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

March 10, 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
- 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)
- Approval of Minutes from previous meeting (2 mins)
- THOTH (30 minutes)
- MlWorkflow & Interop Committee: progress update presentation on the dataset license compliance initiative (20 minutes)
- 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

Challenge with TAC Quorum

- 19 voting members requiring 10 voting members to achieve quorum

- Proposing updating charter to reflect the following changes:
  - A TAC voting member who misses 2 TAC meetings in a row will lose their voting seat until they attend twice in a row.

- Process: Socialize with GB and TAC. Propose amendment to the Charter and have the GB vote on it.

TAC Voting Members

* = still need backup specified on [wiki]

Member Representatives

Member Company or Graduated Project	Membership Level or Project Level	Voting Eligibility	Country	TAC Representative	Designated TAC Representative Alternates
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)
IBM	Premier	Voting Member	USA	Susan Malaika	Saishruthi Swaminathan
Nokia	Premier	Voting Member	Finland	@Michael Rooke	@Jonne Soininen
OPPO	Premier	Voting Member	China	Jimin Jia	
SAS	Premier	Voting Member	USA	*Nancy Rausch	JP Trawinski
Tech Mahindra	Premier	Voting Member	India	Amit Kumar	Prasanna Kulkarni
Tencent	Premier	Voting Member	China	Bruce Tao	Huaming Rao
ZTE	Premier	Voting Member	China	Wei Meng	Liya Yuan
Acumos Project	Graduated Technical Project	Voting Member	USA	Amit Kumar	Prasanna Kulkarni
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 Chair

LF AI & DATA

Minutes approval

LF AI & DATA

Approval of February 10, 2021 Minutes

Draft minutes from the February 10th TAC call were previously distributed to the TAC members via the mailing list

Proposed Resolution:

- That the minutes of the February 10 meeting of the Technical Advisory Council of the LF AI & Data Foundation are hereby approved.

An update from the ML Workflow & Interop Committee dataset license compliance initiative

Howard <huangzhipeng@huawei.com>
Liza lizi4@huawei.com
Gopi Krishnan Rajbahadur <gopi.krishnan.rajbahadur1@huawei.com>

LF AI & DATA

10MAR2022

Dataset license compliance – A progress report for Mlflow and interop committee

Gopi Krishnan Rajbahadur

HUAWEI

gopikrishnanrajbahadur@ gmail.com

@gopirajbahadur

This work would not have been possible without the contributions from Erika Tuck, Li Zi, Dr. Dayi Lin, Dr. Boyuan Chen, Prof. Zhen Ming (Jack) Jiang, Prof. Daniel M. German

Outline

Recap

License compliance process for curated datasets

Challenges

Current progress

Road ahead

Outline

Recap

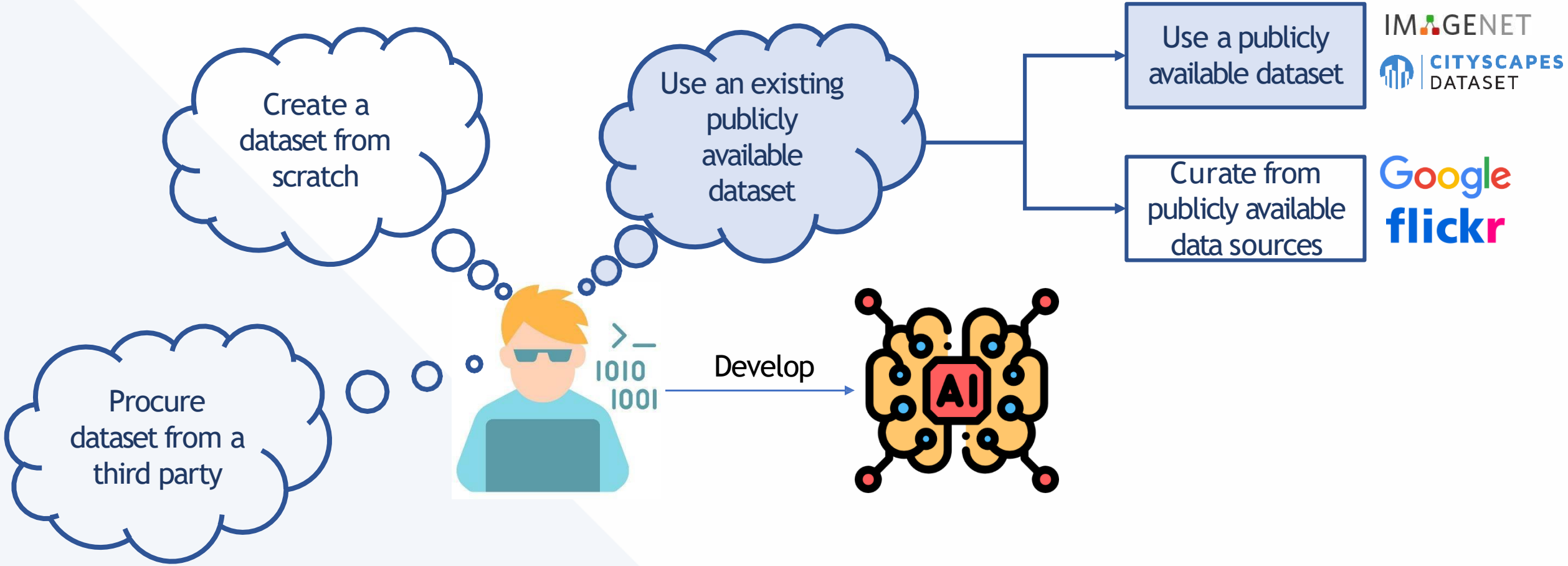
License compliance process for curated datasets

Challenges

Current progress

Road ahead

There are several ways of acquiring the data required to build AI software



Disclaimers

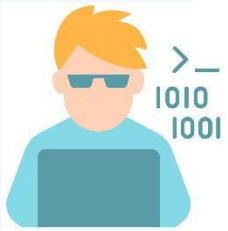
The potential risks that we assess does not necessarily constitute as legal risks. We simply propose an approach to identify potential risks

Whether a dataset's copyright should be extended to a model trained on the given dataset is still an open question and we don't argue one way or another

We loosely define the term dataset license. Unlike OSS, most datasets don't have a definitive license rather they outline terms of use, agreements. For the purposes of this talk, we call them license

The views presented in this presentation are that of the authors and it does not reflect on the views presented by Huawei.

Our approach to assess the potential risks of using publicly available datasets in commercial AI software



AI Engineer

Lawyer

Dataset

Phase 1: License identification

- License extraction
- Provenance extraction
- Lineage extraction

Phase 1: outputs
- Dataset provenance table
- Dataset lineage table

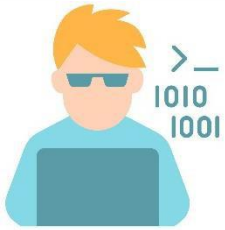
Phase 2: License compliance assessment

- License interpretation
- License compatibility analysis

Phase 2: outputs
- Dataset license interpretation table
- Dataset license compatibility assessment table

Scenario assessment

Example Scenario assessment



AI Engineer

Wants to use

The CIFAR-10 dataset

For the following scenarios

- Commercially distribute the dataset
- Release a product with AI model
- Commercialize the model output

Our approach to assess the potential risks of using publicly available datasets in commercial AI software

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead

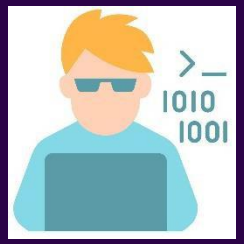
License metadata	Licensor	License name	Dataset name	Dataset version
	Alex Krizhevsky	Custom license	CIFAR-10	N/A
	Credit/Attribution Notice			
	Learning Multiple Layers of Features from Tiny Images, Alex Krizhevsky, 2009.			
	License validity period	Liability /Warranty	Designated third parties	Additional conditions
	N/A	N/A	Only by agreement	None

Data (standalone)	Access	Tagging	Distribute	Re-represent
Rights	✓	✓ (✗)	✓ (✗)	✓ (✗)
Obligations	Cite paper	Cite paper	Cite paper	Cite paper

Data rights in conjunction with model	Bench-mark	Re-search	Publish	In-ternal Use	Commercialization: Out-put	Commercialization: Model	Model Reverse Engineer
Rights	✓	✓	✓	✓	✓ (✗)	✓ (✗)	✓
Obligations	Cite paper	Cite paper	Cite paper	Cite paper	Cite paper	Cite paper	Cite paper

There are risks associated with using CIFAR-10 for any of these scenarios

Example Scenario assessment



AI Engineer

Wants to use

The CIFAR-10 dataset

For the following scenarios

- Commercially distribute the dataset ⊘
- Release a product with AI model ⊘
- Commercialize the model output ⊘

Our potential risk assessment results on studied publicly available datasets

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead

	Commercially distribute the dataset	Release a product with AI model	Commercialize the model output
IMAGENET	✗	✗	✗
CITYSCAPES DATASET	✗	✗	✗
VGG Face Dataset	✓	✗	✗
The CIFAR-10 dataset	✗	✗	✗
COCO Common Objects in Context	✓	✓	✓
Flickr-Faces-HQ Dataset (FFHQ)	✓	✗	✗

Outline

Recap

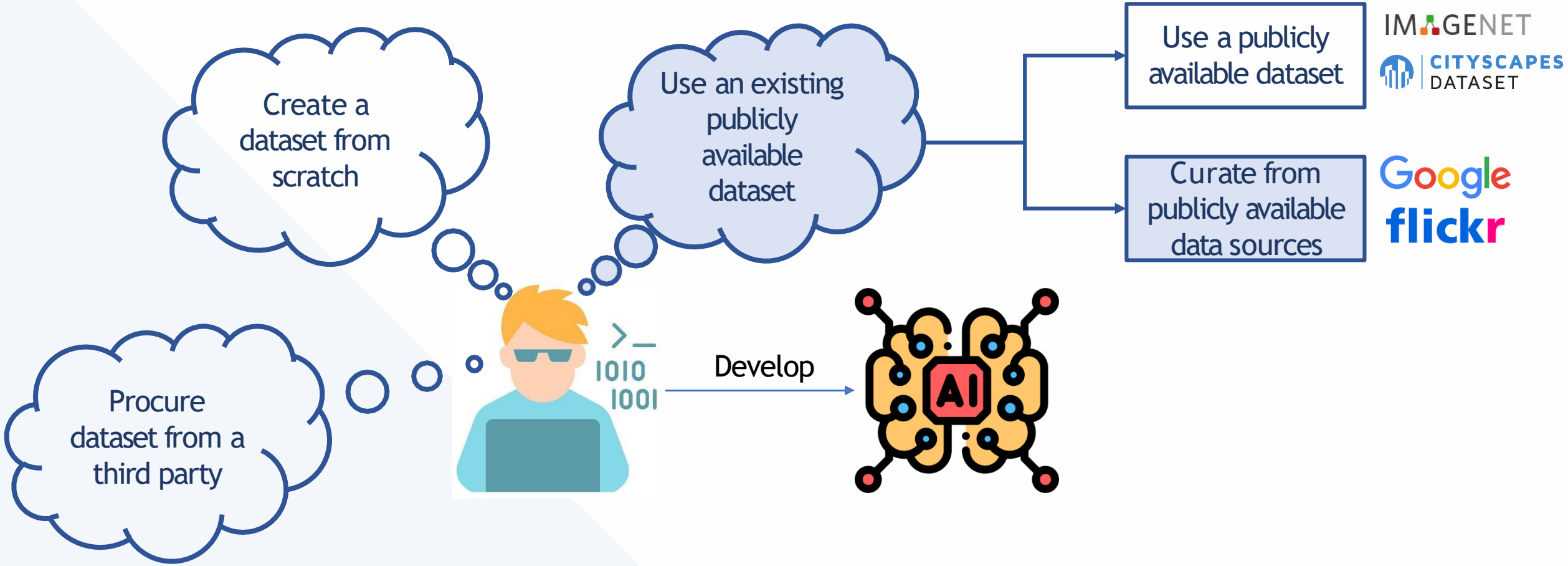
License compliance process for curated datasets

Challenges

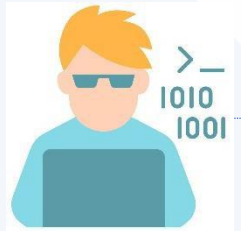
Current progress

Road ahead

There are several ways of acquiring the data required to build AI software



Our approach to assess the potential risks of using datasets created from publicly available data sources



AI Engineer

Lawyer

Data source | Data source | Data source

Phase 1: License identification

- License extraction
- Provenance record
- Lineage record

Phase 1: outputs
- Dataset provenance table
- Dataset lineage table

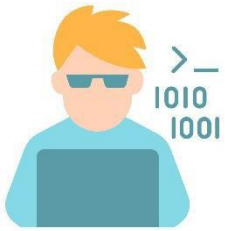
Phase 2: License compliance assessment

- License interpretation
- License compatibility analysis

Phase 2: outputs
- Dataset license interpretation table
- Dataset license compatibility assessment table

Create dataset → Scenario Assessment

Example Scenario assessment



AI Engineer

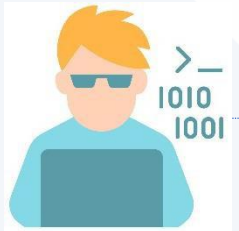
Wants to create dataset from

Google flickr

For the following scenarios

- Commercially distribute the dataset
- Release a product with AI model
- Commercialize the model output

Our approach to assess the potential risks of using datasets created from publicly available data sources



AI Engineer

Lawyer

Data source | Data source | Data source

Phase 1: License identification

- License extraction
- Provenance record
- Lineage record

Phase 1: outputs
- Dataset provenance table
- Dataset lineage table

Phase 2: License compliance assessment

- License interpretation
- License compatibility analysis

Phase 2: outputs
- Dataset license interpretation table
- Dataset license compatibility assessment table

Create dataset → Scenario Assessment

Provenance record

Since the data collection process is controlled by the curator, provenance of the dataset can be created as a record using the schema that we provide

Lineage record

Similar to provenance the curator, can track and record the lineage using our schema

When curating datasets, unless another (pre-curated) dataset is involved, no explicit provenance or lineage extraction is required

Outline

Recap

License compliance process for curated datasets

Challenges

Current progress

Road ahead

Challenges in ensuring dataset license compliance

Provenance related challenges

Lineage related challenges

License related challenges

Unclear licensing range

All the data sources are not specified

Rights and obligations are unclear

Unclear license locations

Identifying the minimum licensable data unit

Multiple license interactions and their effects are unclear

Multiple copies/variants of dataset hosted in different places

Challenges in ensuring dataset license compliance

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead

Provenance related challenges | Lineage related challenges | License related challenges

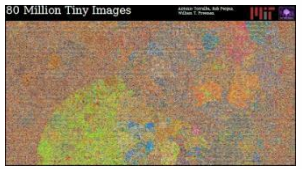
Unclear licensing range

Unclear license locations

Multiple copies/variants of dataset hosted in different places

Google picsearch altavista flickr webshots don't blink Ask.com

200X ??



2008

The CIFAR-10 dataset

2009

- Learning Multiple Layers of Features from Tiny Images, Alex Krizhevsky, 2009.

When using CIFAR-10 license from which year should apply for the data sources?

