

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

August 11, 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)
- › (Possible) OSS update from Mandy Chessell (5 minutes)
- › OpenDataology from BAAI Sandbox Proposal (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

# 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	Jimin Jia	
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	

# Minutes approval



# Approval of July 28, 2022 Minutes

Draft minutes from the July 28 TAC call were previously distributed to the TAC members via the mailing list

## **Proposed Resolution:**

- › That the minutes of the July 28 meeting of the Technical Advisory Council of the LF AI & Data Foundation are hereby approved.

# OpenDataology

 OLF AI & DATA

# OpenDataology - An Open source dataset license compliance project

Proposal to Sandbox

Dr. Gopi Krishnan  
Rajbahadur



 [gopikrishnanrajbahadur@gmail.com](mailto:gopikrishnanrajbahadur@gmail.com)

 @gopirajbahadur

Clement Li



 [luburyana@gmail.com](mailto:luburyana@gmail.com)

 @Lazy\_LZ

Zev Qu



 [quzicheng315@gmail.com](mailto:quzicheng315@gmail.com)

 @qu\_zicheng

# 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.



# Outline



OpenDataology project overview



Sandbox requirements



Collaboration with existing LF and LF-AI Projects



Challenges



Road ahead

# Outline



OpenDataology project overview



Sandbox requirements



Collaboration with existing LF and LF-AI Projects

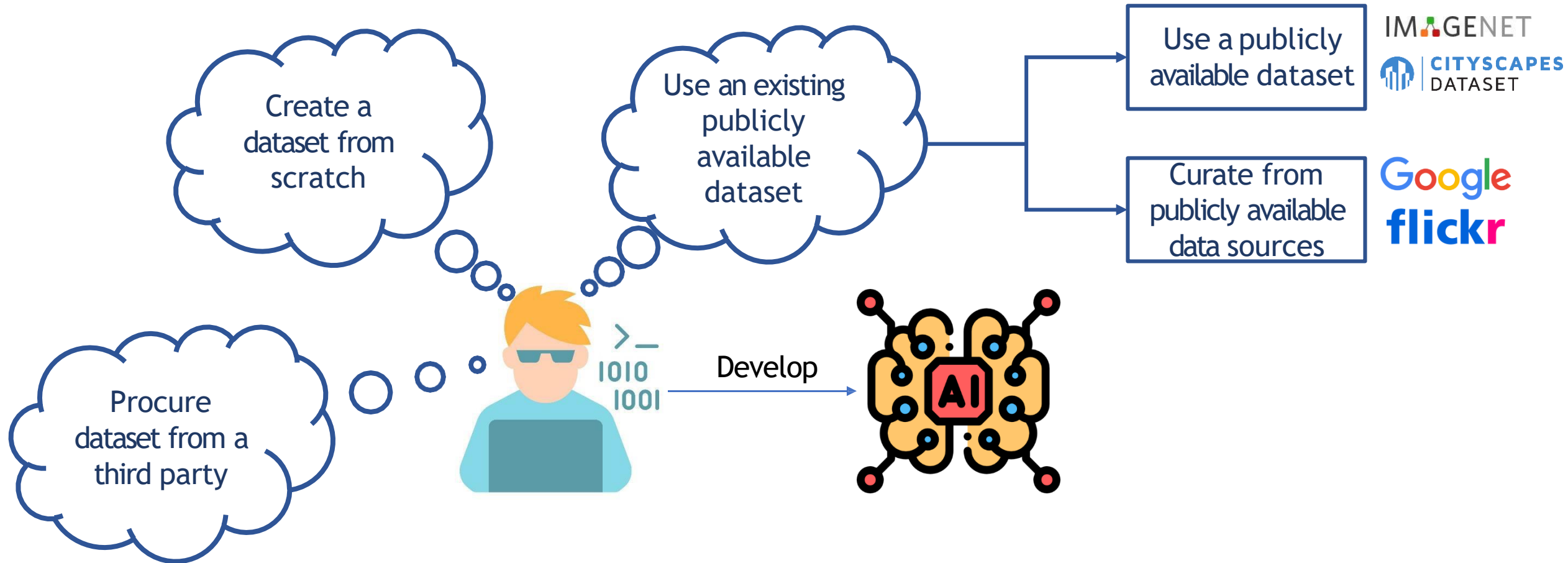


Challenges



Road ahead

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



# OpenDataology - An open source dataset license compliance project



The rights on the dataset that the users are entitled to

The actions that one must perform to enjoy those rights

- Cite the dataset
- Distribute the dataset (or the AI software) under the same license
- Do not use it for commercial purposes

OpenDataology assess the potential license compliance related risks associated with using a publicly available dataset to build commercial AI software. We do so using a license compliance analysis procedure that we propose and a crowd-sourced platform

	COPY & PUBLISH	ATTRIBUTION REQUIRED	COMMERCIAL USE	MODIFY & ADAPT	CHANGE LICENSE
PUBLIC DOMAIN	✓	✗	✓	✓	✓
CC BY	✓	✓	✗	✓	✓
CC BY-SA	✓	✓	✗	✓	✗
CC BY-ND	✓	✓	✗	✗	✗
CC BY-NC	✓	✓	✗	✓	✓
CC BY-NC-SA	✓	✓	✗	✓	✗
CC BY-NC-ND	✓	✓	✗	✗	✗

✓ You can redistribute (copy, publish, display, communicate, etc.)  
 ✓ You have to attribute the original work  
 ✓ You can use the work commercially  
 ✓ You can modify and adapt the original work  
 ✓ You can choose license type for your adaptations of the work.

IMAGENET

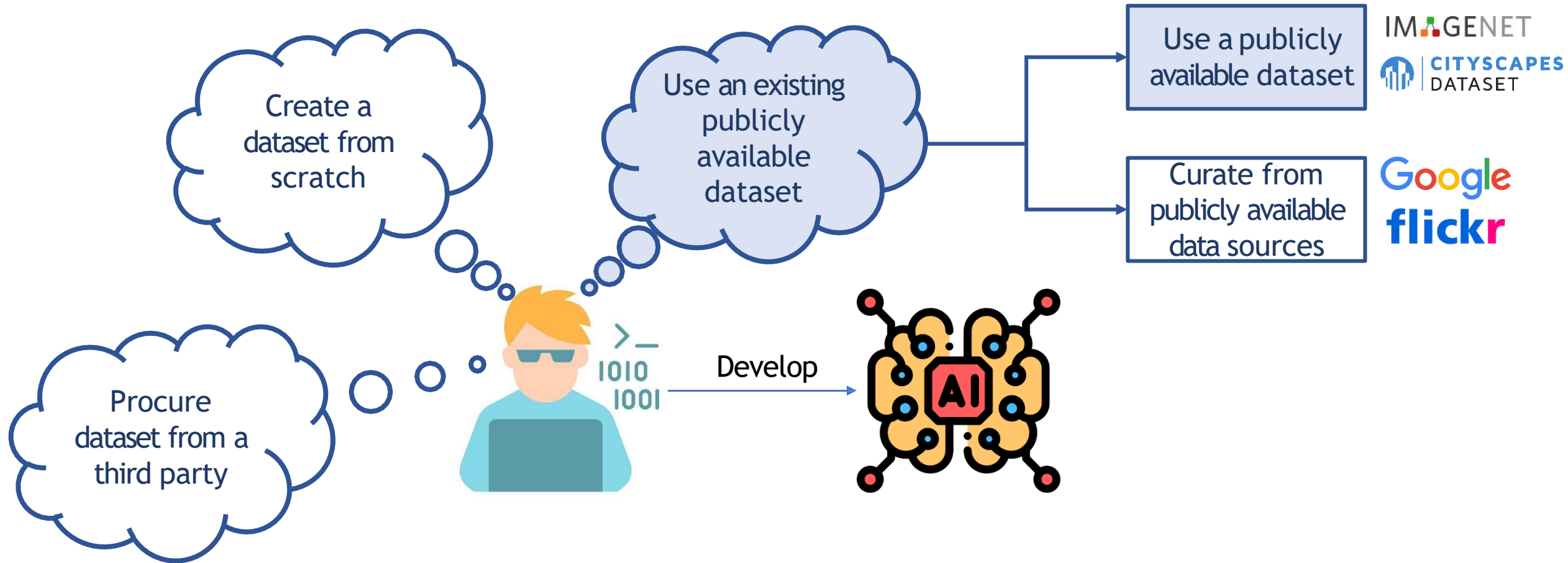
4. Researcher may provide research associates and colleagues with access to the Database provided that they first agree to be bound by these terms and conditions.

