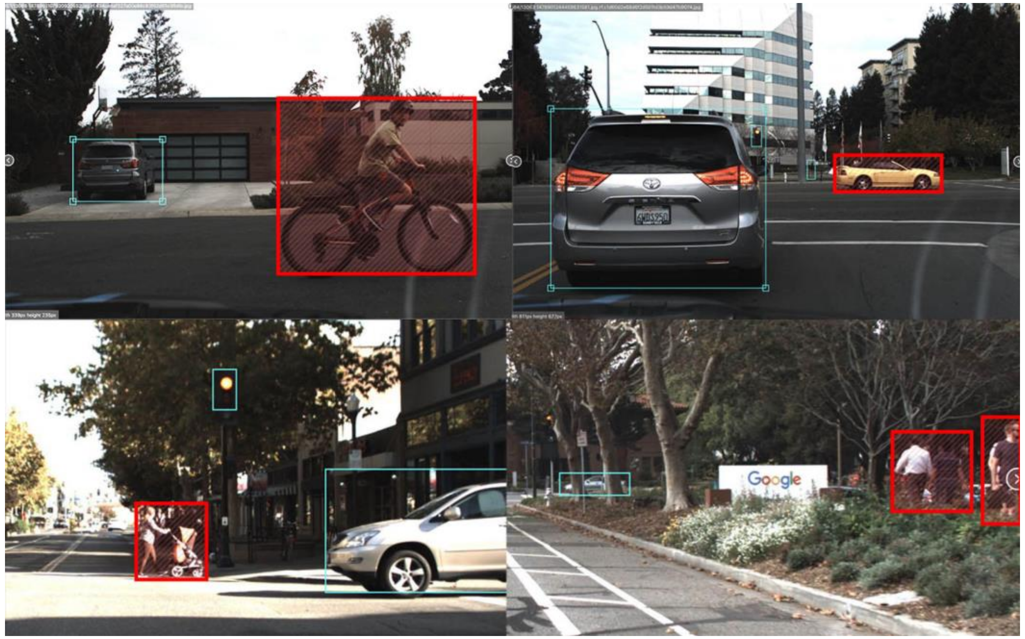
Incompleteness and inaccuracy in metadata and annotations may lead to biased model performance


FiftyOne
GROUNDTRUTH_DETECTIONS:VEHICLE
GROUNDTRUTH_DETECTIONS:AUTO PART
Vitz

id · 5f2b2c248ce384d9f0719805

filepath · /Users/tylerganter/data/open-images-v4-test-500/images/0e95ec4fbb331e04.jpg

An example image from Google's Open Images dataset: The back wheel is not annotated even though the model successfully detects it.



About 11.5% images of the Udacity Self Driving Car Dataset have missing labels.

BARRIER 2 - Legal

Inappropriate licensing increases vulnerability of both publishers and users.

UNCLEAR LICENSING

- No suitable license for new generated dataset.
- No specific license for open datasets

COMPLEXITY

- Hard to understand various licenses with different versions and interpretations

CONSTRAINTS

- Purposes for reusing
- Threat of lawsuits from privacy and security risks

RIGHT FOR REUSE AND REDISTRIBUTION

- Ownership of the datasets
- Implicit definition of some scenarios

BARRIER 3 - Resources

More support are needed for open datasets contributions.

TECHNICAL SUPPORT

modifying, publishing,
maintaining datasets
lack of guiding principles

FUNDING

the whole lifecycle cost
from publishing to maintaining

PROFESSIONAL SERVICES

legal consulting services,
strategic and business planning before
publishing data

What's Coming Next?

Project OpenBytes is aiming to establish an open data community to enable standardized, licensed and interoperable data sharing.