2022

Challenges in ensuring dataset license compliance

Unclear licensing range

Unclear license locations

Multiple copies/variants of dataset hosted in different places

Sentiment Analysis

Sentiment Treebank

License is provided with the downloaded dataset in the README file

GLUE

License is provided in the GitHub page

IMAGENET

License is provided along with the website

Challenges in ensuring dataset license compliance

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead

Provenance related challenges | Lineage related challenges | License related challenges

Unclear licensing range

Unclear license locations

Multiple copies/variants of dataset hosted in different places

The CIFAR-10 dataset

PyTorch | Keras | kaggle | GitHub

DeepAI | TensorFlow

Challenges in ensuring dataset license compliance

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead

Provenance related challenges | Lineage related challenges | License related challenges

All the data sources are not specified

Identifying the minimum licensable data unit

Google, picsearch, flickr, webshots (don't blink), Ask.com, altavista

The CIFAR-10 dataset

These data sources are not specified in the CIFAR-10 report

Challenges in ensuring dataset license compliance

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead

Provenance related challenges | Lineage related challenges | License related challenges

All the data sources are not specified

Identifying the minimum licensable data unit

Google picsearch altavista

flickr webshots don't blink Ask.com

The CIFAR-10 dataset

- Learning Multiple Layers of Features from Tiny Images, Alex Krizhevsky, 2009.

Determining which among these is the minimum license unit is a hard problem

Challenges in ensuring dataset license compliance

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead

Provenance related challenges | Lineage related challenges | License related challenges

Rights and obligations are unclear

Multiple license interactions and their effects are unclear

The CIFAR-10 dataset

Please cite it if you intend to use this dataset.

- Learning Multiple Layers of Features from Tiny Images, Alex Krizhevsky, 2009.

IMAGENET

[RESEARCHER_FULLNAME] (the "Researcher") has requested permission to use **the ImageNet database** (the "Database") at **Princeton University and Stanford University.** In exchange for such permission, Researcher hereby agrees to the following terms and conditions:

1. Researcher shall use the Database only for non-commercial research and educational purposes.
2. Princeton University and Stanford University make no representations or warranties regarding the Database, including but not limited to warranties of non-infringement or fitness for a particular purpose.
3. Researcher accepts full responsibility for his or her use of the Database and shall defend and indemnify the ImageNet team, Princeton University, and Stanford University, including their employees, Trustees, officers and agents, against any and all claims arising from Researcher's use of the Database, including but not limited to Researcher's use of any copies of copyrighted images that he or she may create from the Database.
4. Researcher may provide research associates and colleagues with access to the Database provided that they first agree to be bound by these terms and conditions.
5. Princeton University and Stanford University reserve the right to terminate Researcher's access to the Database at any time.
6. If Researcher is employed by a for-profit, commercial entity, Researcher's employer shall also be bound by these terms and conditions, and Researcher hereby represents that he or she is fully authorized to enter into this agreement on behalf of such employer.
7. The law of the State of New Jersey shall apply to all disputes under this agreement.

No clear mention if the dataset can be used for commercial purposes

No clear mention if the model that was trained using the dataset for non-commercial purpose can be used commercially

Challenges in ensuring dataset license compliance

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead

Provenance related challenges | Lineage related challenges | License related challenges

Rights and obligations are unclear

Multiple license interactions and their effects are unclear

Google

Do I still need permission to use an image I found on Google Image Search?

Yes you do need permission in order to use it. Google does not own the images found via Google Search. The "Usage rights" Search tool is provided to help you find images which may be suitable for your use. It is not a grant of permission to use the images.

You must contact the owner of the image (typically whoever first posted the image on the web) and obtain his/her permission in order to use it, especially if you intend to use it publicly or commercially. Using an image without the written permission of the copyright owner can turn out to be very expensive!

flickr

4. Restrictions

You agree that you will not (i) modify or alter the Flickr Materials; (ii) create derivative works of the Flickr Materials; (iii) decompile, disassemble, decode or reverse engineer the Flickr Materials, translate the Flickr Materials or otherwise attempt to learn the source code, structure, algorithms or internal ideas underlying the Flickr Materials or reduce the Flickr Materials by any other means to a human-perceivable form; or (iv) bypass, delete or disable any copy protection mechanisms or any security mechanisms in the Flickr Materials.

Except as otherwise expressly permitted herein, you may not use the Services or the Flickr Materials to engage in any of the following prohibited activities:

- the collection, copying or distribution of any portion of the Flickr Materials;
- any resale, commercial use, commercial exploitation, distribution, public performance or public display of the Services or the Flickr Materials;
- modifying or otherwise making any derivative uses of the Services or the Flickr Materials;
- scraping or otherwise using any data mining, robots or similar data gathering or extraction methods on or in connection with the Services;
- with the exception of User Content made available by users for download, the downloading of any portion of the Flickr Materials or any information contained therein; or

The CIFAR-10 dataset

Please cite it if you intend to use this dataset.

- Learning Multiple Layers of Features from Tiny Images, Alex Krizhevsky, 2009.

Outline

Recap

License compliance process for curated datasets

Challenges

Current progress

Road ahead

Data license compliance project – Areas of interest

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead

Establish Dataset license compliance process and requirements

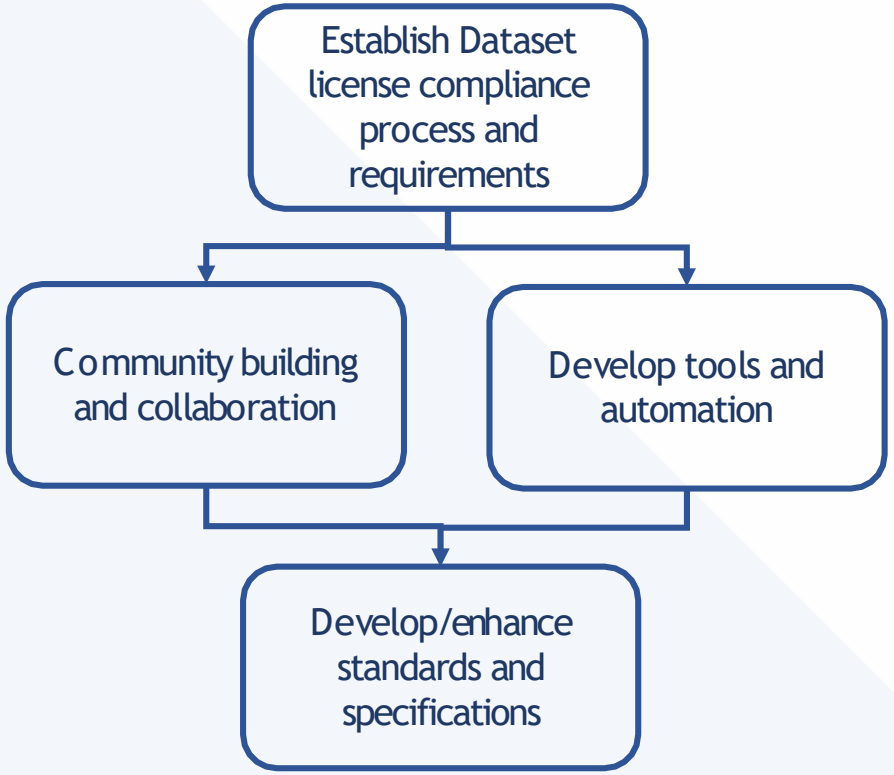
Develop tools and automation

Community building and collaboration

Develop/enhance standards and specifications

Data license compliance project – Current progress

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



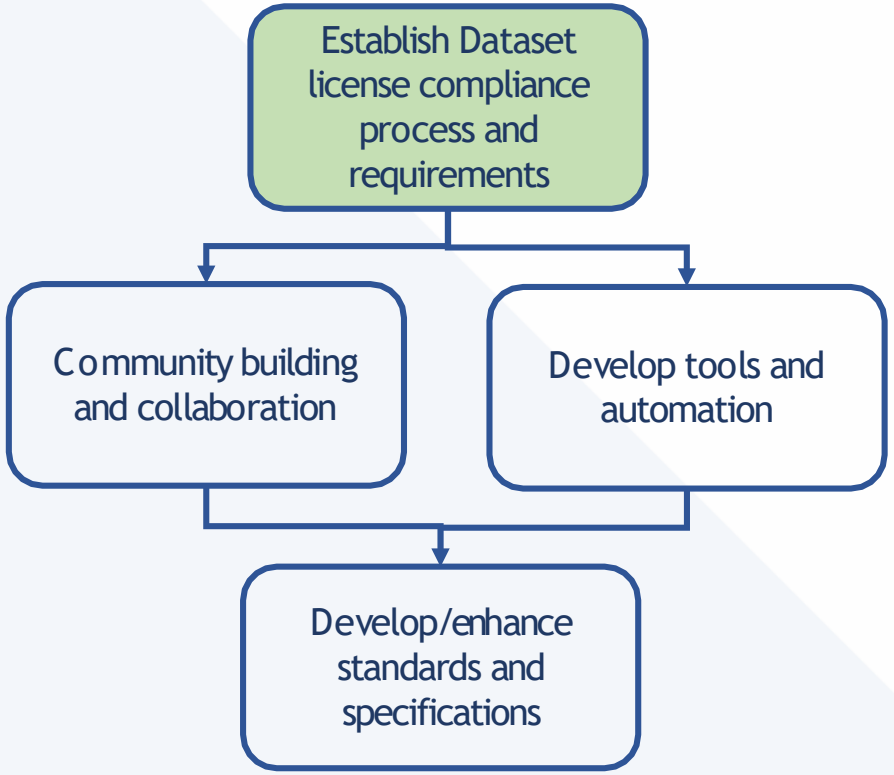
Current core contributors

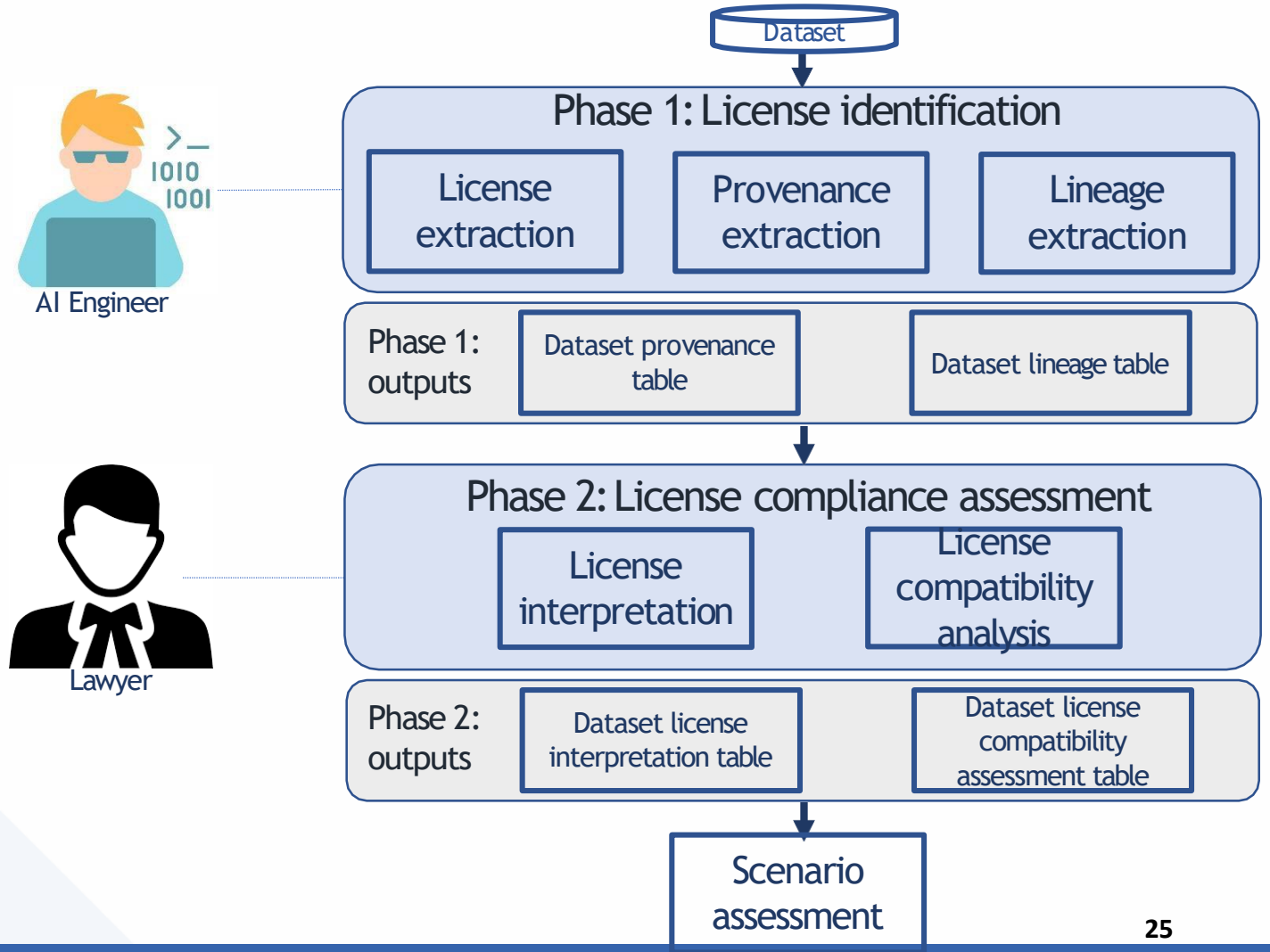
(In alphabetic order)

Boyuan Chen
Daniel M. German
Dayi Lin
Erika Tuck
Gopi Krishnan Rajbahadur
Li Zi
Song Liu
Zhengcai You
Zichen Qui
Zhen Ming (Jack) Jiang
Zhipeng Huang

Data license compliance project – Current progress

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



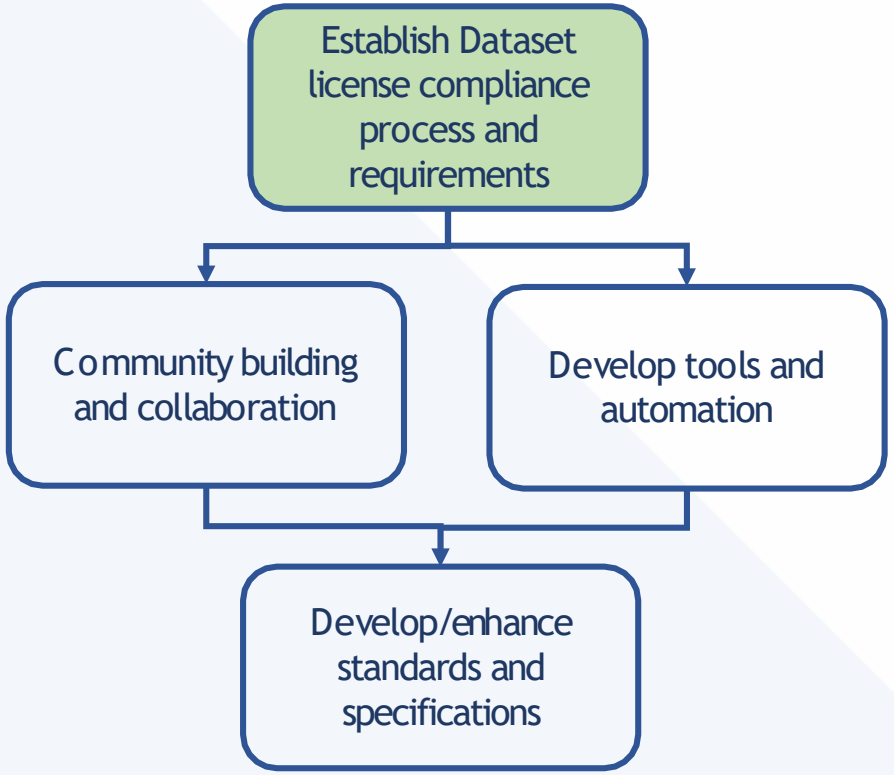


Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Data license compliance project – Current progress