CITYSCAPES DATASET

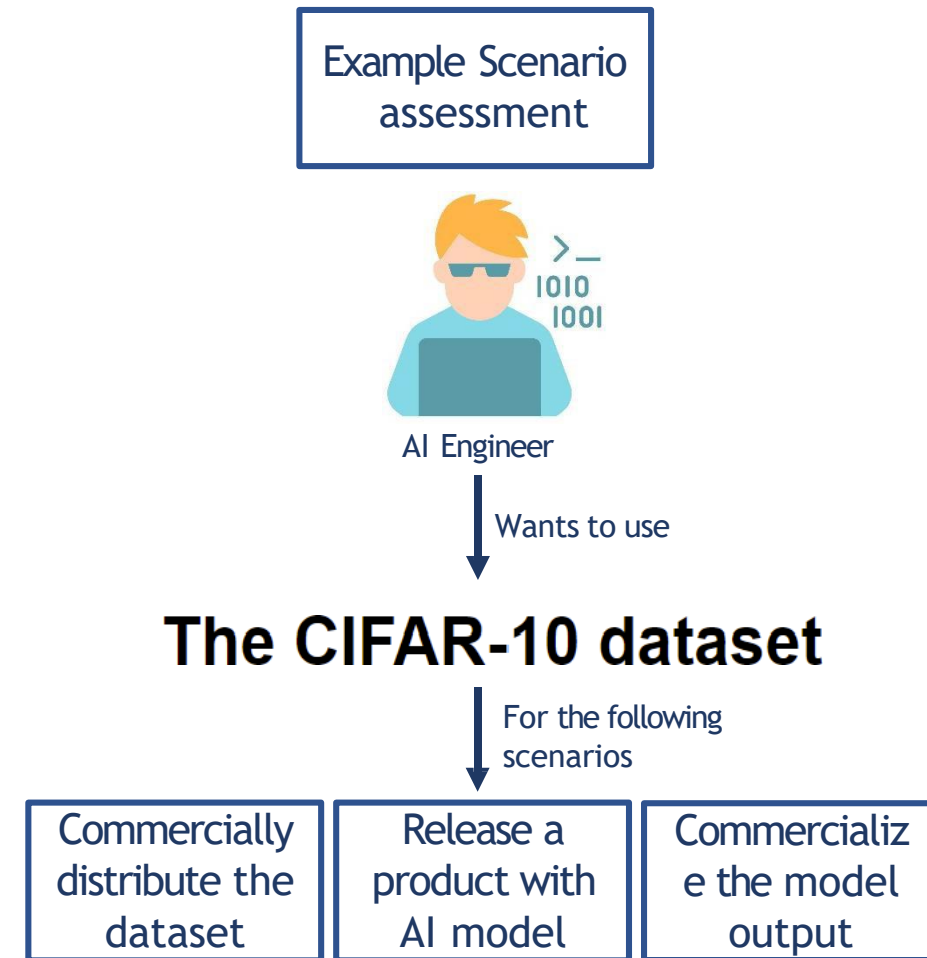
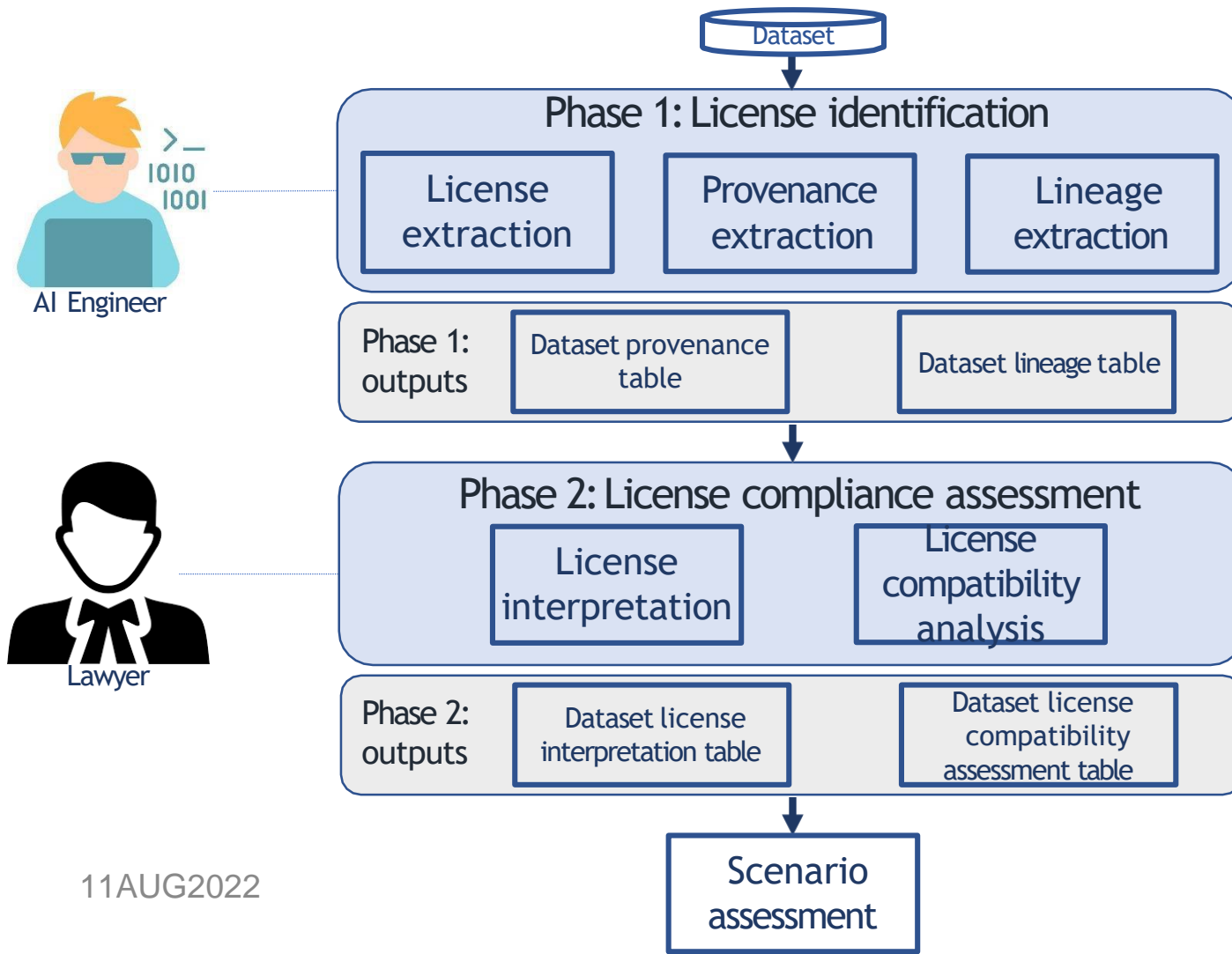
2. That you include a reference to the Cityscapes Dataset in any work that makes use of the dataset. For research papers, cite our preferred publication as listed on our [website](#); for other media cite our preferred publication as listed on our [website](#) or link to the [Cityscapes website](#).



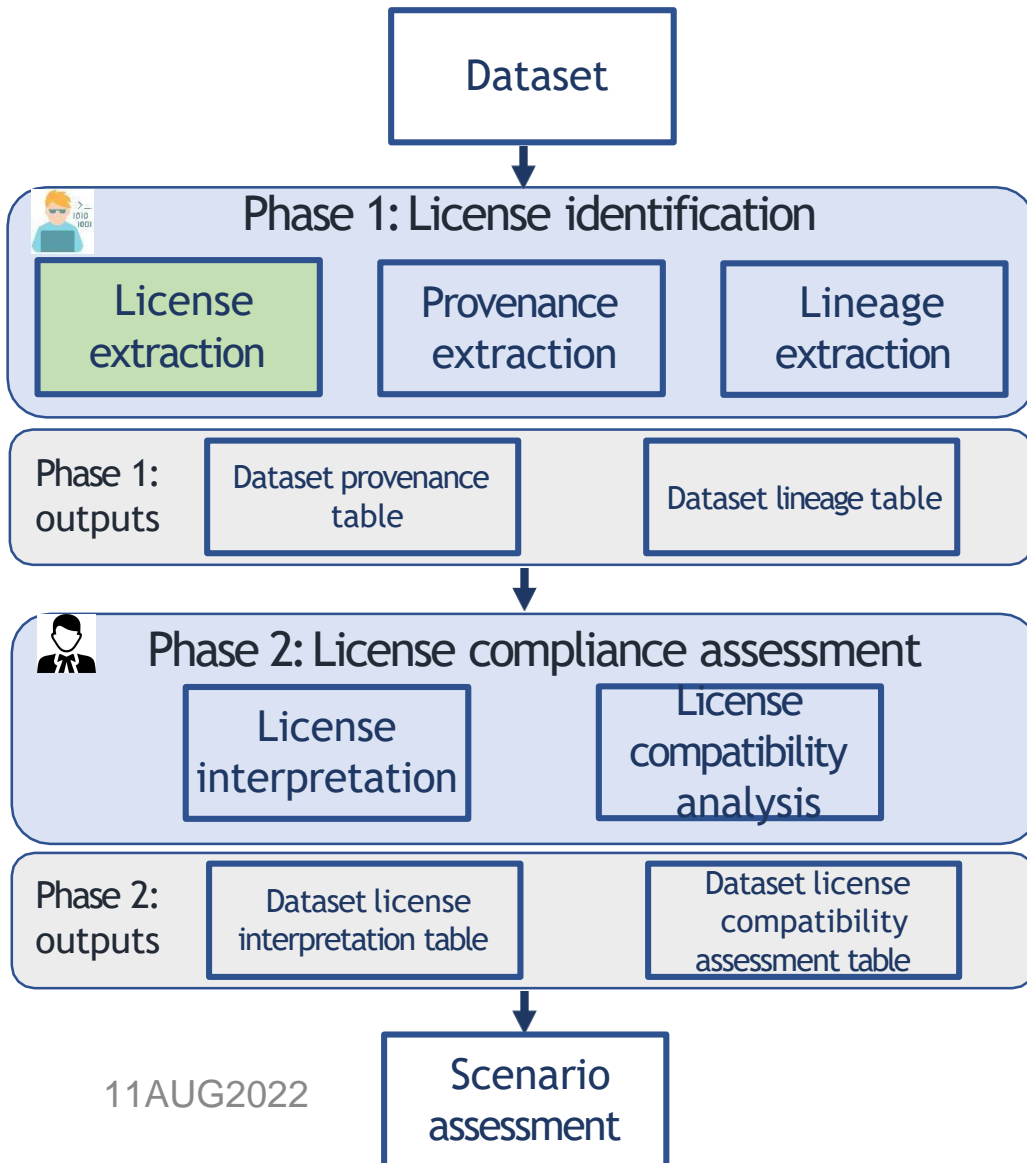
# There are several ways of acquiring the data required to build AI software