Licensing

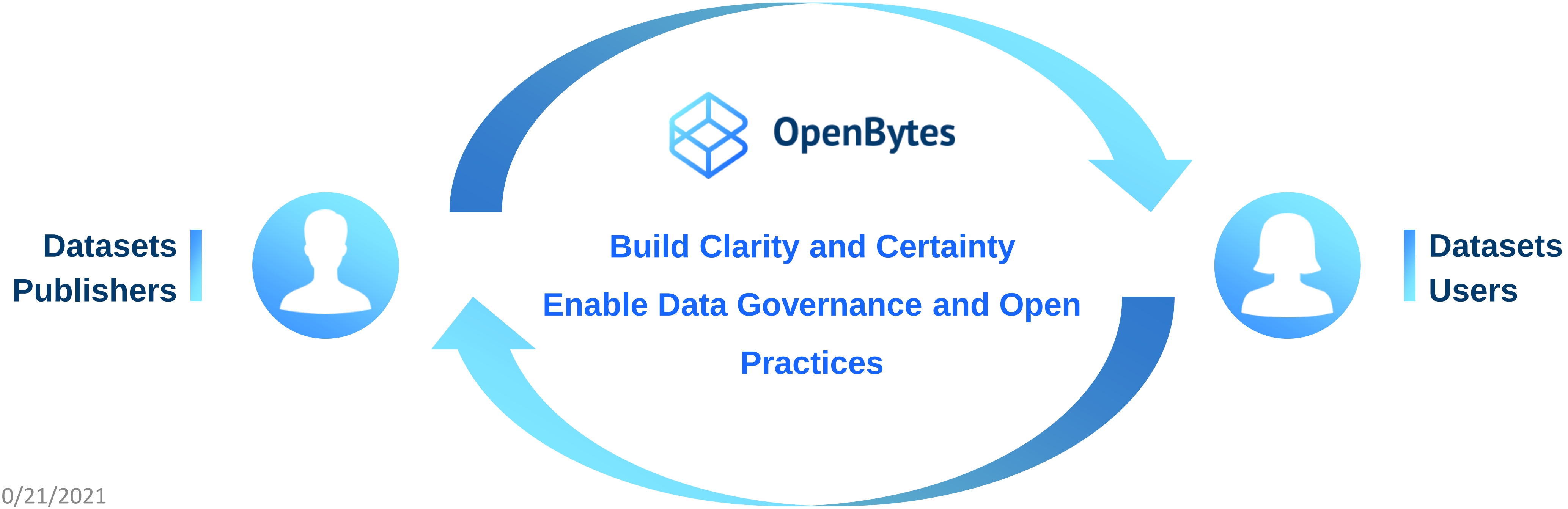
We will market the licenses specific to open datasets that address problems of copyright and distribution, reducing contributors' liability risks, supporting legal protection of the data.

Standards

We will formulate open dataset standards that specify and require how datasets should be published, shared and exchanged.

Resources

We will collect the communities' efforts to provide funding, professional services, and technical supports with dataset participants.



Datasets Publishers

OpenBytes

Build Clarity and Certainty

Enable Data Governance and Open Practices

Datasets Users

Potential Integration with LF AI & Data

Project OpenBytes aims to integrate with LF AI & Data by enriching data sharing community and connecting supports in data distribution

Enrich Data Sharing Community

DiDi

NYU

Datasets

Promote collaborative innovation through opening data

HOROVOD

Petastorm

Model

Standardized dataset accelerate model training

Connect Supports in Data Distribution

COMMUNITY DATA LICENSE AGREEMENT

License

Promote and Collaborate with open data license agreements

LabelImg

intel CVAT

Data quality

Introduce technical supports such as labelling

Potential Integration with LF AI & Data

Project OpenBytes hopes to collaborate with the LF AI & Data community to jointly promote open dataset and enable faster deployment of AI.

OpenBytes

LF AI & DATA

- Collaborative development on standardized open datasets
- Wider adoption for open datasets
- Overcome AI barriers for businesses

- Raise awareness and promoting open datasets
- Community-driven support
- Potential Partnership with members and projects