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



Can I use this publicly available dataset to build commercial AI software?-A Case Study on Publicly Available Image Datasets

GOPI KRISHNAN RAJBAHADUR, Centre for Software Excellence, Huawei Canada, Canada
ERIKA TUCK, Lassonde School of Engineering, York University, Canada
LI ZI, Huawei China, Canada
DAYI LIN, Centre for Software Excellence, Huawei Canada, Canada
BOYUAN CHEN, Centre for Software Excellence, Huawei Canada, Canada
ZHEN MING (JACK) JIANG, Lassonde School of Engineering, York University, Canada
DANIEL M. GERMAN, University of Victoria, Canada

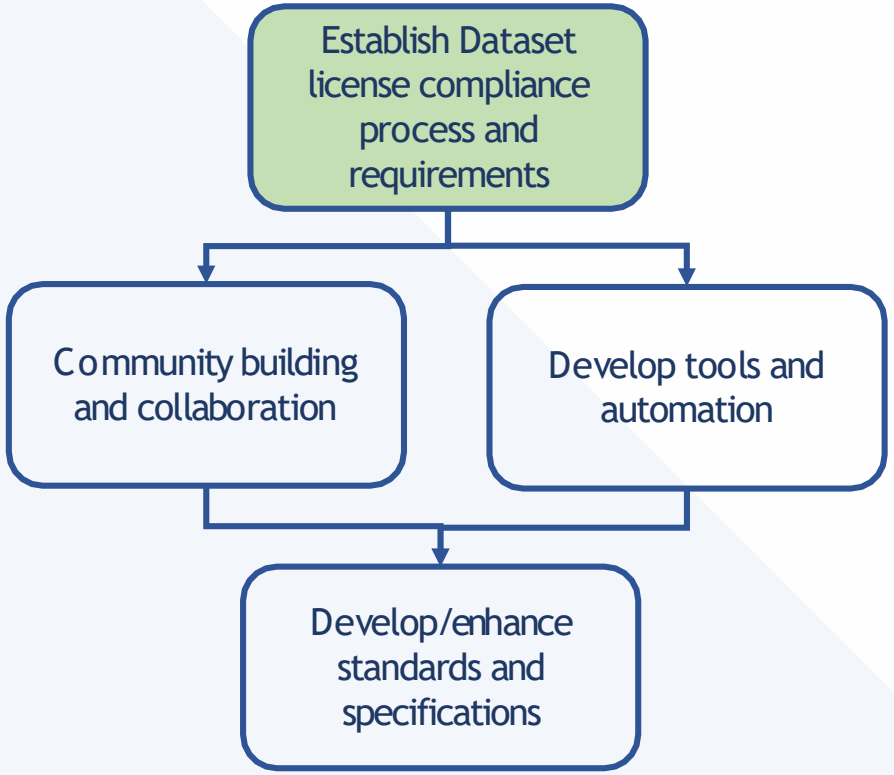
Link: https://arxiv.org/abs/2111.02374

Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Data license compliance project – Current progress

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



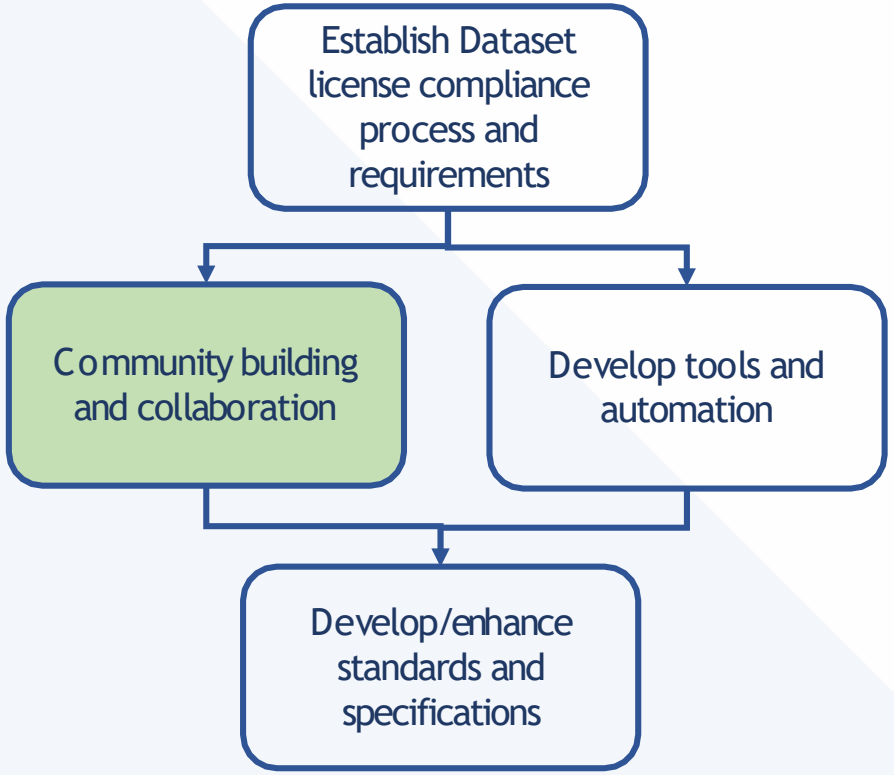
	Commercially distribute the dataset	Release a product with AI model	Commercialize the model output
IMAGENET	⊘	⊘	⊘
CITYSCAPES DATASET	⊘	⊘	⊘
VGG Face Dataset	✓	⊘	⊘
The CIFAR-10 dataset	⊘	⊘	⊘
COCO Common Objects in Context	⊘	⊘	⊘
Flickr-Faces-HQ Dataset (FFHQ)	✓	⊘	⊘

Current core contributors

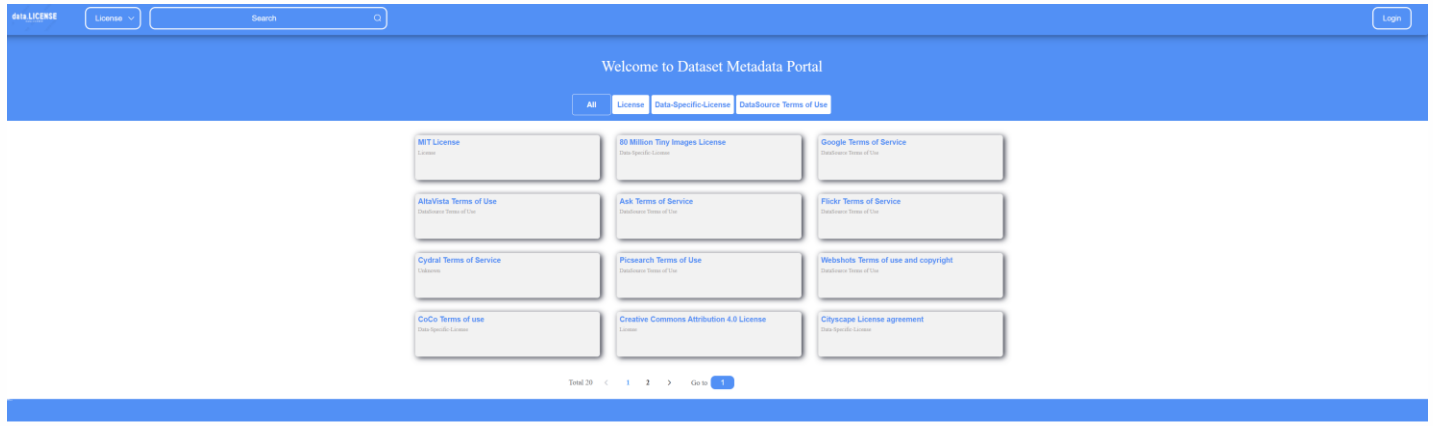
Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Data license compliance project – Current progress

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



We developed an initial version of a portal that documents dataset's license, meta-data (provenance and lineage details per our schema) and license decomposition and analysis that we have conducted
Link: http://140.83.83.152:30800/#/dataSetInfo?id=1

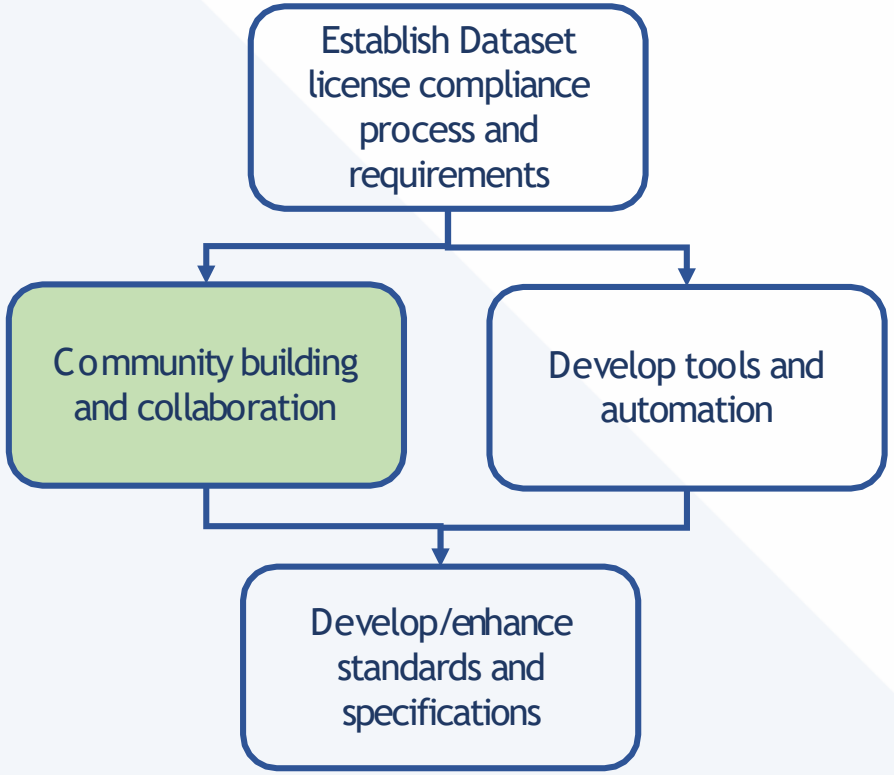


Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

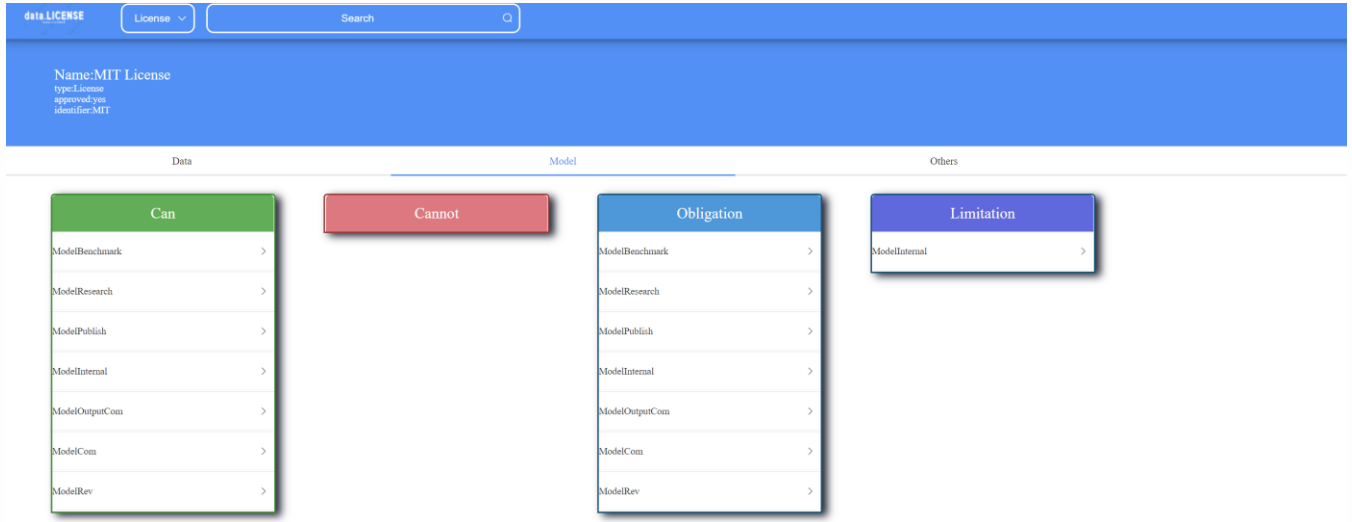
Data license compliance project – Current progress

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



We developed an initial version of a portal that documents dataset's license, meta-data (provenance and lineage details per our schema) and license decomposition and analysis that we have conducted

Link: http://140.83.83.152:30800/#/dataSetInfo?id=1

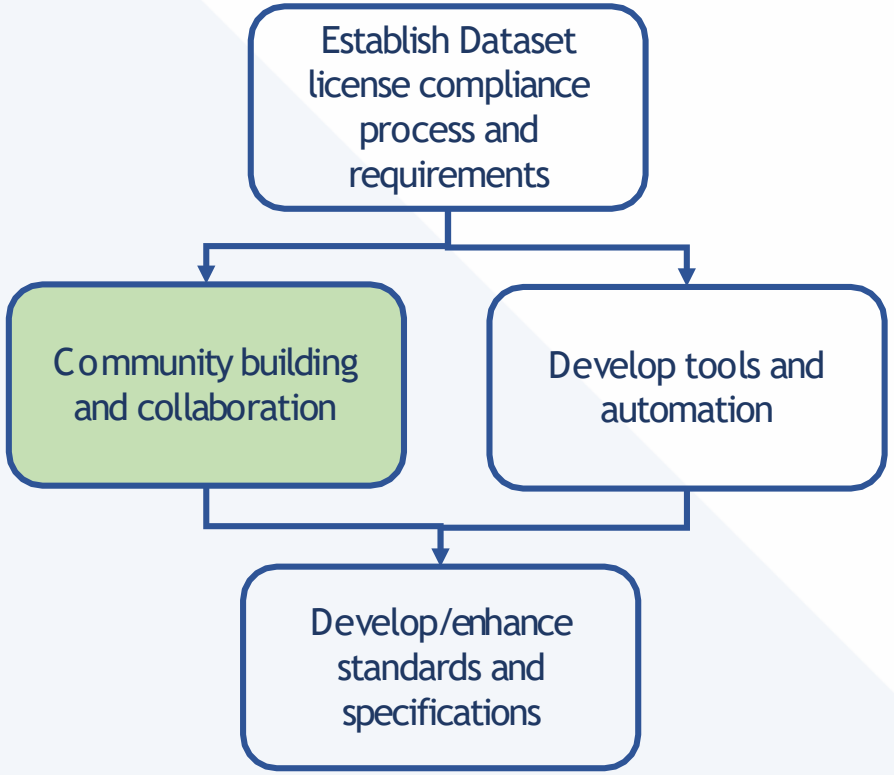


Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Data license compliance project – Current progress

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



We developed an initial version of a portal that documents dataset's license, meta-data (provenance and lineage details per our schema) and license decomposition and analysis that we have conducted
Link: http://140.83.83.152:30800/#/dataSetInfo?id=1