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



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

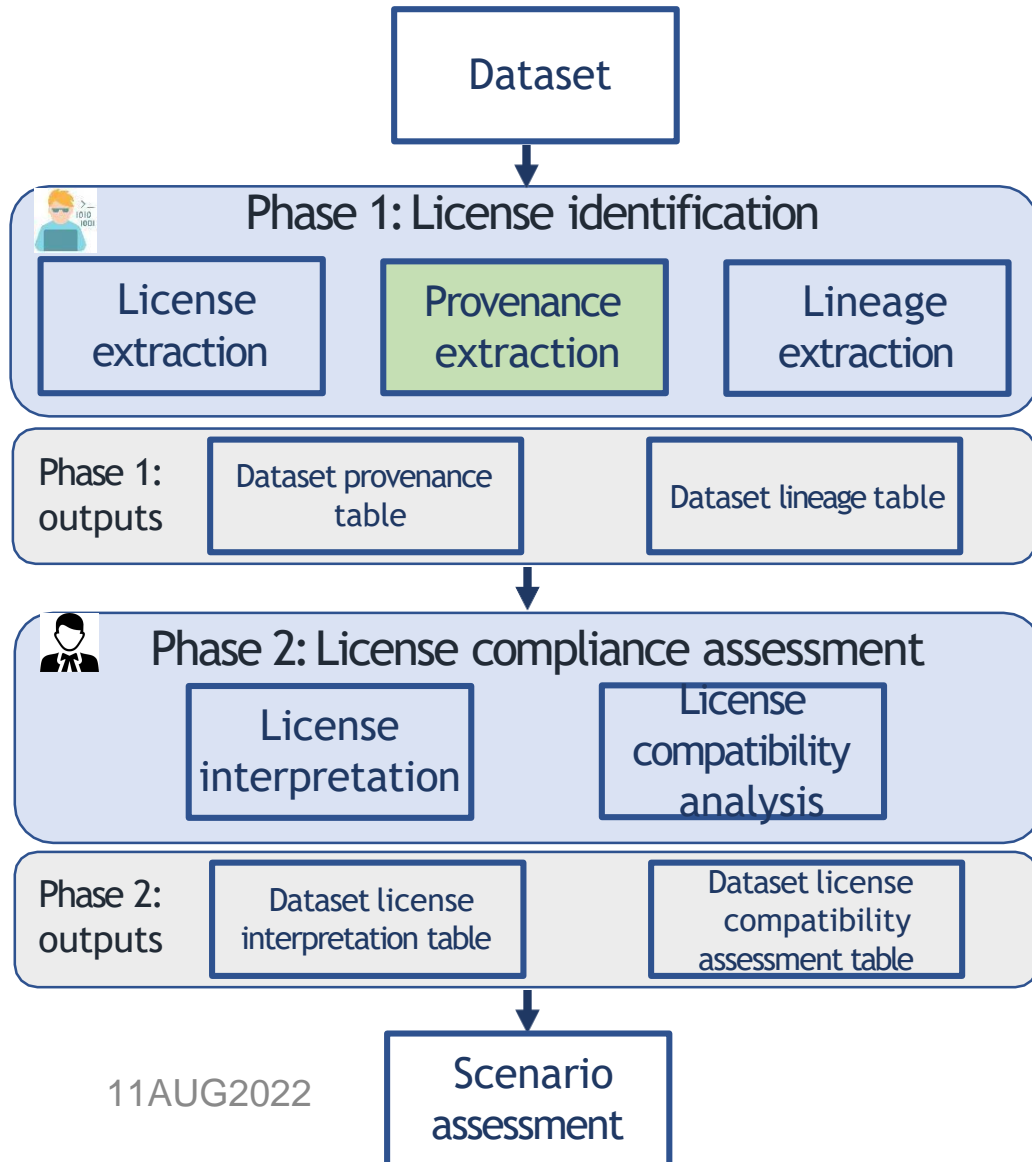


CIFAR-10 License (available on official website)

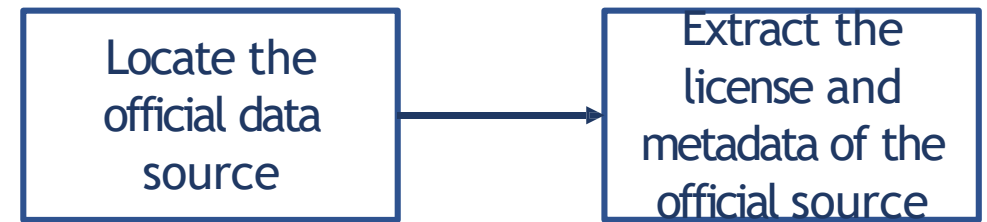
Please cite it if you intend to use this dataset.

- [Learning Multiple Layers of Features from Tiny Images](#), Alex Krizhevsky, 2009.

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

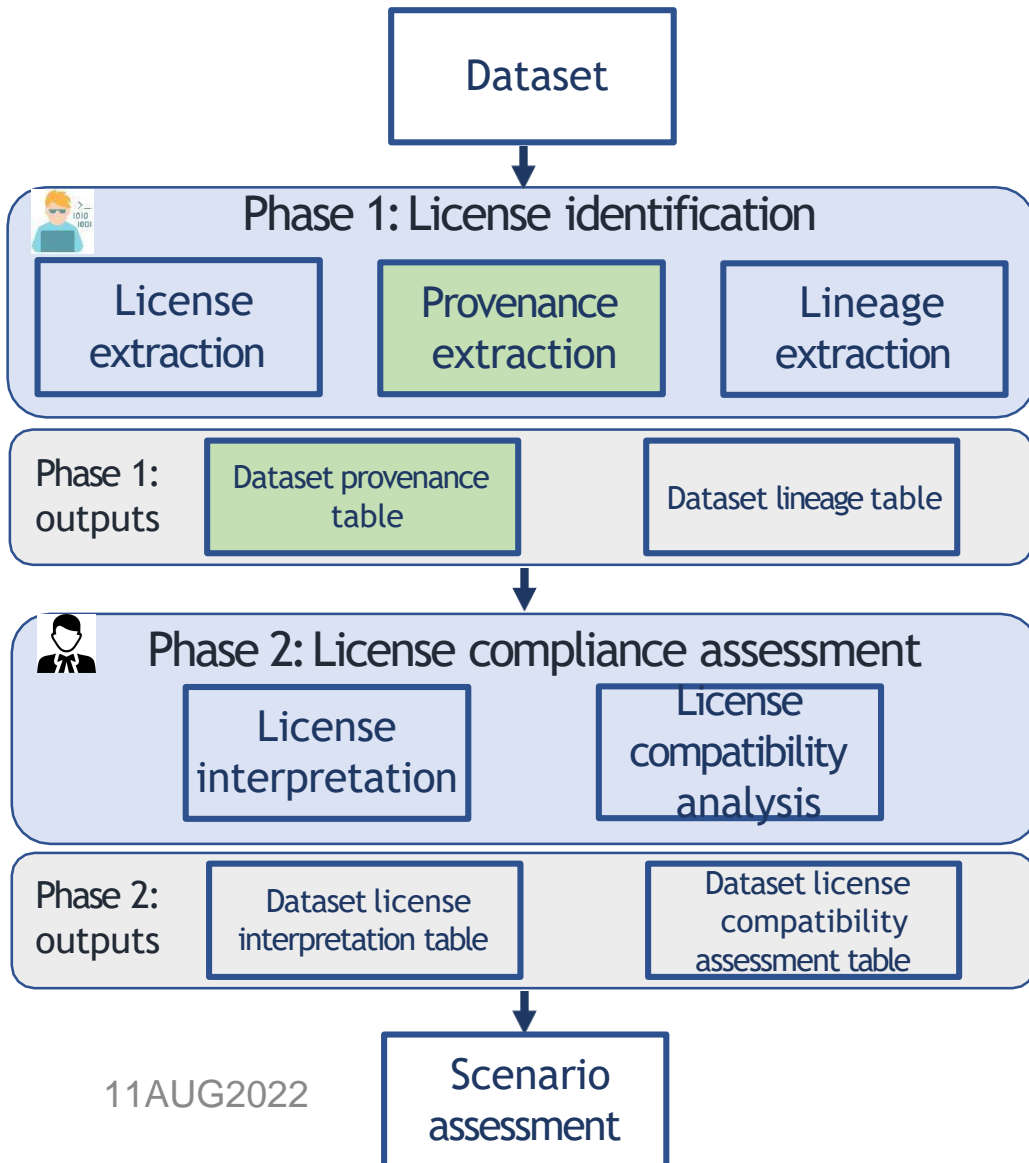


## Provenance extraction sub-steps



Provenance extraction step helps us mitigate **non-standard license location** and **unknown dataset origin problem**

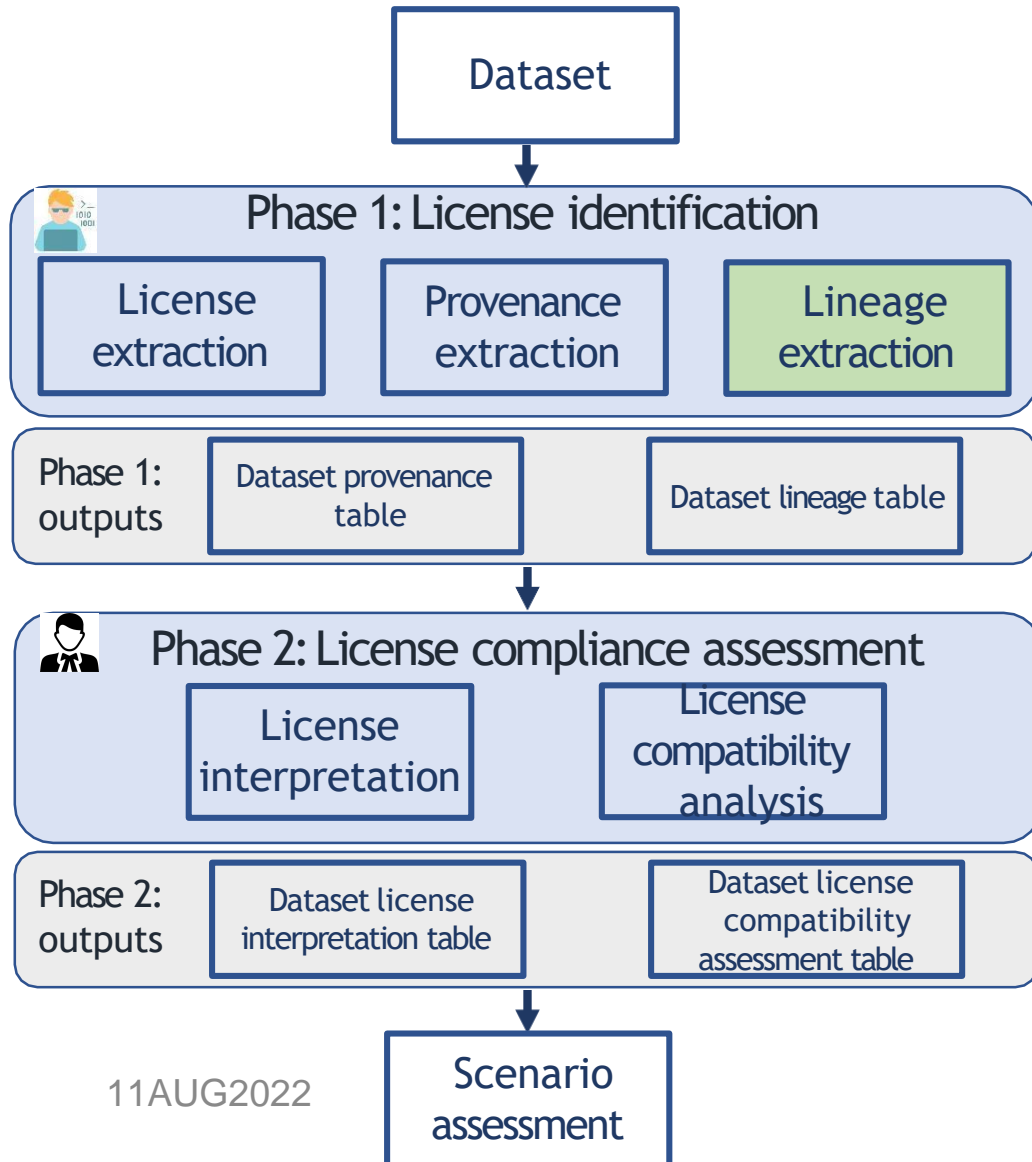
# Our approach to assess the potential risks of using publicly available datasets in commercial AI software