THANK YOU

Edward Cui | edward.cui@graviti.com

GRAVITI

OpenBytes approval

Minutes approval

LF AI & DATA

Meeting minutes approval – Voting members

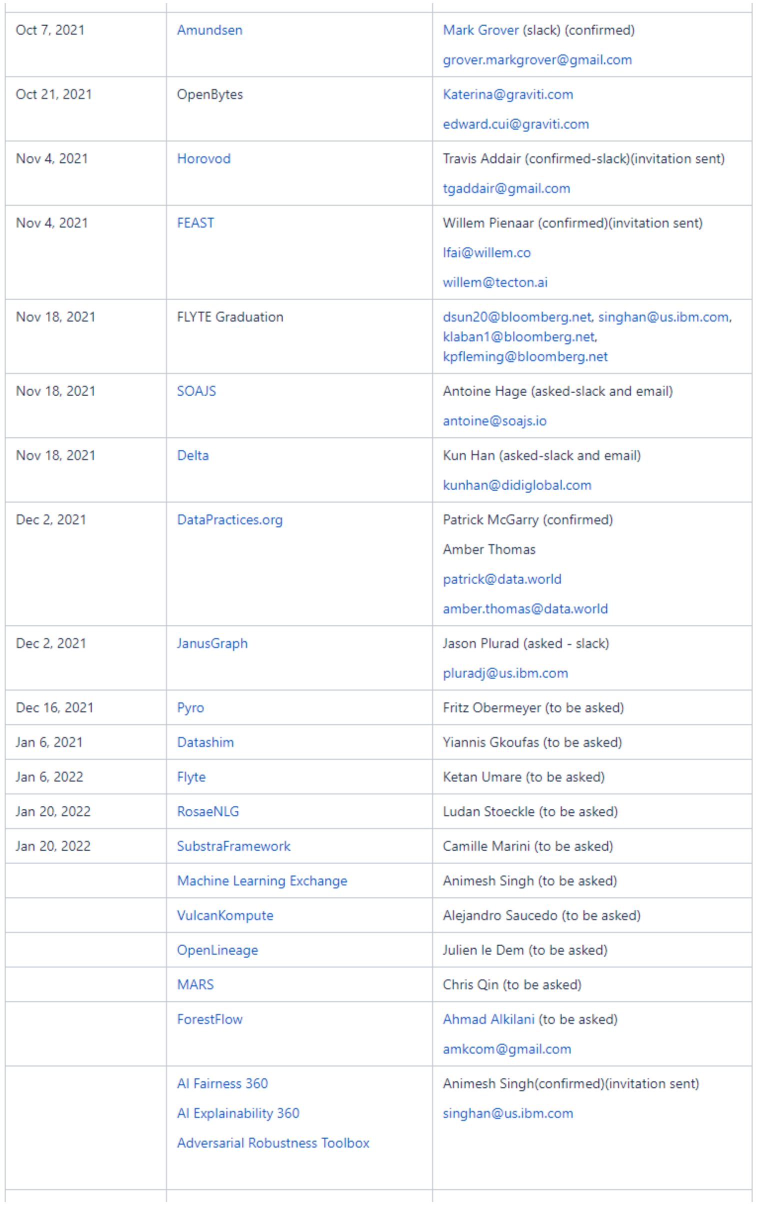
September, October

Upcoming TAC Meetings

Upcoming TAC Meetings (Tentative)

- Oct 21, 2021: AI Fairness 360, AI Explainability 360, Adversarial Robustness Toolbox
- Nov 4, 2021: Horovod, FEAST; Outreach Committee update

Please send agenda topic requests to tac-general@lists.lfaidata.foundation



Oct 7, 2021	Amundsen	Mark Grover (slack) (confirmed) grover.markgrover@gmail.com
Oct 21, 2021	OpenBytes	Katerina@graviti.com edward.cui@graviti.com
Nov 4, 2021	Horovod	Travis Addair (confirmed-slack)(invitation sent) tgaddair@gmail.com
Nov 4, 2021	FEAST	Willem Pienaar (confirmed)(invitation sent) lfai@willem.co willem@tecton.ai
Nov 18, 2021	FLYTE Graduation	dsun20@bloomberg.net, singhan@us.ibm.com, klaban1@bloomberg.net, kpfleming@bloomberg.net
Nov 18, 2021	SOAJS	Antoine Hage (asked-slack and email) antoine@soajs.io
Nov 18, 2021	Delta	Kun Han (asked-slack and email) kunhan@didiglobal.com
Dec 2, 2021	DataPractices.org	Patrick McGarry (confirmed) Amber Thomas patrick@data.world amber.thomas@data.world
Dec 2, 2021	JanusGraph	Jason Plurad (asked - slack) pluradj@us.ibm.com
Dec 16, 2021	Pyro	Fritz Obermeyer (to be asked)
Jan 6, 2021	Datashim	Yiannis Gkoufas (to be asked)
Jan 6, 2022	Flyte	Ketan Umare (to be asked)
Jan 20, 2022	RosaeNLG	Ludan Stoeckle (to be asked)
Jan 20, 2022	SubstraFramework	Camille Marini (to be asked)
	Machine Learning Exchange	Animesh Singh (to be asked)
	VulcanKompute	Alejandro Saucedo (to be asked)
	OpenLineage	Julien le Dem (to be asked)
	MARS	Chris Qin (to be asked)
	ForestFlow	Ahmad Alkilani (to be asked) amkcom@gmail.com
	AI Fairness 360 AI Explainability 360 Adversarial Robustness Toolbox	Animesh Singh(confirmed)(invitation sent) singhan@us.ibm.com

Schedule: https://wiki.lfaidata.foundation/pages/editpage.action?pageId=43286684

Open Discussion

LF AI & DATA

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 with any questions about The Linux Foundation's policies or the notices set forth on this slide.

LF AI & DATA