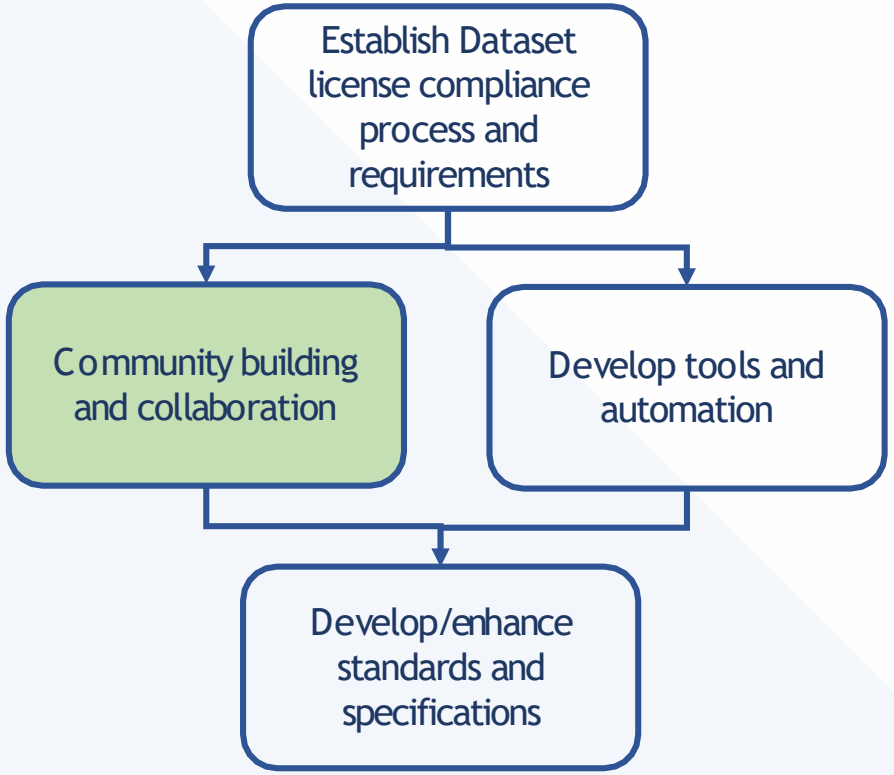
MetaData					
Name	CIFAR-10	Version	N/A	License ID	1
License Name	MIT License	Licensor	Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton	License From	Present on the official dataset website
License Location	https://www.cs.toronto.edu/~kriz/cifar.html	Origin	https://www.cs.toronto.edu/~kriz/cifar.html	Downloaded	N/A
Outlet	N/A	Size	163MB (python version); 175MB (Matlab version); 162MB (binary version)	Format	.tar.gz
Personal	unknown	Additional	N/A	Offensive	Yes
Comply		Collect	Subset of 80 Million Tiny Images	Available	1
License content	<license> <lname>cifar paper citation</lname> <hash>651A4DCDA5635BF26914F7B219B66D57</hash> </license>				
Description	"The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images"				
Collection process	"The CIFAR-10 and CIFAR-100 are labeled subsets of the 80 million tiny images dataset. They were collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton."				
Collection process	"The CIFAR-10 and CIFAR-100 are labeled subsets of the 80 million tiny images dataset. They were collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton."				

Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

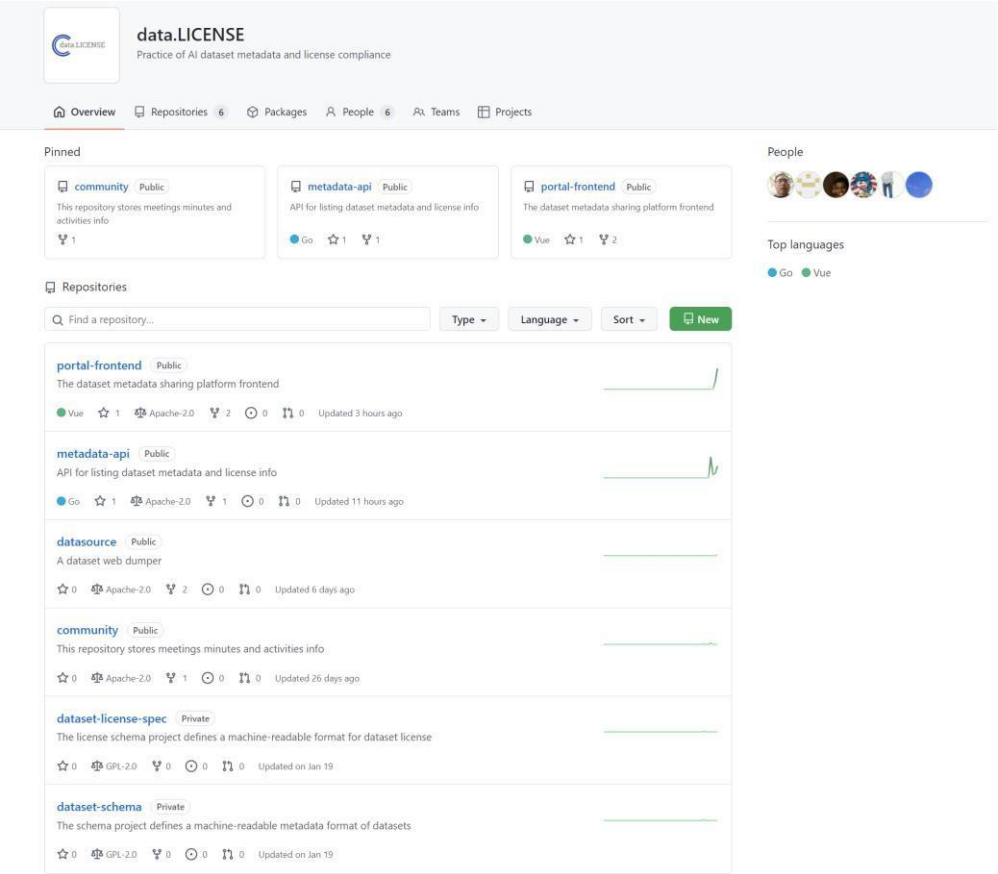
Data license compliance project – Current progress

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



GitHub

Open for collaboration and contributions

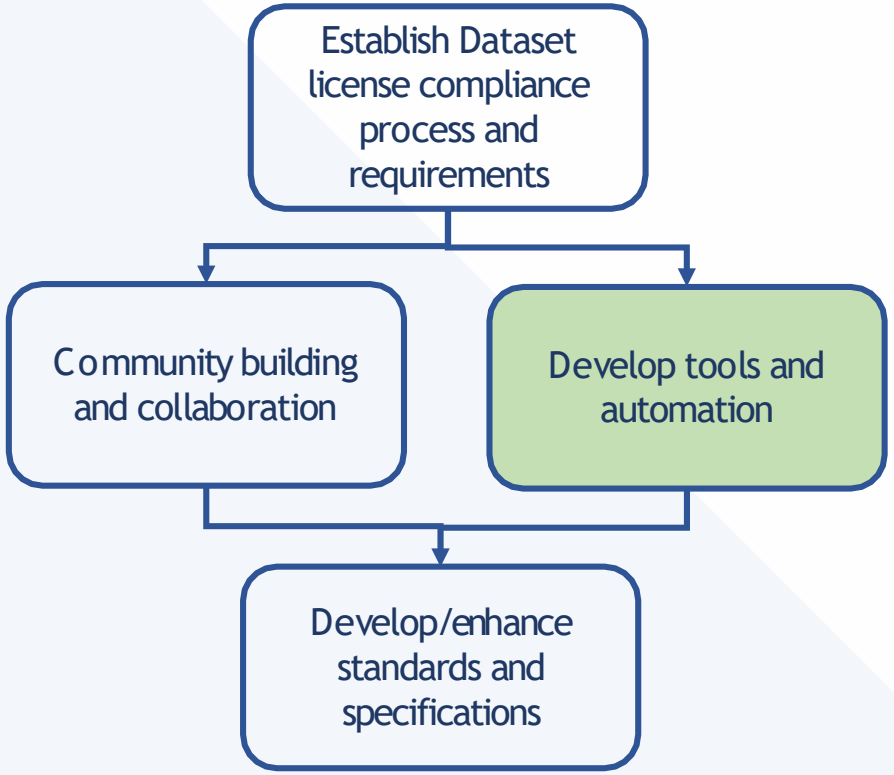


Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Data license compliance project – Current progress

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



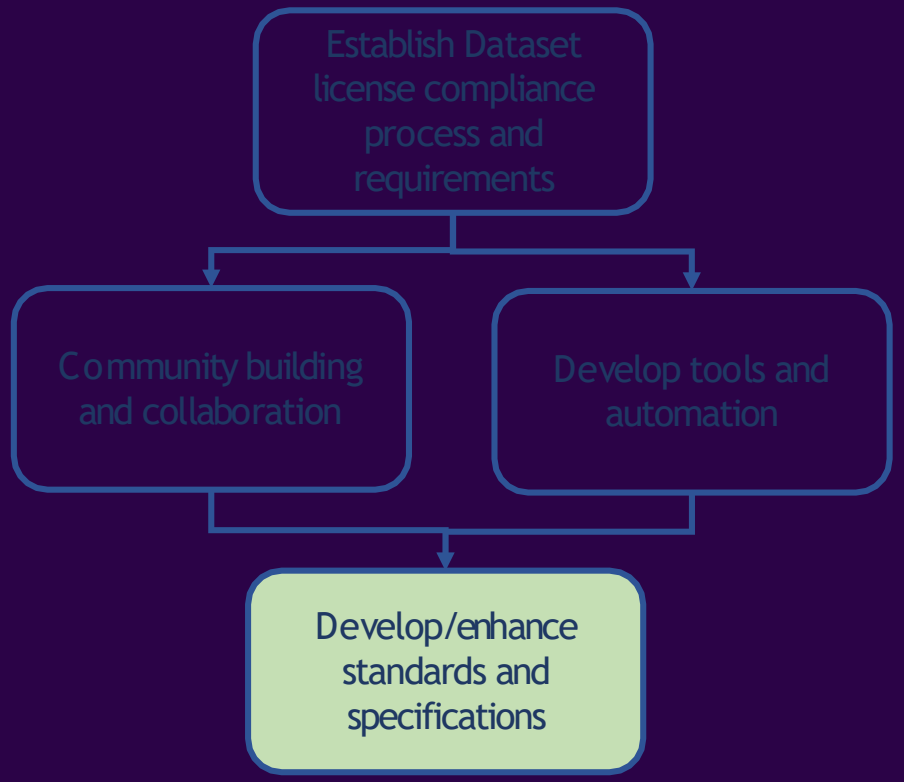
Generate machine readable, serializable formats that enable compatibility with SPDX

Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Data license compliance project – Current progress

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



Dataset-related details	**Dataset name**	**Dataset version**	**Origin date**	**Origin**
	CIFAR-10	N/A	2009	https://www.cs.toronto.edu/~kriz/cifar.html
	Description of dataset		**Description of data collection process**	
	The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images		The CIFAR-10 and CIFAR-100 are labeled subsets of the 80 million tiny images dataset. They were collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton.	
	Downloaded outlet	**Is outlet licensed?**	**Is dataset publicly available?**	**Additional notes**
	N/A	N/A	Yes	This dataset is a subset of another dataset called 80 Million Tiny Images
License-related details	**Where license was found**		**License location**	**License content**
	Present on the official dataset website		https://www.cs.toronto.edu/~kriz/cifar.html	(not pasting content due to space)
Metadata	**Hashcode**		**Size**	**Format**
	MD5: c58f30108f718f92721af3b95e74349a (Python version)		163MB (Python version)	tar.gz

License metadata	**Licensor**		**License name**		**Dataset name**		**Dataset version**
	Alex Krizhevsky		Custom license		CIFAR-10		N/A
	Credit/Attribution Notice						
	Learning Multiple Layers of Features from Tiny Images, Alex Krizhevsky, 2009.						
	License validity period		**Liability /Warranty**		**Designated third parties**		**Additional conditions**
	N/A		N/A		Only by agreement		None
Data (standalone)	**Access**		**Tagging**		**Distribute**		**Re-represent**
Rights	✓		✓		✓		✓
Obligations	Cite paper		Cite paper		Cite paper		Cite paper
Data rights in conjunction with model					**Commercialization**		**Model Reverse Engineer**
	Bench-mark	**Re-search**	**Publish**	**In-ternal Use**	**Out-put**	**Model**	
Rights	✓	✓	✓	✓	✓	✓	✓
Obligations	Cite paper	Cite paper	Cite paper	Cite paper	Cite paper	Cite paper	Cite paper

We propose initial version of the standard to record details about a dataset's provenance, lineage and license that will enable anyone to conduct dataset license compliance analysis.

We welcome feedback!

Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Outline

Recap

License compliance process for curated datasets

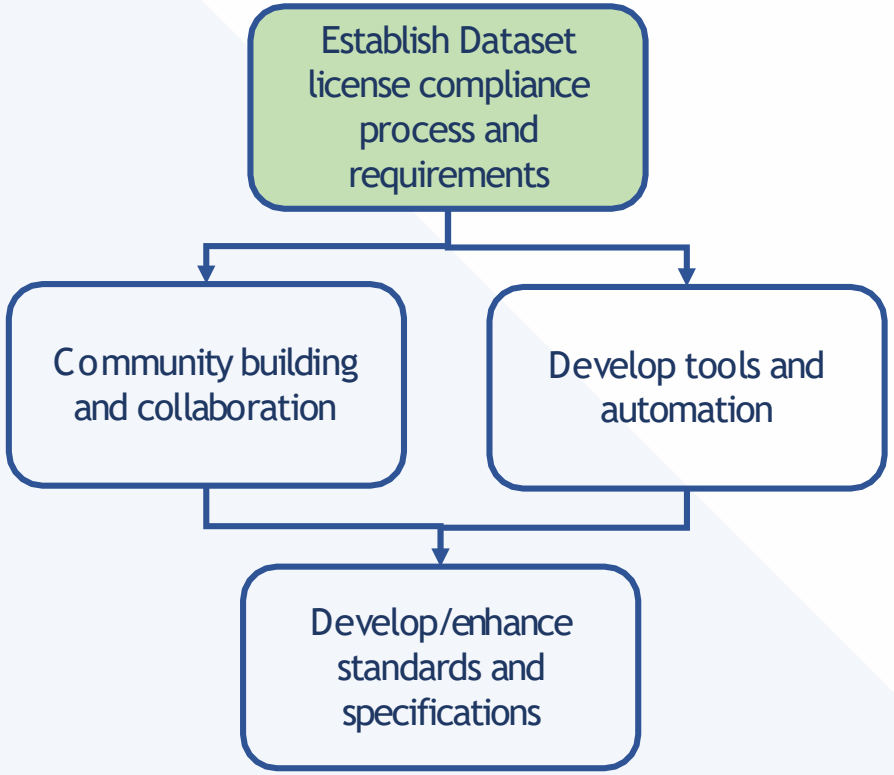
Challenges

Current progress

Road ahead

Data license compliance project – Look ahead

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



The first step is to establish dataset compliance process for various requirements like