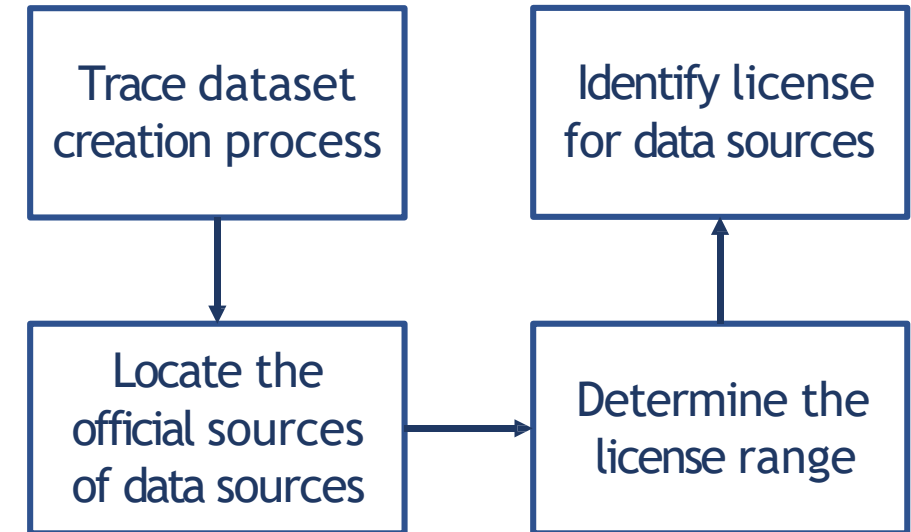
CIFAR-10's dataset provenance table

<b>Dataset-related details</b>	<b>Dataset name</b>	<b>Dataset version</b>	<b>Origin date</b>	<b>Origin</b>
	CIFAR-10	N/A	2009	<a href="https://www.cs.toronto.edu/~kriz/cifar.html">https://www.cs.toronto.edu/~kriz/cifar.html</a>
	<b>Description of dataset</b>		<b>Description of data collection process</b>	
	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.	
<b>Downloaded outlet</b>	<b>Is outlet licensed?</b>	<b>Is dataset publicly available?</b>	<b>Additional notes</b>	
N/A	N/A	Yes	This dataset is a subset of another dataset called 80 Million Tiny Images	
<b>License-related details</b>	<b>Where license was found</b>		<b>License location</b>	<b>License content</b>
	Present on the official dataset website		<a href="https://www.cs.toronto.edu/~kriz/cifar.html">https://www.cs.toronto.edu/~kriz/cifar.html</a>	(not pasting content due to space)
<b>Metadata</b>	<b>Hashcode</b>		<b>Size</b>	<b>Format</b>
	MD5: c58f30108f718f92721af3b95e74349a (Python version)		163MB (Python version)	tar.gz

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

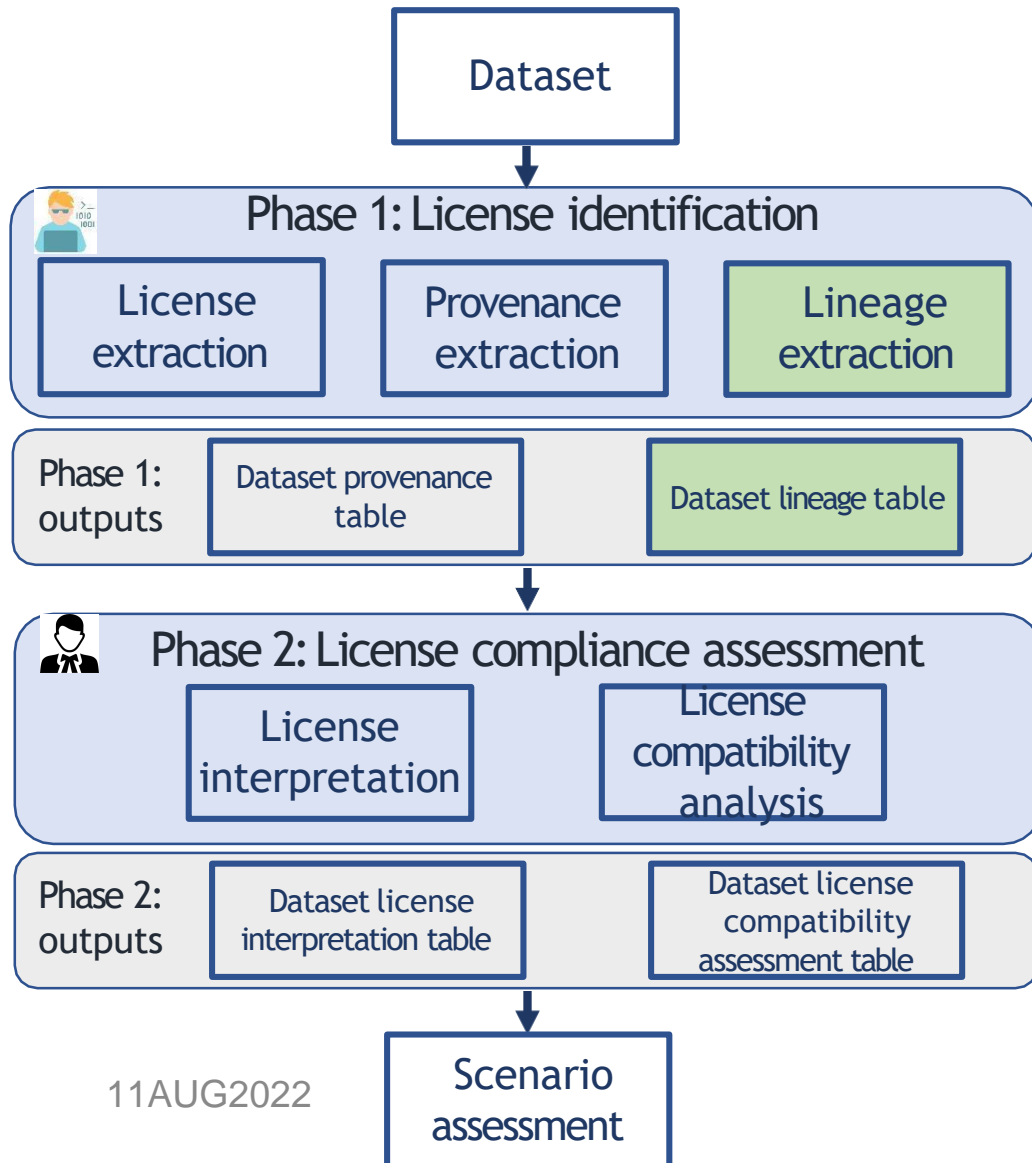


## Lineage extraction sub-steps

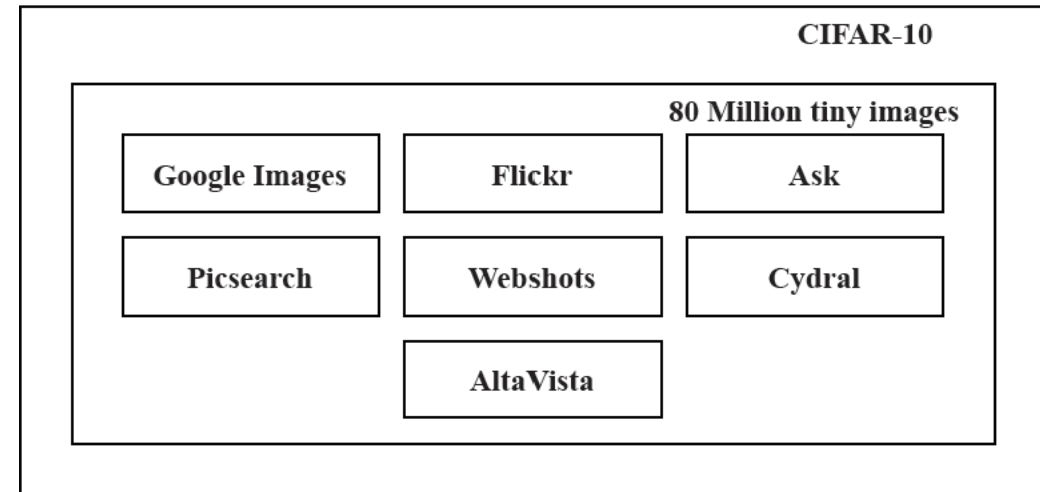


Lineage extraction step helps us mitigate **unknown-data problem**

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

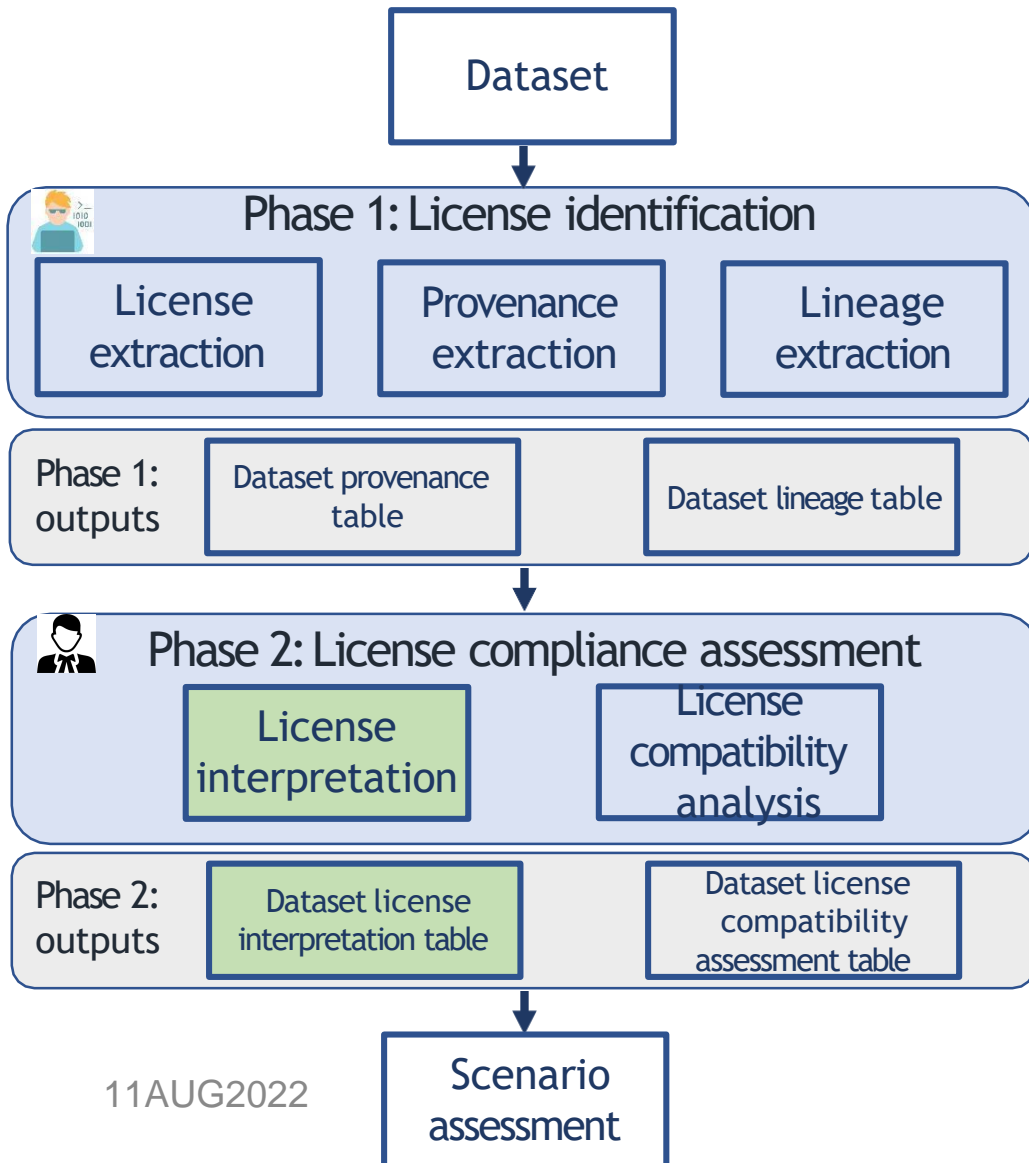


CIFAR-10's dataset lineage table



Provenance details are recorded for each of the data source

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

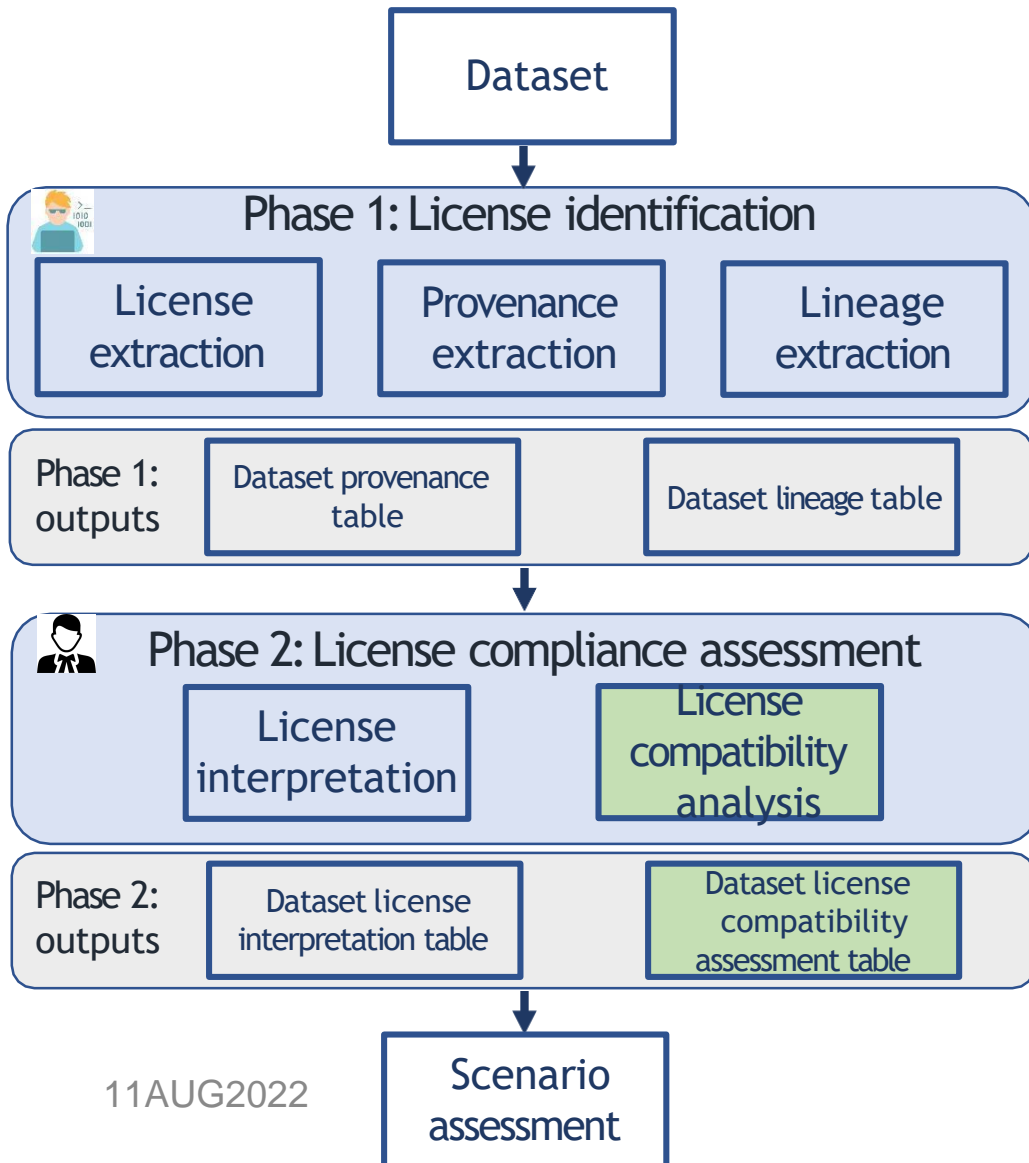


CIFAR-10's dataset license interpretation table  
(Based on enhanced Montreal Data License)

License metadata	Licensor		License name		Dataset name		Dataset version	
		Alex Krizhevsky		Custom license		CIFAR-10		N/A
License metadata	<b>Credit/Attribution Notice</b>							
	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	
<b>Rights</b>	✓		✓		✓		✓	
Obligations	Cite paper		Cite paper		Cite paper		Cite paper	
Data rights in conjunction with model	Benchmark	Re-search	Publish	Internal Use	Commercialization		Model Reverse Engineer	
					Output	Model		
<b>Rights</b>	✓	✓	✓	✓	✓	✓	✓	
Obligations	Cite paper	Cite paper	Cite paper	Cite paper	Cite paper	Cite paper	Cite paper <b>24</b>	



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



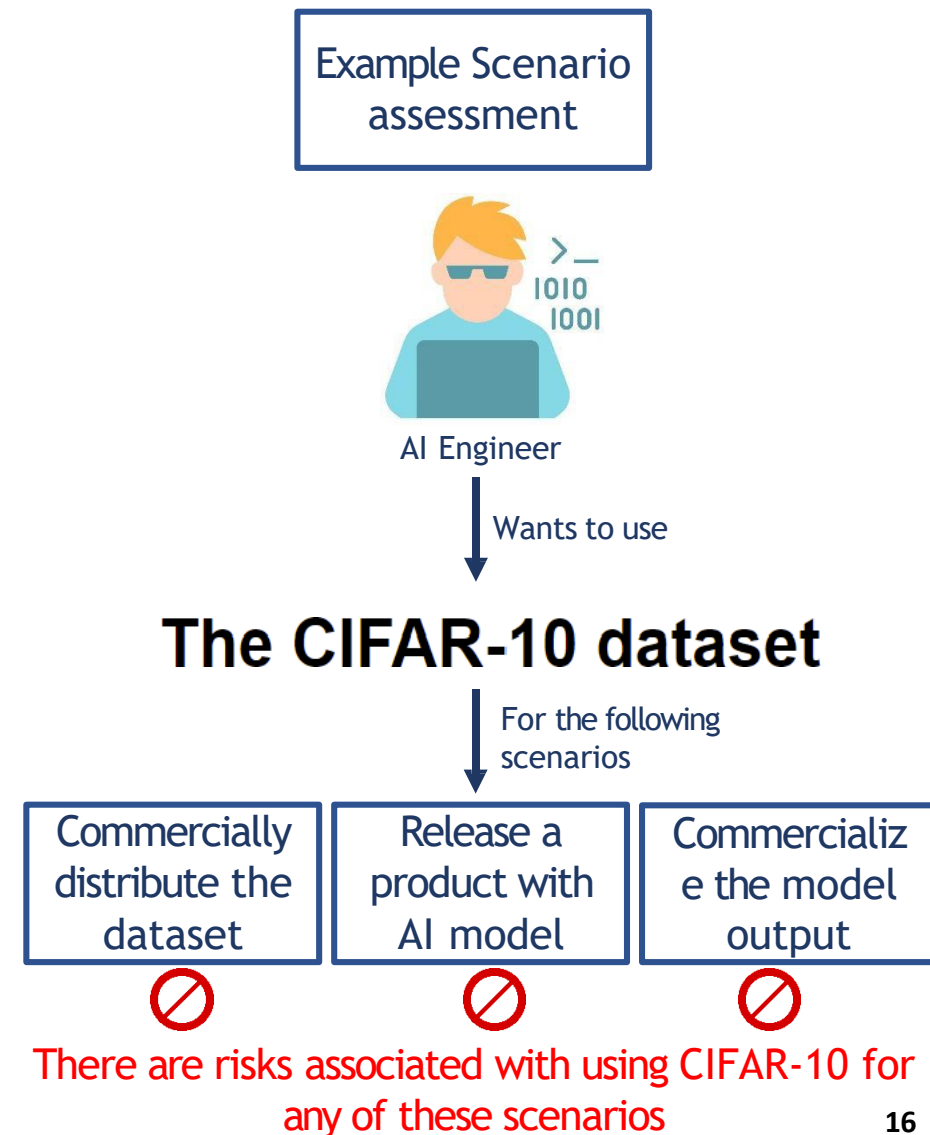
CIFAR-10's dataset license compatibility table  
(Based on analyzing the license of all data sources)

	Licensor		License name	Dataset name	Dataset version		
License metadata	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	✓		✓ (X)	✓ (X)	✓ (X)		
Obligations	Cite paper		Cite paper	Cite paper	Cite paper		
Data rights in conjunction with model	Bench- mark	Re- search	Publish	In- ternal Use	Commercialization		Model Reverse Engineer
					Out- put	Model	
Rights	✓	✓	✓	✓	✓ (X)	✓ (X)	✓
Obligations	Cite paper	Cite paper	Cite paper	Cite paper	Cite paper	Cite paper	Cite paper