License compliance

Copyright compliance

Privacy compliance

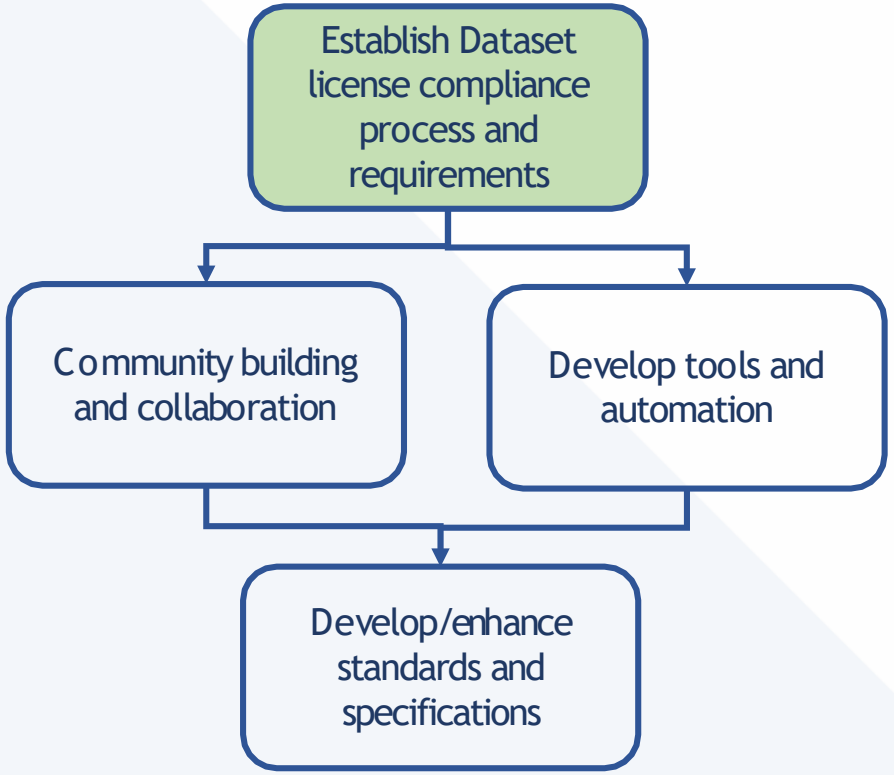
Ethics compliance

Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Data license compliance project - Look ahead

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



The first step is to establish dataset compliance process for various requirements like

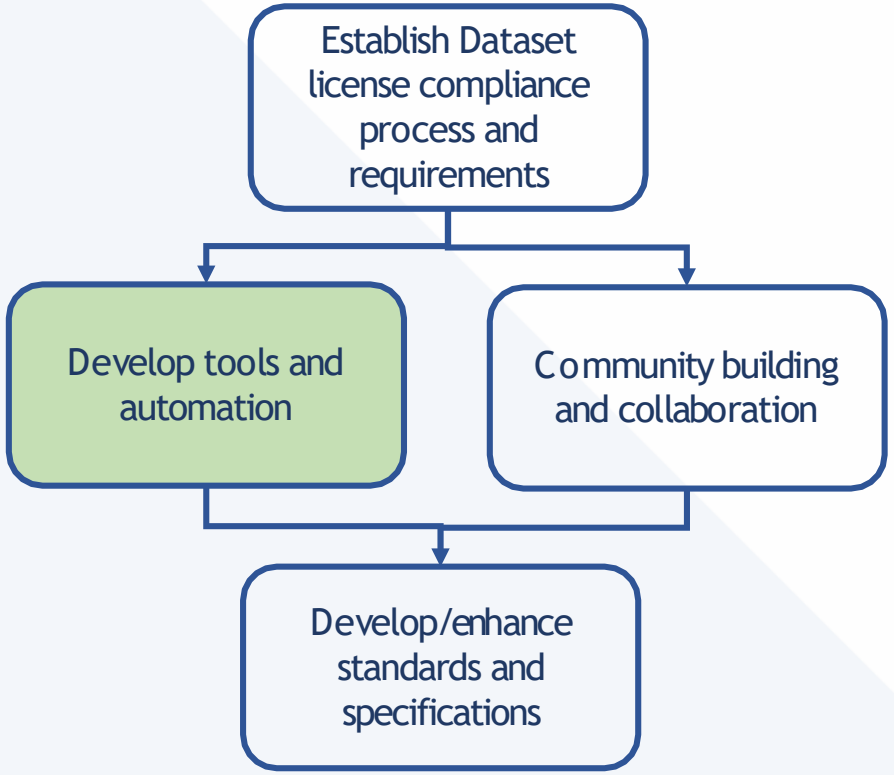
- 2022-Q2 License compliance
- 2022-Q4 Copyright compliance
- 2024-Q4 Privacy compliance — HELP?
- 2025-Q2 Ethics compliance — HELP?

Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Data license compliance project - Look ahead

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



We aim to develop various tools and automation procedures such as

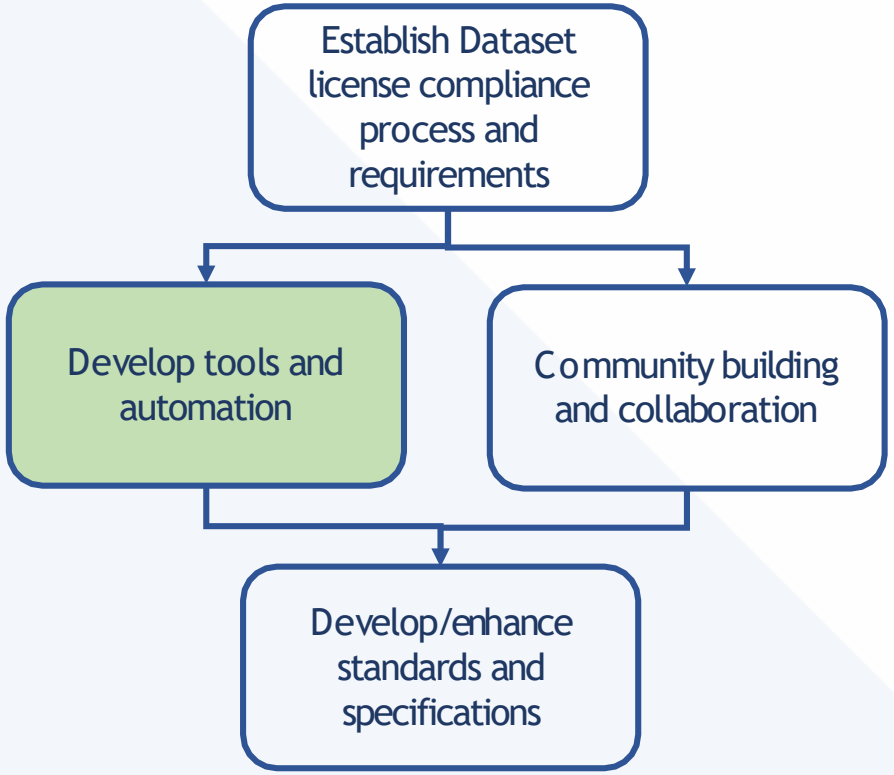
- Automated license generator
- Ensuring compliance through data clone detection
- Automated provenance extraction
- Automated lineage extraction
- License Compliance process automation

Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Data license compliance project - Look ahead

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



We aim to develop various tools and automation procedures such as

Automated license generator

A tool that helps users specify the rights and obligations and generate a license based on the chosen right license

Automated provenance extraction

Automated lineage extraction

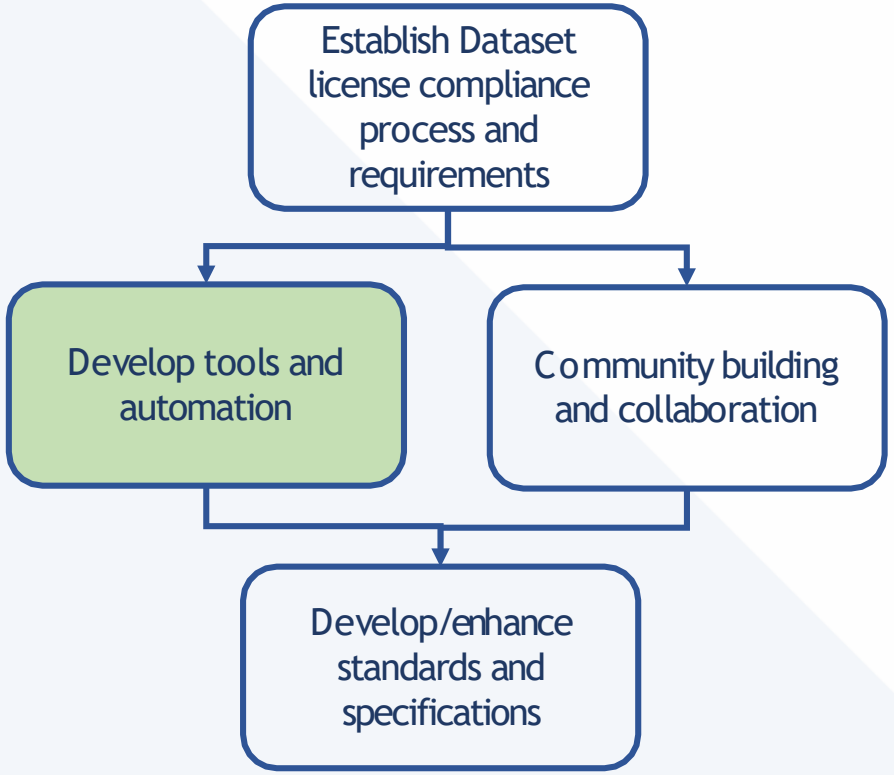
Tools that helps users extract and document the provenance and lineage details of datasets automatically using NLP on relevant documents and websites

Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

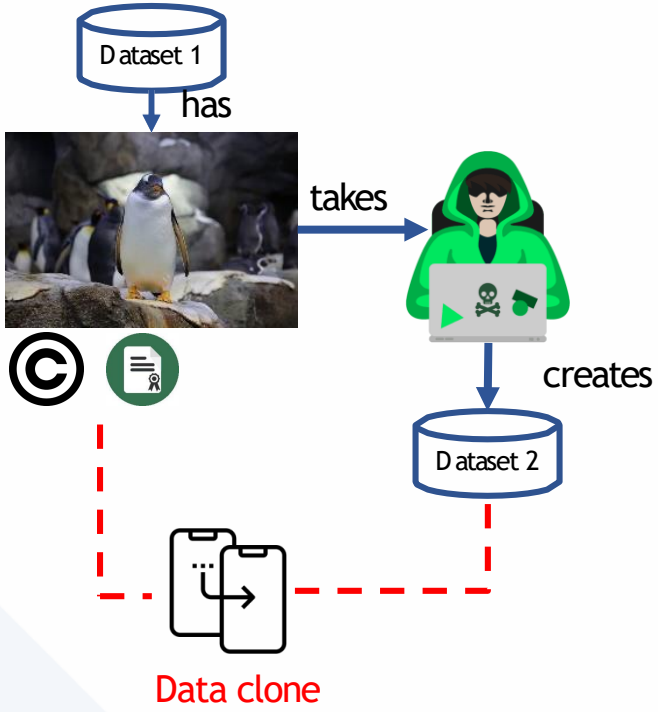
Data license compliance project - Look ahead

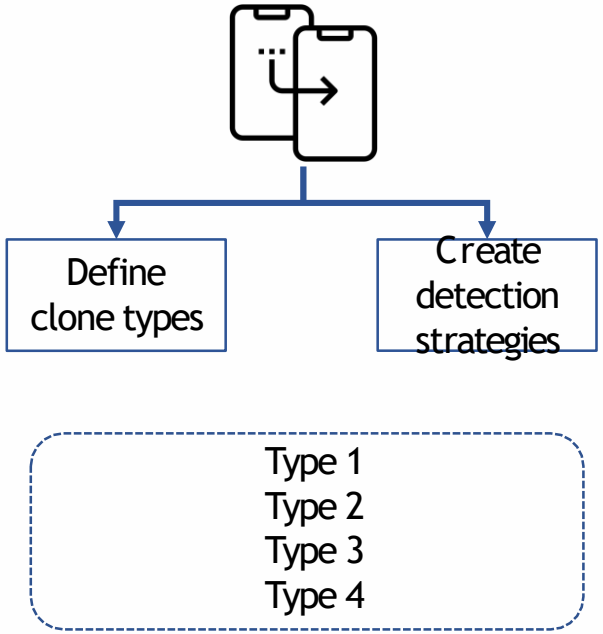
Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



We aim to develop various tools and automation procedures such as

Ensuring compliance through data clone detection



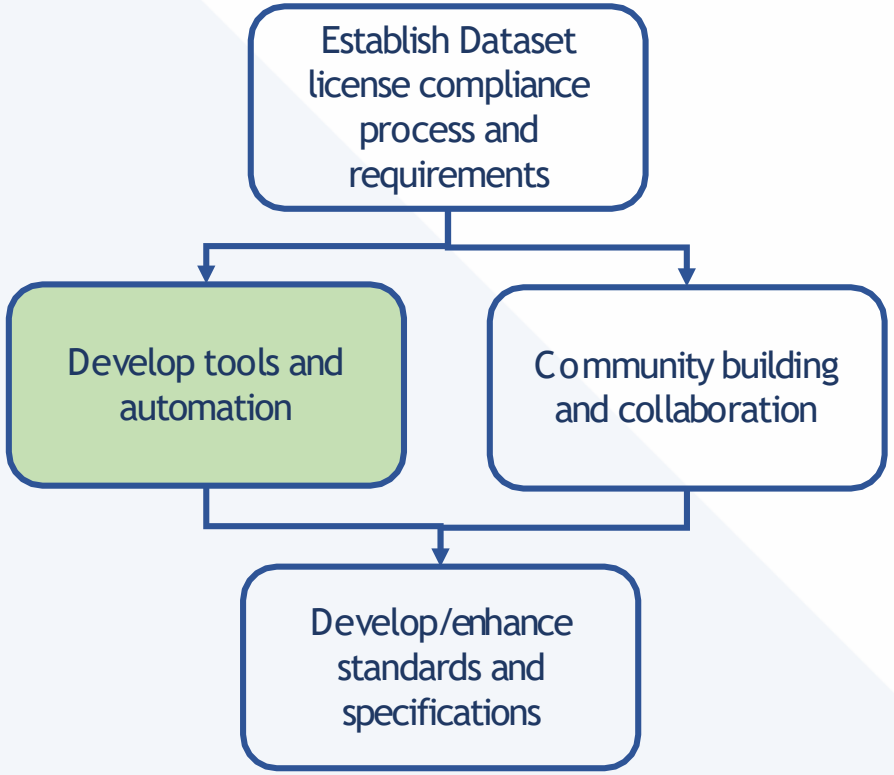


Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Data license compliance project – Look ahead