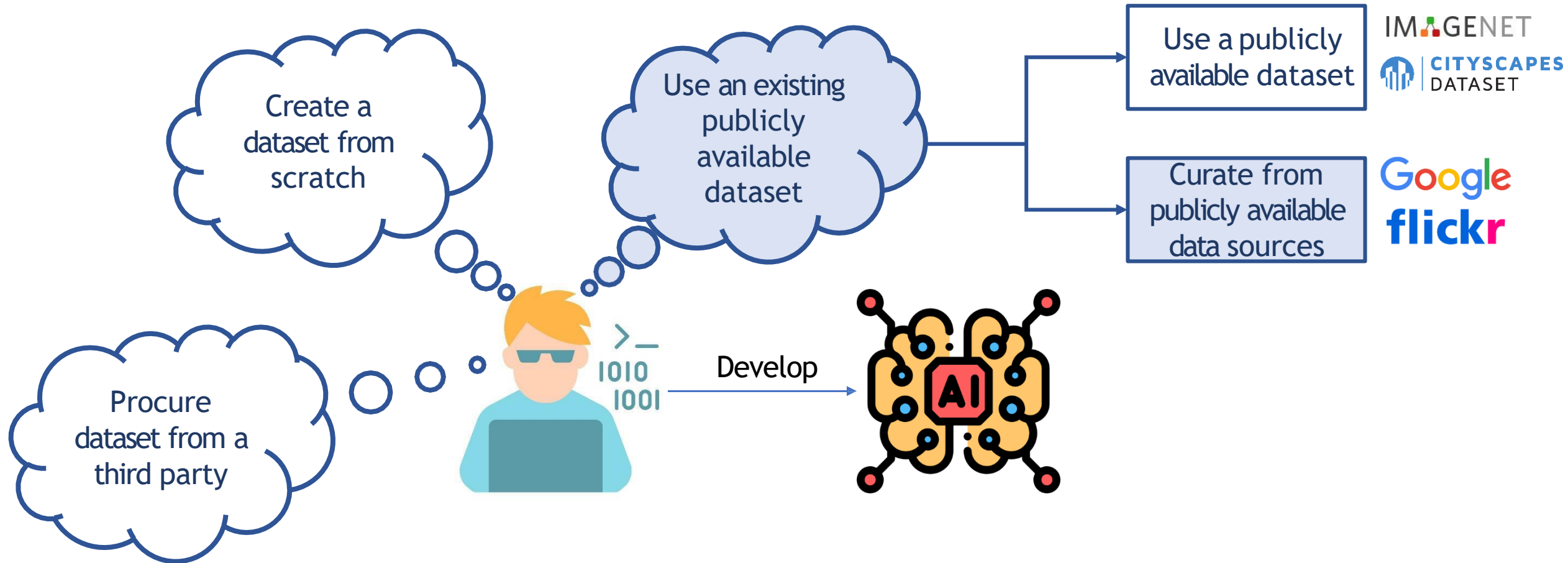
# Our approach to assess the potential risks of using publicly available datasets in commercial AI software

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	✓		✓ (X)		✓ (X)		✓ (X)	
Obligations	Cite paper		Cite paper		Cite paper		Cite paper	
Data rights in conjunction with model	Benchmark	Re-search	Publish	Internal Use	Commercialization		Model Reverse Engineer	
					Output	Model		
Rights	✓	✓	✓	✓	✓ (X)	✓ (X)	✓	
Obligations	Cite paper	Cite paper	Cite paper	Cite paper	Cite paper	Cite paper	Cite paper	

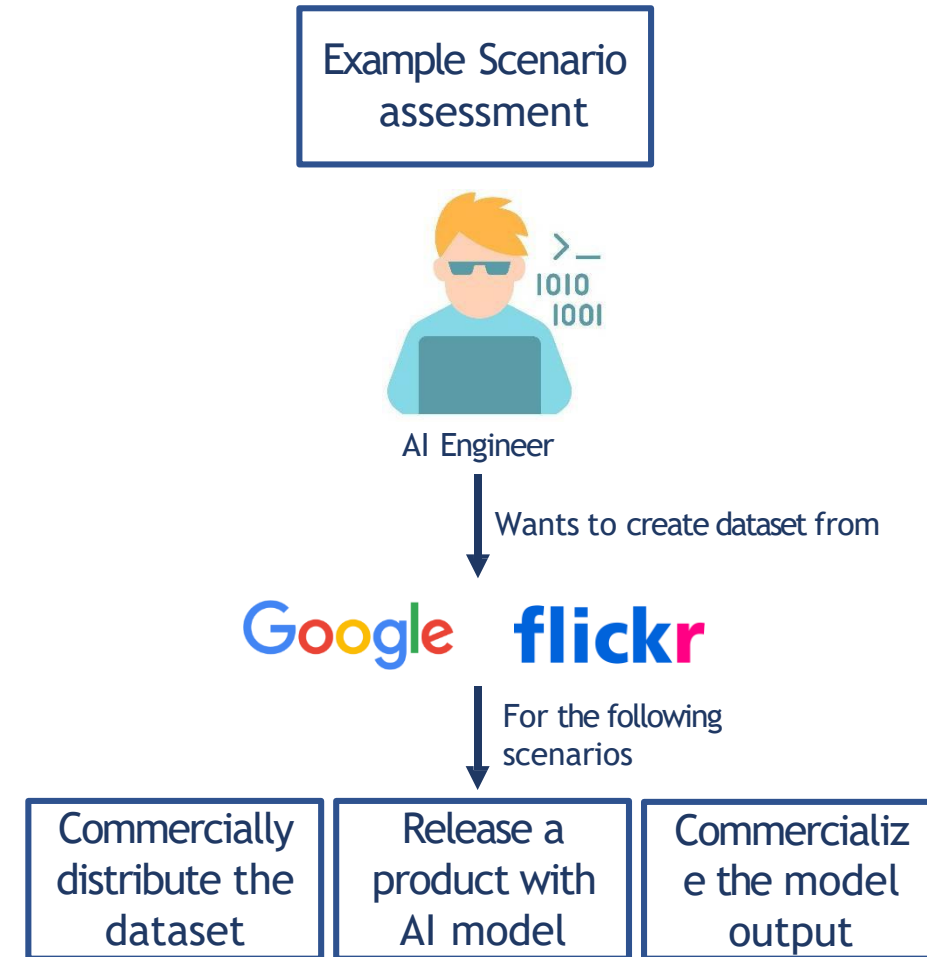
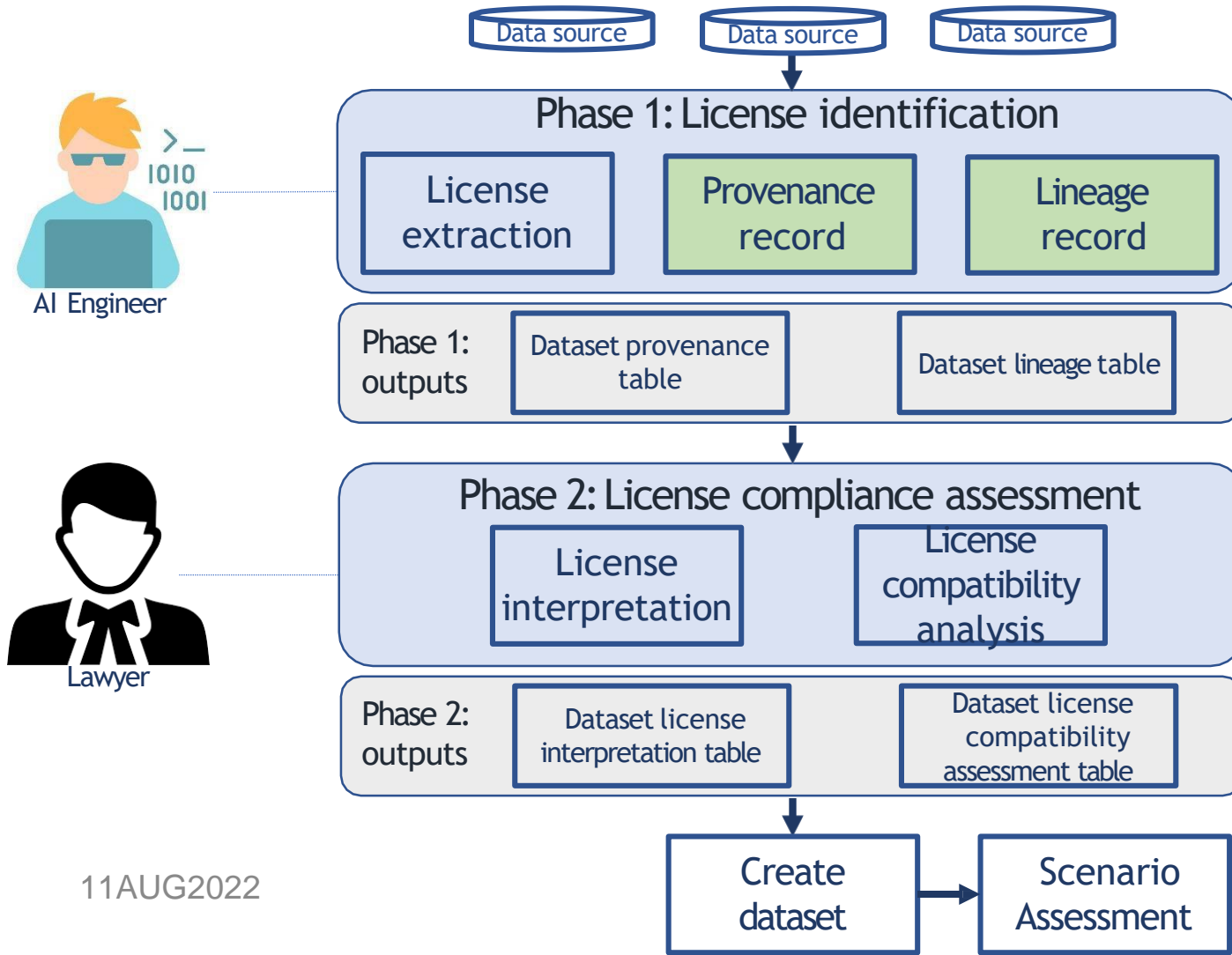
11AUG2022



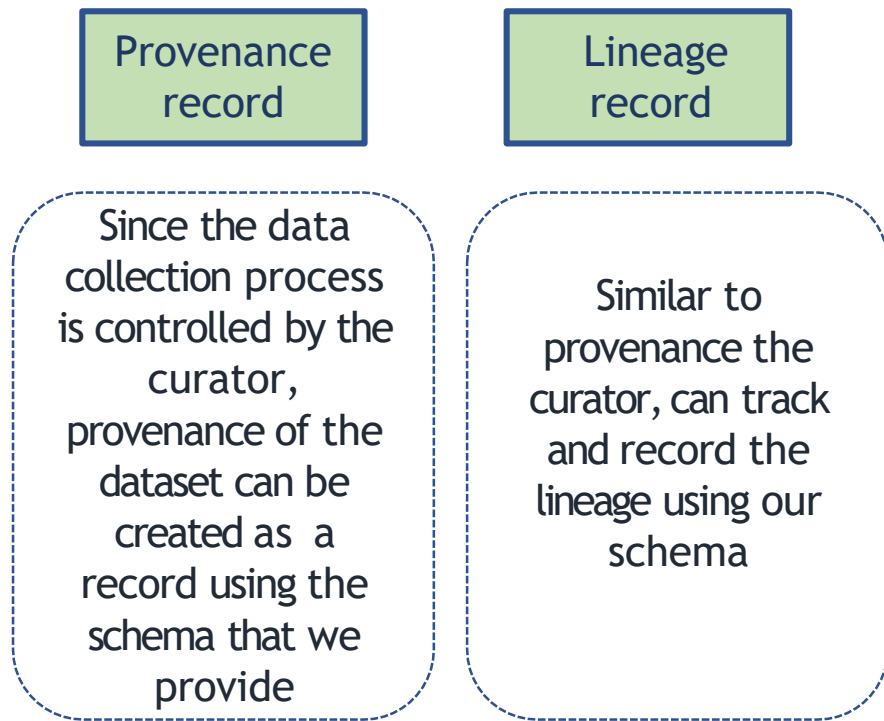
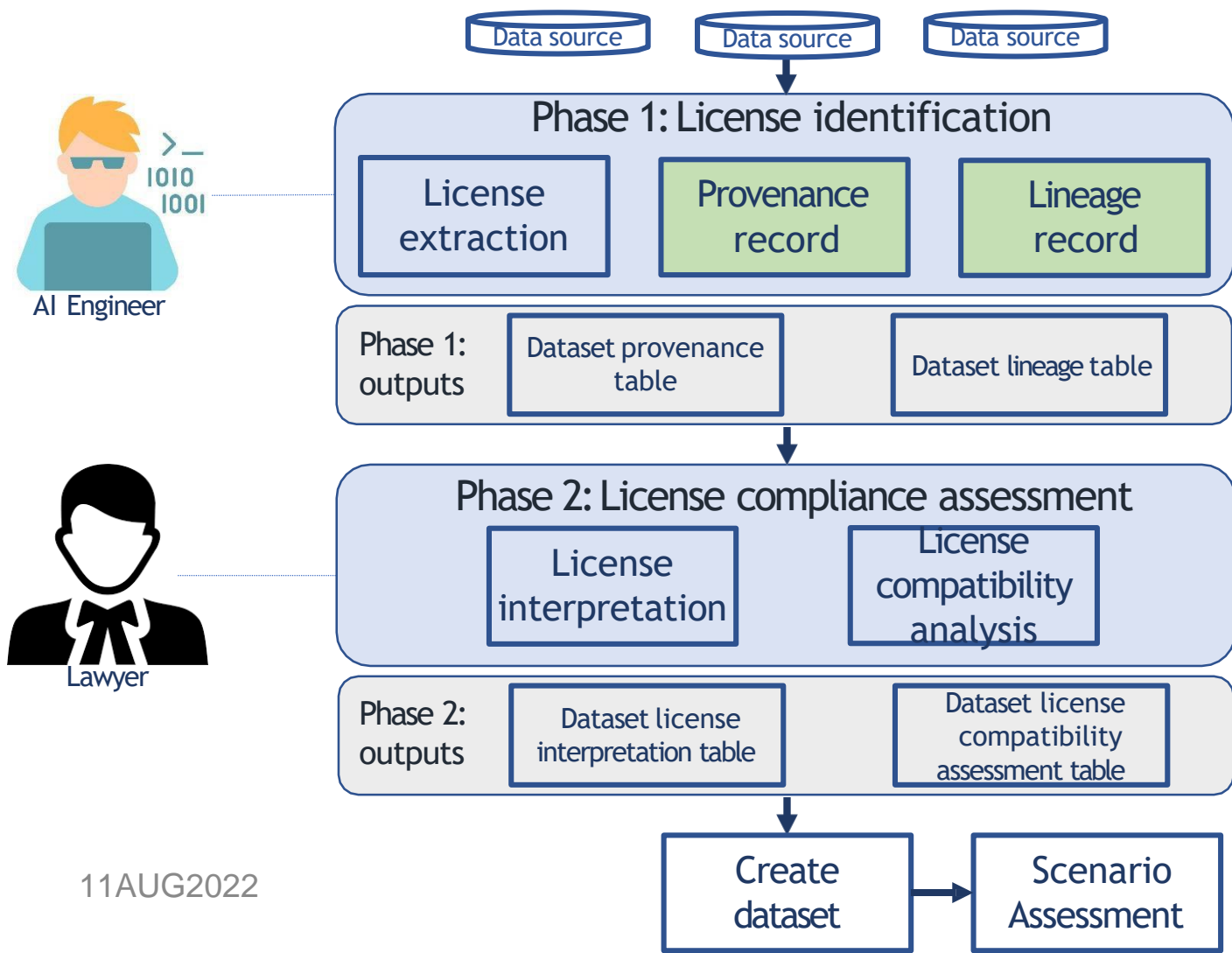
# There are several ways of acquiring the data required to build AI software



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

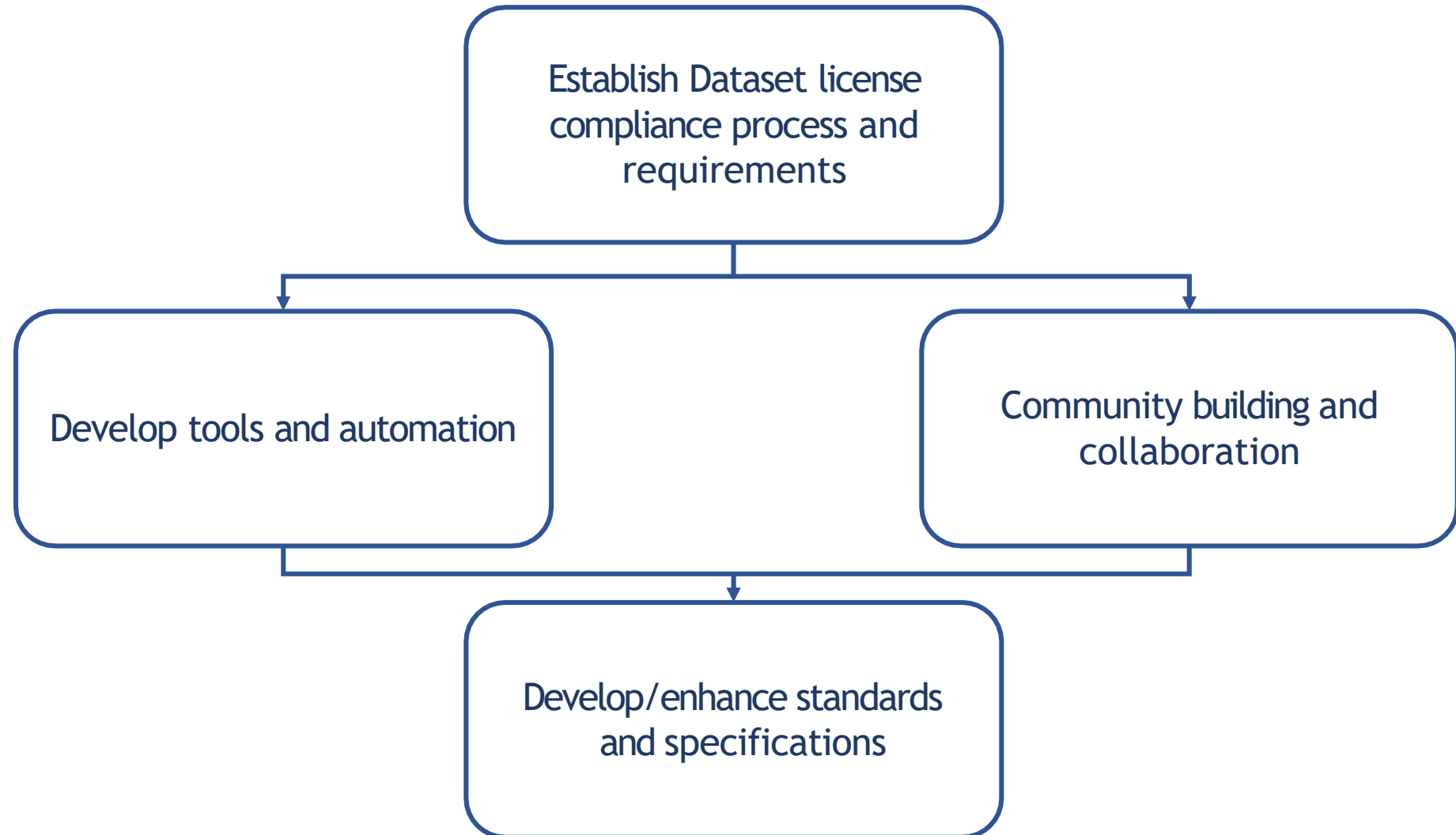


# Our approach to assess the potential risks of using datasets created from publicly available data sources

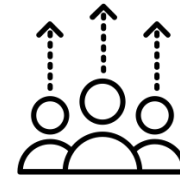
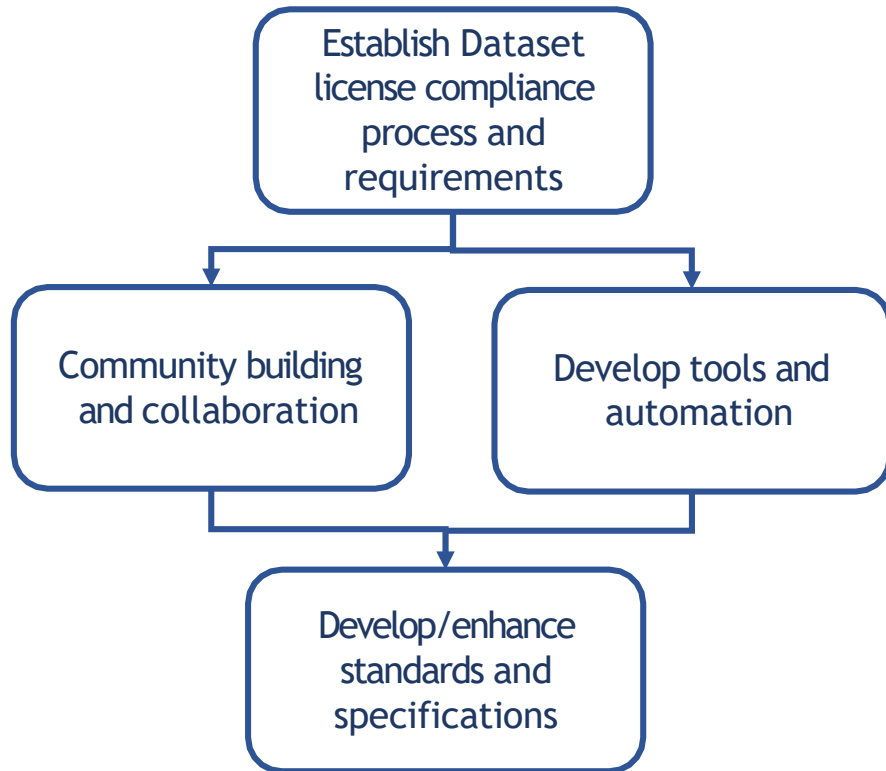


**When curating datasets, unless another (pre-curated) dataset is involved, no explicit provenance or lineage extraction is required**