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



We aim to develop various tools and automation procedures such as

2022-Q2
Automated license generator

2023-Q1
Ensuring compliance through data clone detection

2023-Q4
Automated provenance extraction

2024-Q1
Automated lineage extraction

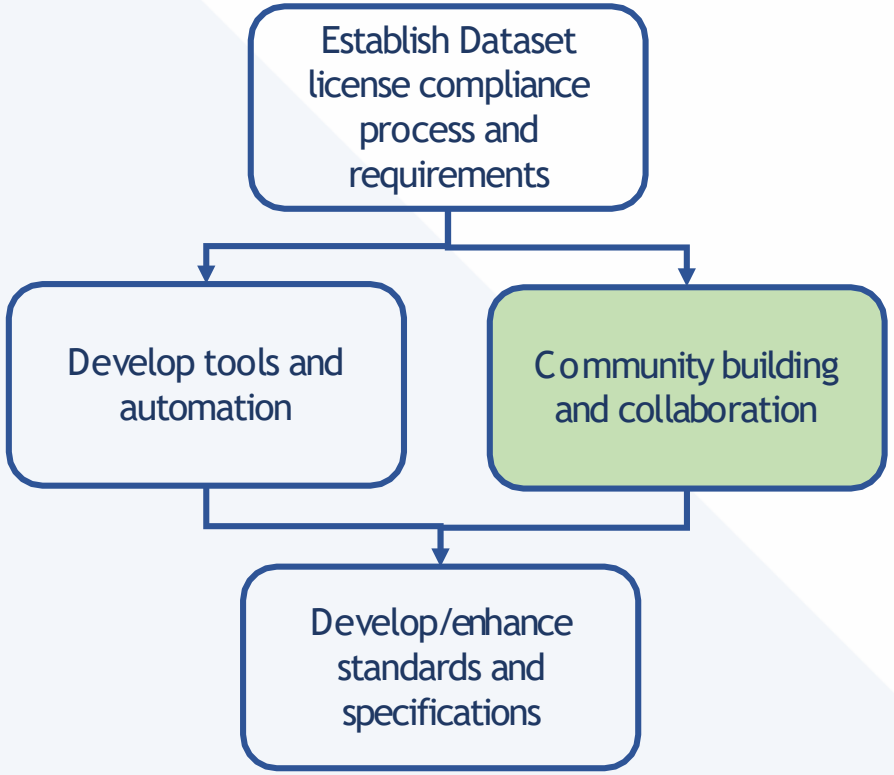
2024-Q4
License Compliance process automation

Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Data license compliance project – Look ahead

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



We aim to develop various tools and automation procedures such as

Invite contributors and onboard them

Invite legal experts to help contribute

Establish moderation and governance policy

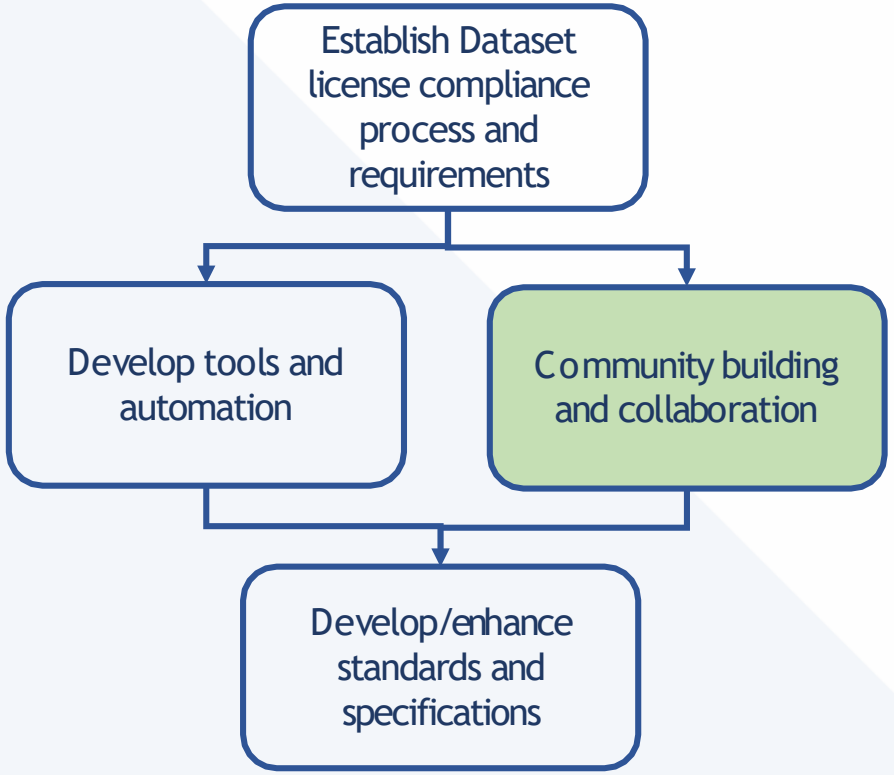
Establish wiki and forum for active discussion

Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Data license compliance project – Look ahead

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



We aim to develop various tools and automation procedures such as

2022-Q2
Invite contributors and onboard them

2022-Q4
Invite legal experts to help contribute

2022-Q3
Establish moderation and governance policy

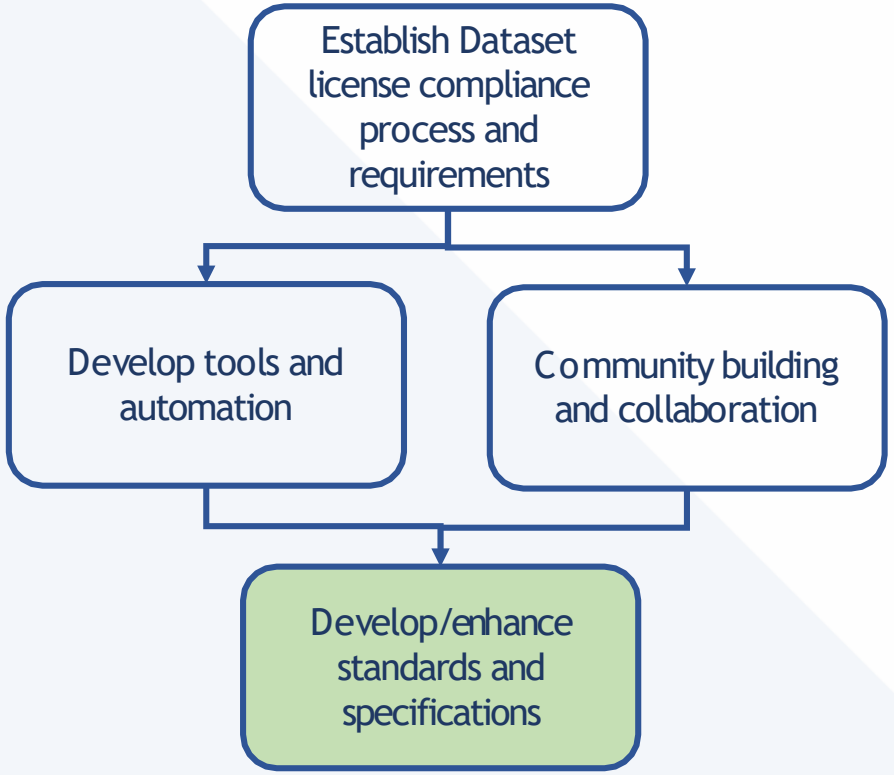
2022-Q3
Establish wiki and forum for active discussion

Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Data license compliance project – Look ahead

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



We aim to develop various tools and automation procedures such as

SPDX, SBOM — Enhance existing standards

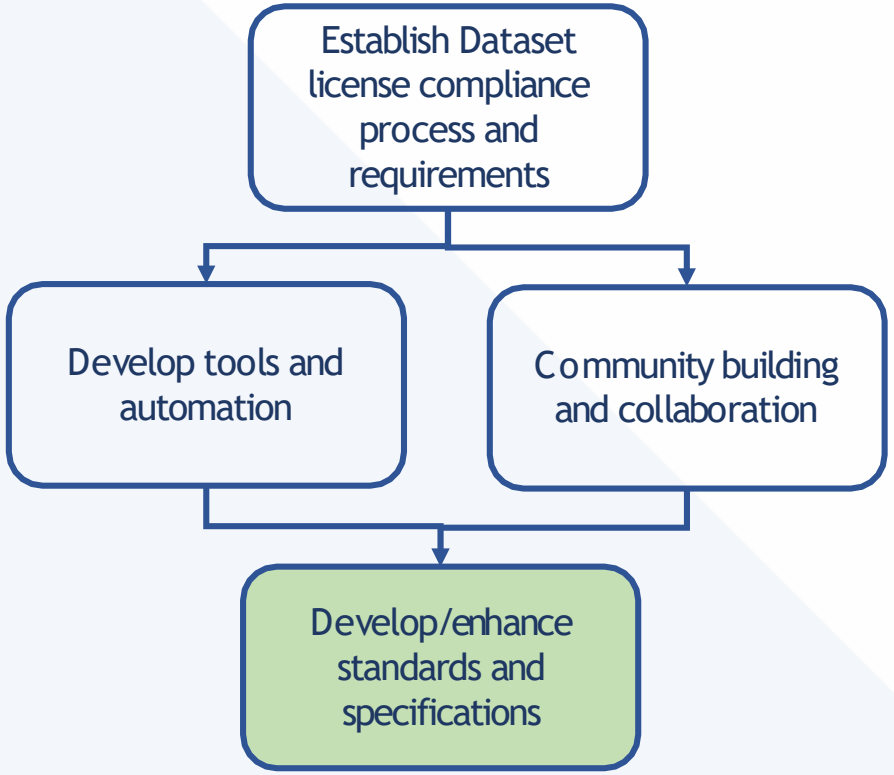
ISO, IEEE SA STANDARDS ASSOCIATION — Create new standards

Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Data license compliance project – Look ahead

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead

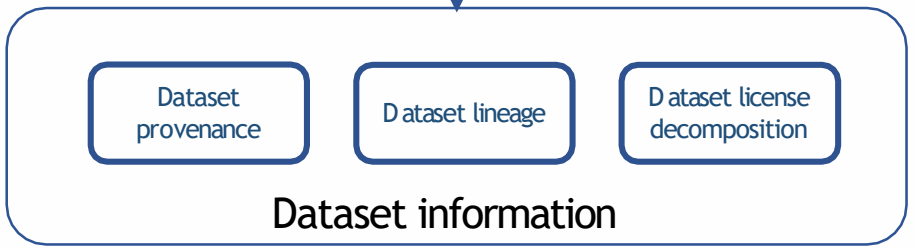


We aim to develop various tools and automation procedures such as

SPDX SBOM

Enhance existing standards

Enhance with

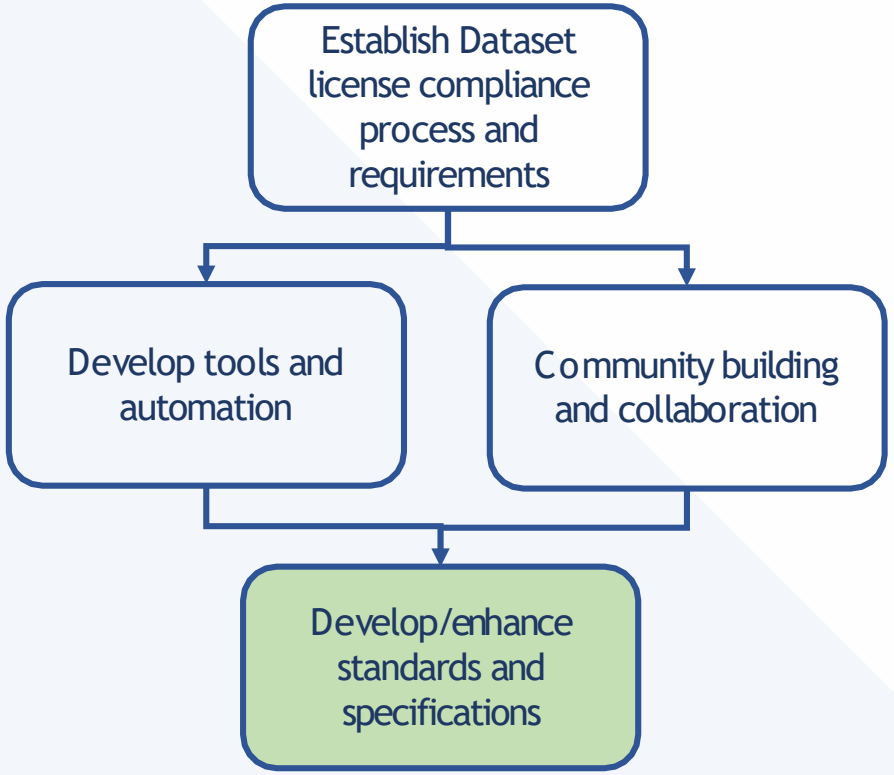


Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Data license compliance project – Look ahead

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead



We aim to develop various tools and automation procedures such as

SPDX / **SBOM** — 2022-Q4
Enhance existing standards

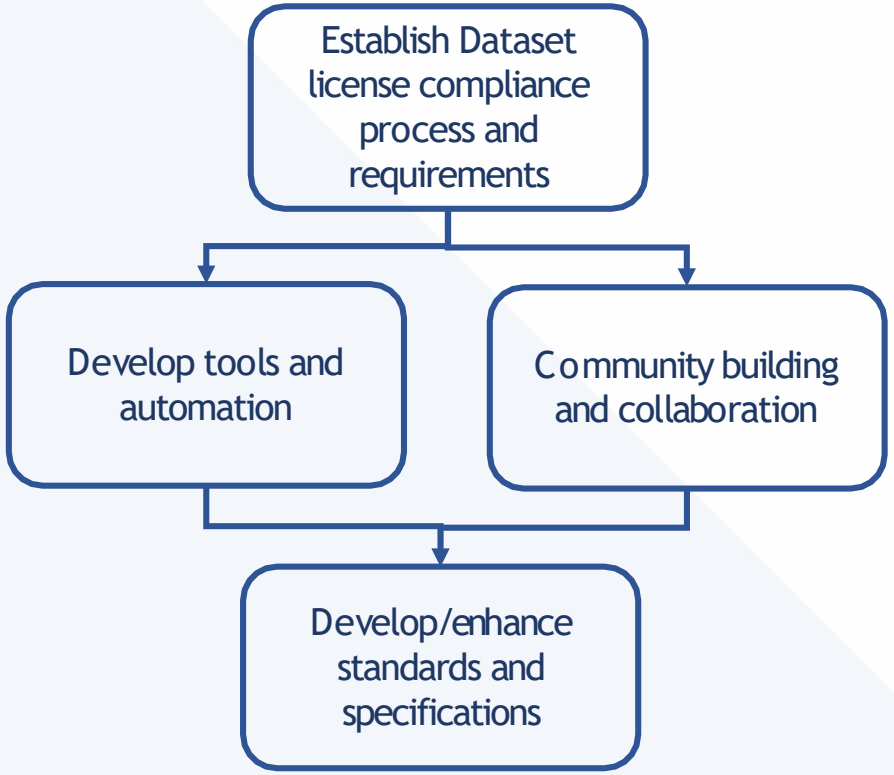
ISO / **IEEE SA** STANDARDS ASSOCIATION — 2023-Q4
Create new standards

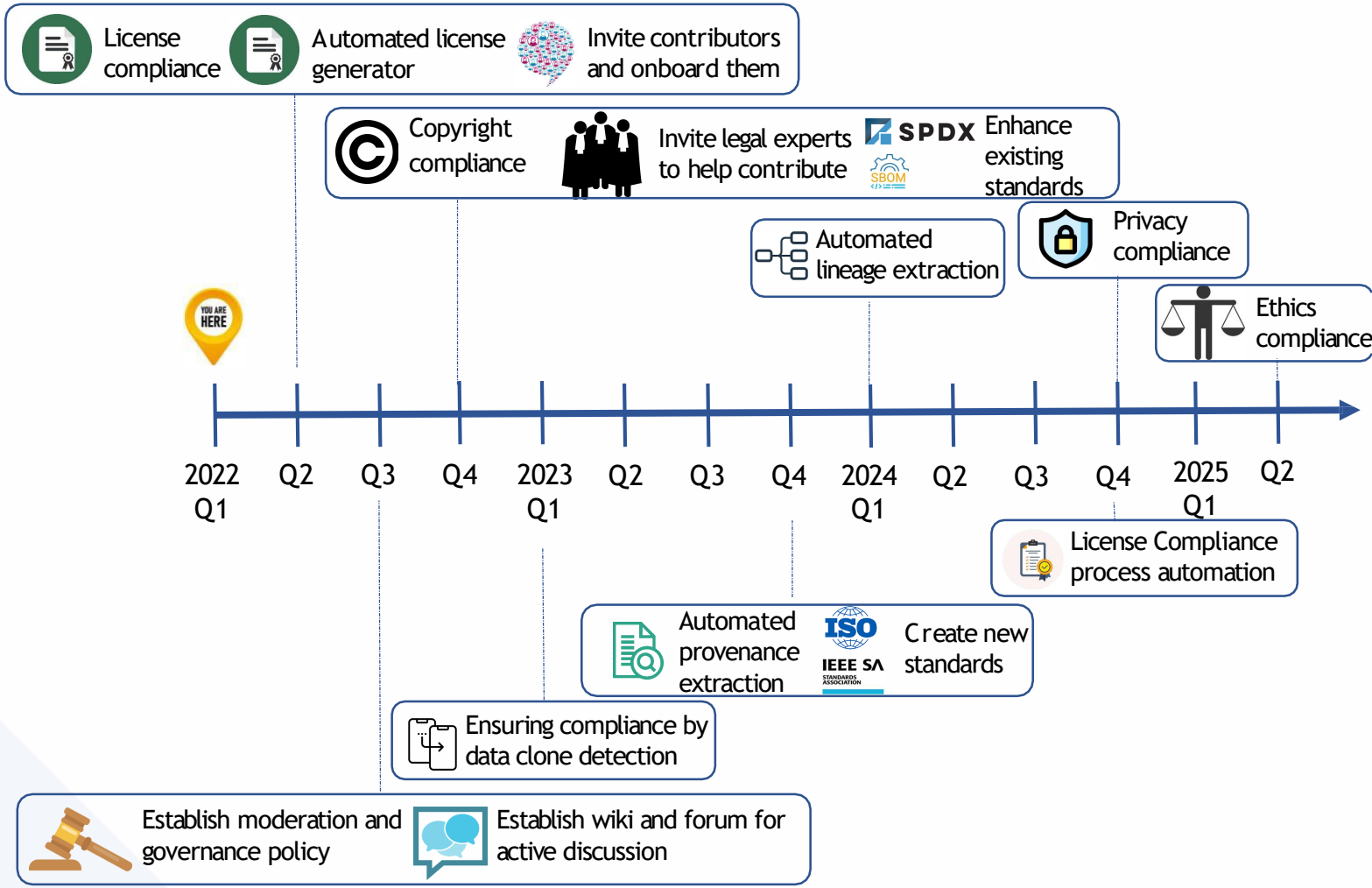
Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

Data license compliance project – Look ahead

Recap | License compliance process for curated datasets | Challenges | Current progress | Road ahead





Current core contributors

Boyuan Chen, Daniel M. German, Dayi Lin, Erika Tuck, Gopi Krishnan Rajbahadur, Li Zi, Song Liu, Zhengcai You, Zichen Qui, Zhen Ming (Jack) Jiang, Zhipeng Huang

THOTH

https://wiki.anuket.io/display/HOME/Thoth

Thoth is an Egyptian God of Learning and Reckoning.

Egyptian God's name was chosen to match with the Parent-Project's name (Anuket)

6 is the number of Thoth, and Ibis/Beak-of-Ibis is one of the symbols of Thoth – Our Logo captures both !!!!!

WHY THOTH?

YES: Frameworks, Tools, ML-model implementations, etc. (landscape of LF-AI)

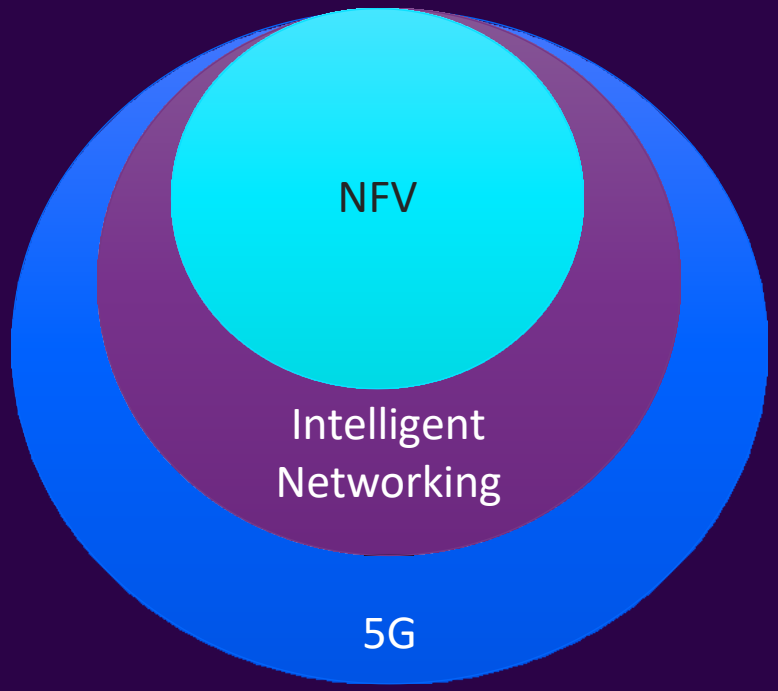
NO: OS ML-Models for NFV problems

AI/ML FOR NFV PROBLEMS

- Detection & Analysis
  - Patterns, Trends, Correlation, etc.
  - Anomalies, Causality, etc.
- Predictions
  - Failure
  - Resource availability
    - Traffic Engineering
    - Application Placement
    - Auto-Scaling
    - SLA-Mgmt.
- Capability Planning
- Detailed List: https://github.com/opnfv/thoth/blob/master/research-studies/ml-problems-techniques-nfv.md

THE SCOPE

AI/ML PROBLEMS



THOTH OVERVIEW

Decision Driven Data Analytics for NFV Usecases

Software Development

Develop Source code – Models and Tools.

Research Studies

The nature of the domain demands systematic studies to take educated decisions.

Collaborate

Collaboration with Telcos, academic researchers & OSS projects with Testbeds.

Model As A Service

Providers share dataset & the problem in hand, Thoth will build, assess and deliver the ML model.

NOUNS

ML Problem

ML Models & Support Tools

Training and Testing Dataset

VERBS

Research

Collaborate

Development & Deployment

ROADMAP (TENTATIVE)

2021

Jan Feb Mar Apr May June July Aug Sep Oct Nov Dec

Discussions with different stakeholders

Start!

Publish Research Studies

Publish Failure Prediction Model

Phase-1 of Log Analysis and GANs

F/W for MaaS

2022

Jan Feb Mar Apr May June July Aug Sep Oct Nov Dec

Phase-2 of Log Analysis and GANs

Models for Cloud Native FP: Containers Anomaly Detection

Models for Cloud Native: Loss Characterization. FP: Service

Initial Proposal for Data Modelling

Research Study on Data

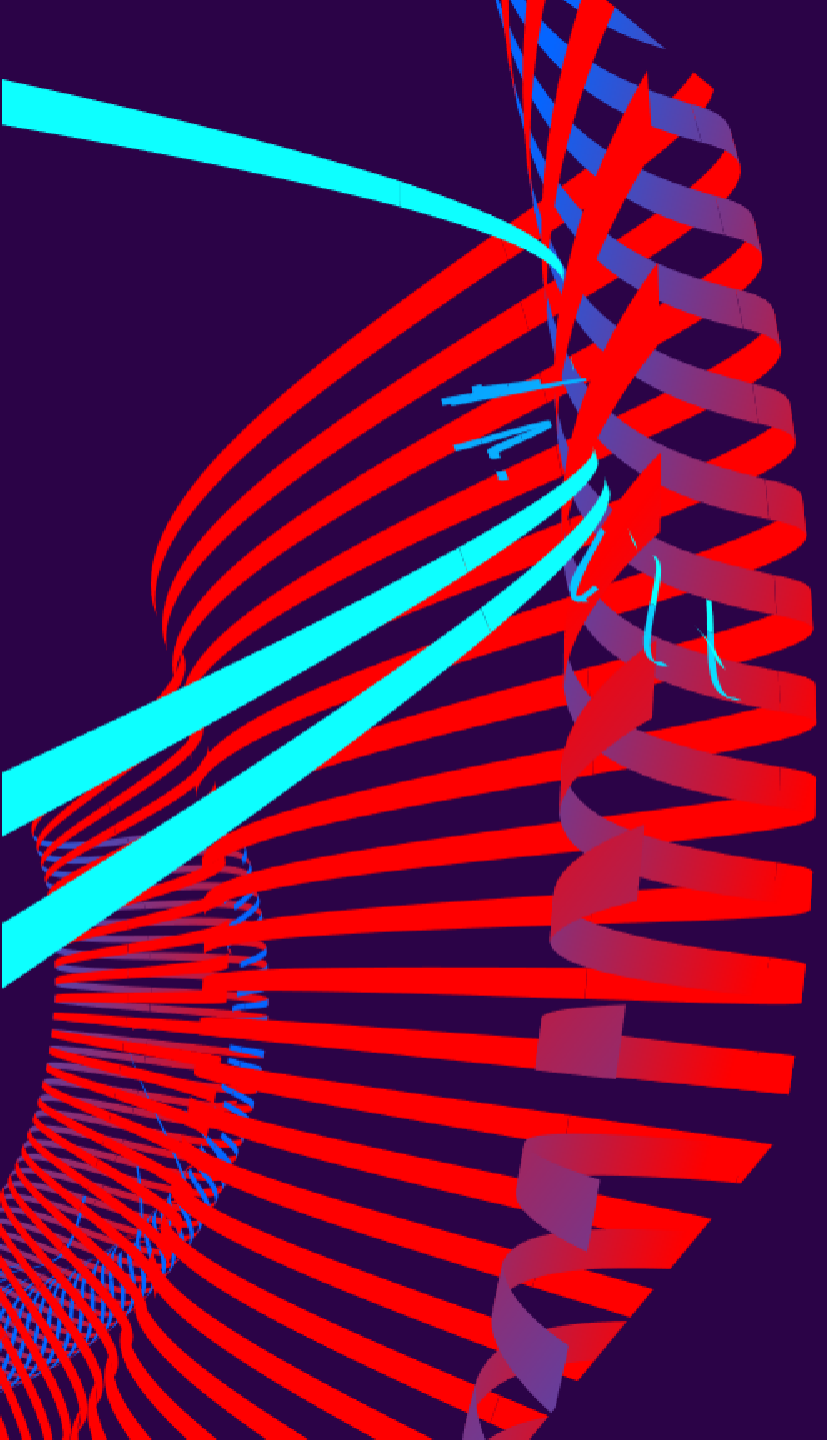
V1 for MaaS: Supervised

Moselle Release

V2 for MaaS: Unsupervised

V3 for MaaS: Reinforcement

Model As A Service



RESEARCH STUDIES

STATUS

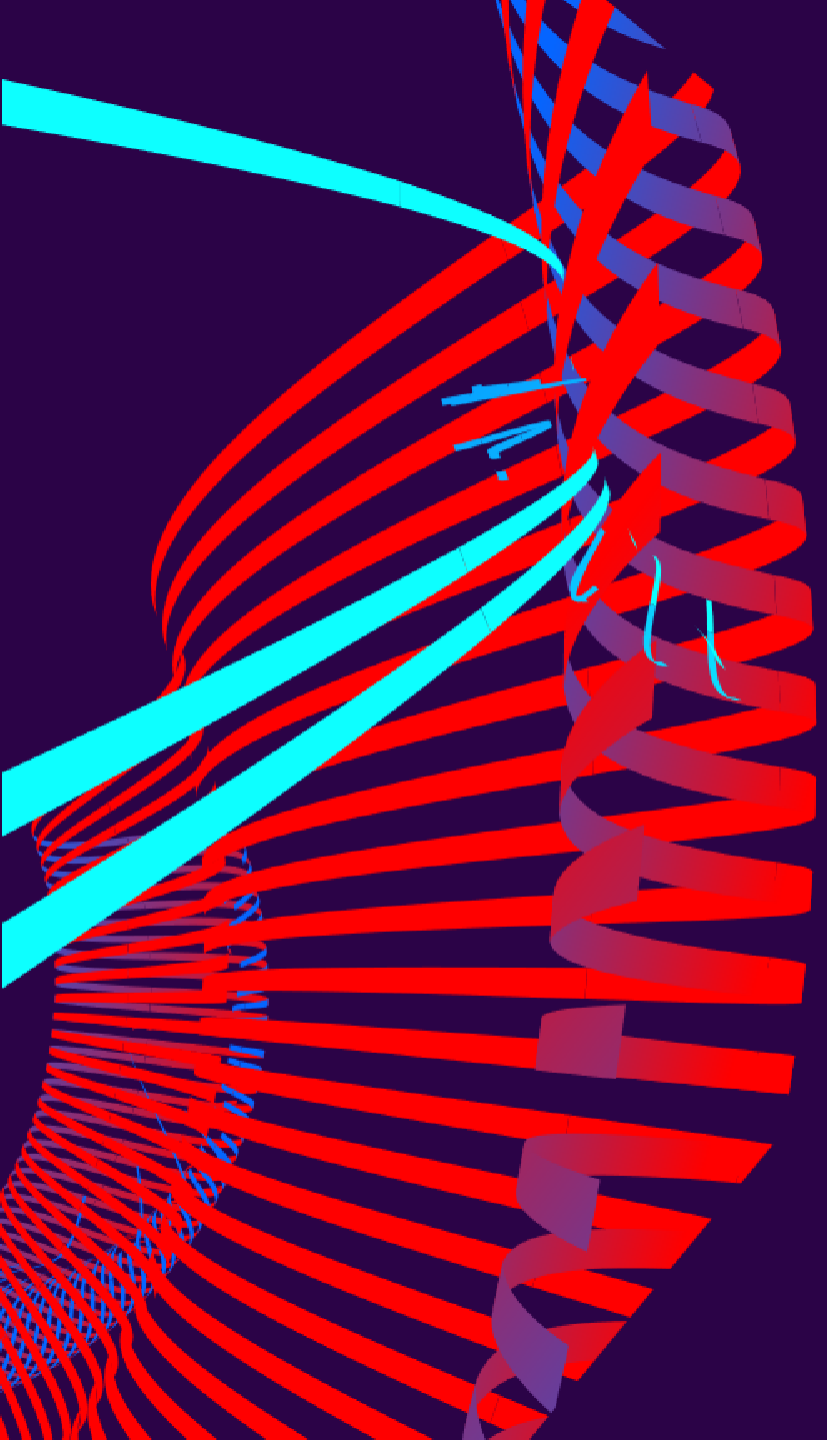
PUBLISHED

State of Art: Machine Learning Problems in domain of NFV, and corresponding techniques.

State of Art: Opensource Projects for AI/ML for NFV.

W.I.P

Sources of Data – their formats and meaning.



MODELS

ABOUT MODELS

Opensource – Python, Jupyter Notebooks.

No constraints on Data access – Filesystem, Databases, Repositories, etc.

No Constraints of frameworks, tools and Libraries. Ex: though Tensorflow is used for initial models – its not mandatory.

No Constraints on Problem-Domain or ML Technology - Contributors can decide which problem they want to solve and in turn which model they want to build - focus on novelty and better-performance.

MODELS
[PHASE-1 TARGET]

ANALYSIS

Log Analysis
Correlation

DETECTION

Anomaly

PREDICTION

Failure

GENERATION

Synthetic
Telemetry Data

MODEL STATUS

PUBLISHED

VM Failure Prediction:

Decision Tree, LSTM, LSTM-Attention, LSTM-Correlation

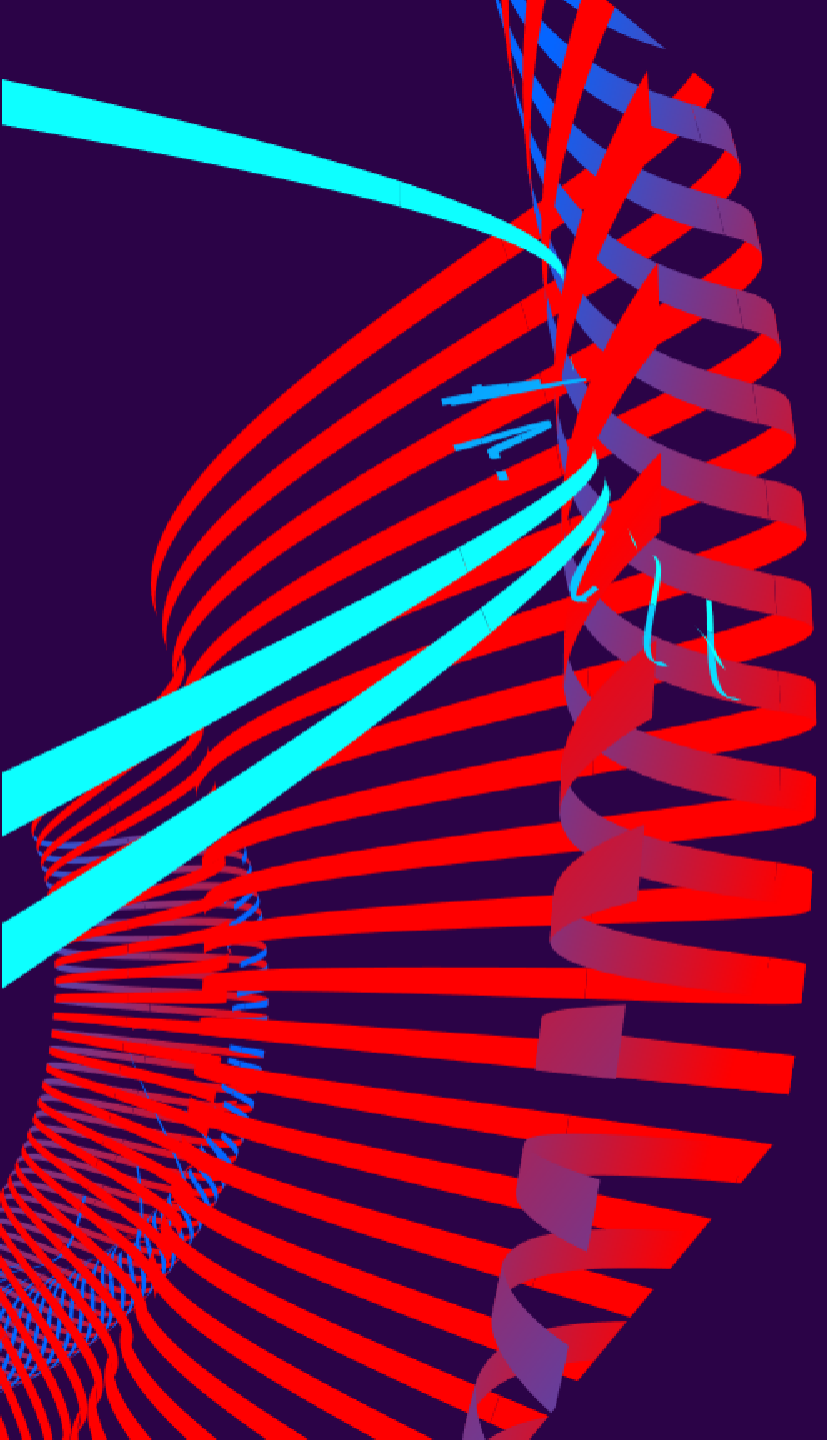
W.I.P

Google BERT for Log Analysis.

GANs for Synthetic Data Generation

CONTRIBUTORS

Students



DATA

AI/ML FOR NFV: DATA

- Q: What is it?
  - Answer is simple!
    - Metrics & Logs
    - Per-Component Statistics (Ex: vSwitch, SDN-Controller, VIM, Orchestrator, etc.)
- Actually, it is not so simple. For each problem ..
  - What are all the metrics?
  - What are all the logs?
  - What are all the Components and specific stats ?.

AIML FOR NFV: FAILURE PREDICTION DATA

Type	Parameter	Where do we find the "Failure" data?		
		Openstack Logs	Kubernetes Logs	Metrics
Links	Down or Removed	l2/l3-agents, neutron-*, virtual-switch/bridge-*	virtual-switch/bridge-*, kube-proxy, CNI logs,	Infrastructure Metrics. *** Inference of Failure from the Metrics ***
VM	Failed to start Failed to boot Shutdown Crash, Hang, Panic Unresponsive*	Links + Nova-*, libvirt, neutron-server, glance, cinder,	Links+Pod-logs	
Container		(open)	Links + OS layer – syslog, boot.log, kern.log etc. Kubernetes Layer – container Logs	
Node	Unresponsive/Unreachable/Service Failure, Crash/hang/Panic	System Logs, Service Logs (ex: nova, neutron, kubelet, kube-proxy) , SNMP events,		
App	Unresponsive, unreachable, crash/failure	Above + (open)	Above+(open)	

SHO WSTO PPER ….

As predicted, availability of the dataset.

SOLVING THE 'DATASET' PROBLEM

REQUEST

Collaborate with …

Research-labs and Operations Teams in Telcos, Vendors and Service providers, LFN Labs, Other Opensource Projects with Testbeds,

GENERATE

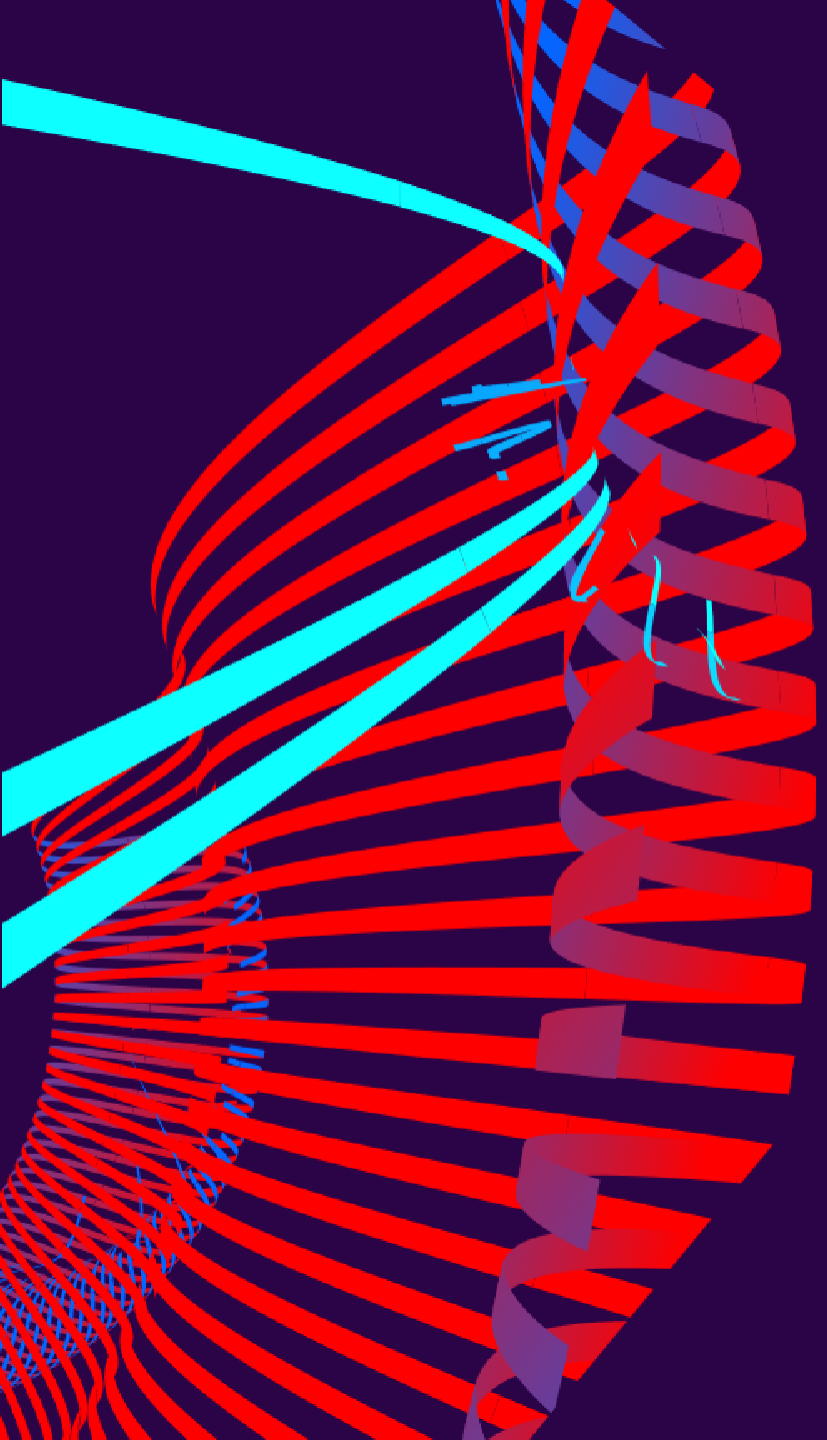
Create Testbeds …

Openstack, Kubernetes, opensource tools.

EMULATE

Synthetic Data

Again, using ML! (GANs)



TOOLS

STATUS

PUBLISHED

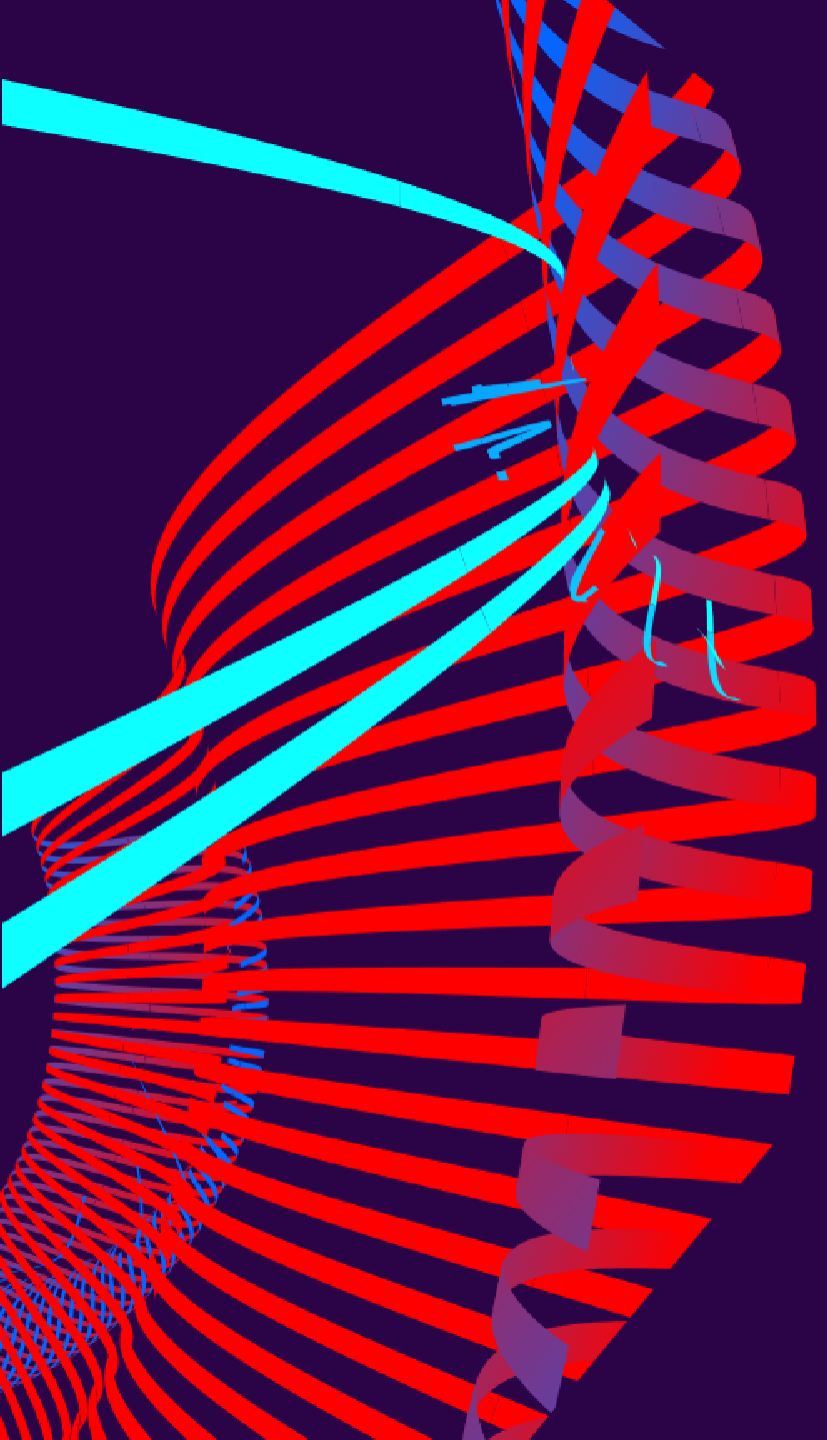
Model Selector : Q&A based CLI-Wizard tool to suggest the user which ML-approach would be better for the Data and the problem in hand.

Data Extractor: Extract Data (time-based or size-based) from well-known databases – Prometheus, Elasticsearch

W.I.P

Data Anonymizer.

TVLV Workload Generator – Synthetic Data Generation.



FAQs

Q &A

- Do you use any existing ML-Frameworks?

Not yet. We are still trying different frameworks. We wanted to start with LF-Acumos, but, getting it installed and running proved very time consuming.

- Can your models be run on any existing ML frameworks.

May not be AS IS as the models are built using Jupyter Notebooks. However, as we use Tensorflow, and majority of frameworks support Tensorflow, the integration should be straight-forward.

- Where do you get your datasets to train and test your models?

We have three-prong approach to solve data-problem. We have started with using the data shared by some Telcos (Ex: Orange).

Q &A

- Do you test your models with ‘Standard’ Dataset?

Unlike other domains, in Networking in general and NFV in particular, there aren’t any standard dataset. The current model is tested with the dataset used by many researchers, and we plan to (a) use the datasets that are well used by others (b) share the dataset with other researchers.

- Restrict the data from NFVI only or any usecases specific to services/workloads would be supported?

We don’t impose any restrictions – the availability of data and contributors’ interest defines everything. However, we expect Novelty and/or performance improvement in the model we publish.

- I want to contribute; do you have any open problems?

https://wiki.anuket.io/display/HOME/Call+for+Contributions+-+Potential+works+for+contributors

THANKS

sridharkn@u.nus.edu

Upcoming TAC Meetings

LF AI & DATA

Upcoming TAC Meetings

- March 24, 2022: Interpretable Deep Learning (tentative)
- April 7, 2022 – RosaeNLG Annual Review

Please note we are requesting special topics for future meetings.

If you have a topic idea or agenda item, please send agenda topic requests to tac-general@lists.lfaidata.foundation

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

LF AI & DATA