# OpenDataology- Areas of interest



# OpenDataology - Current progress

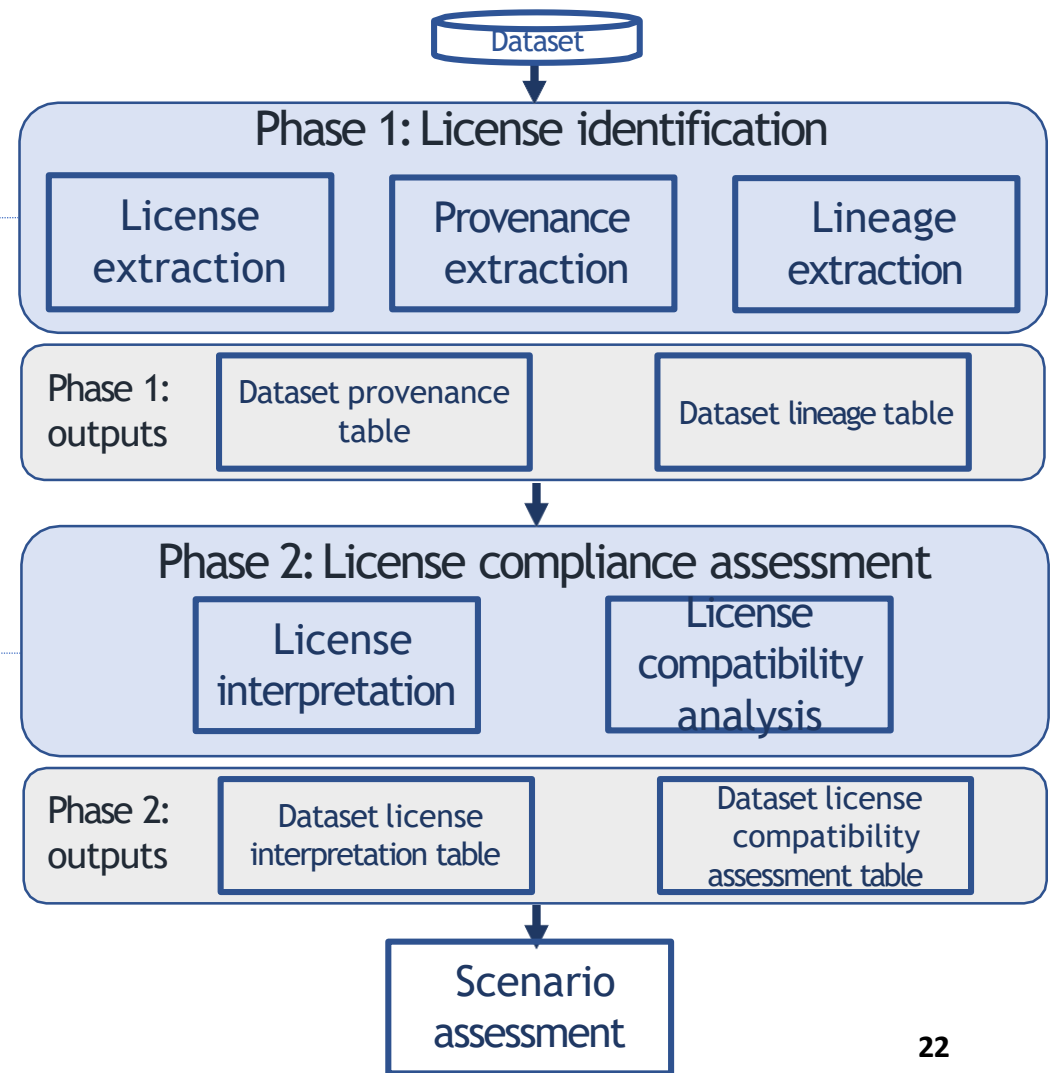
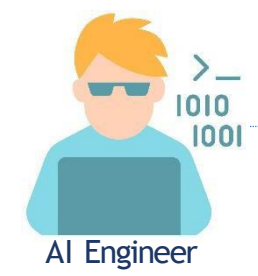
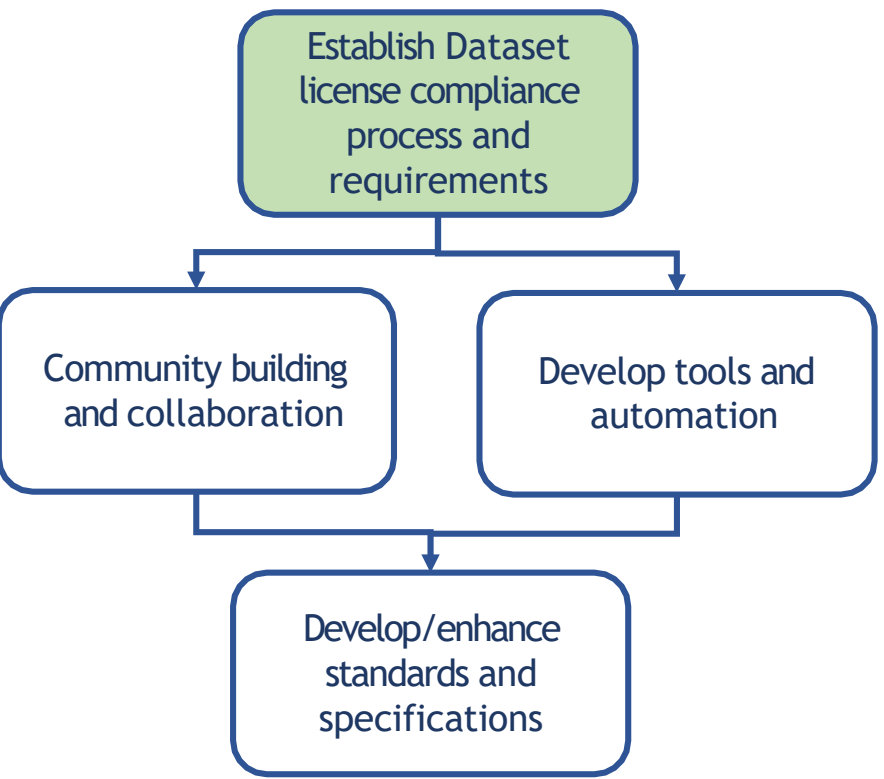


## Current core contributors

(In alphabetic order)

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

# OpenDataology - Current progress





# OpenDataology - Current progress



Recap



License compliance  
process for curated datasets



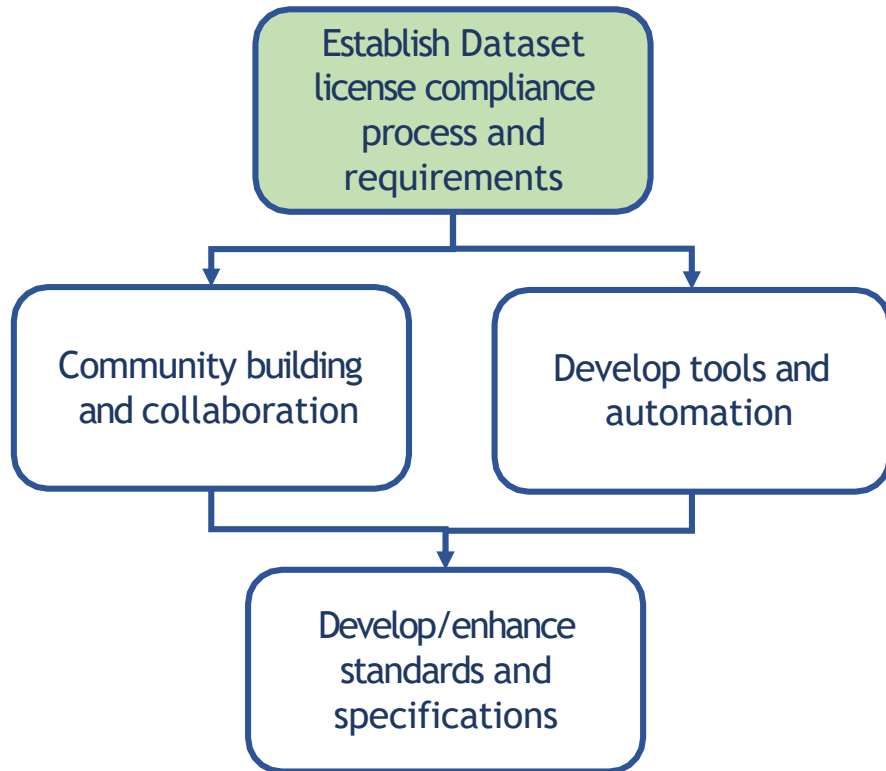
Challenges



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>

# OpenDataology - Current progress



Recap



License compliance process for curated datasets



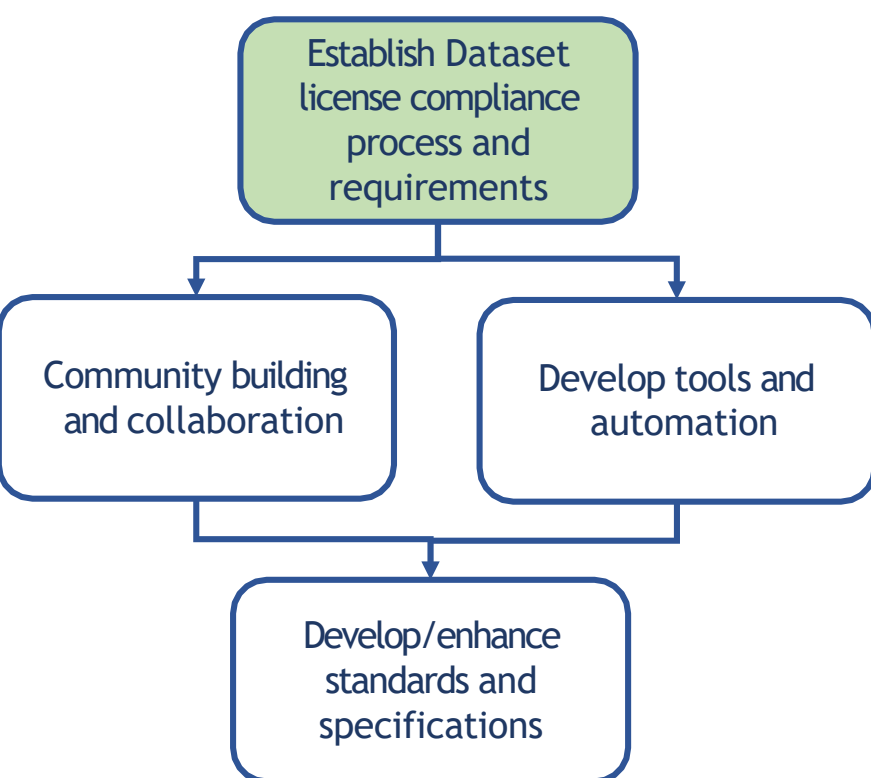
Challenges



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)	✓	⊘	⊘

# OpenDataology - Current progress



Recap



License compliance process for curated datasets



Challenges



Current progress



Road ahead

Establish Dataset license compliance process and requirements

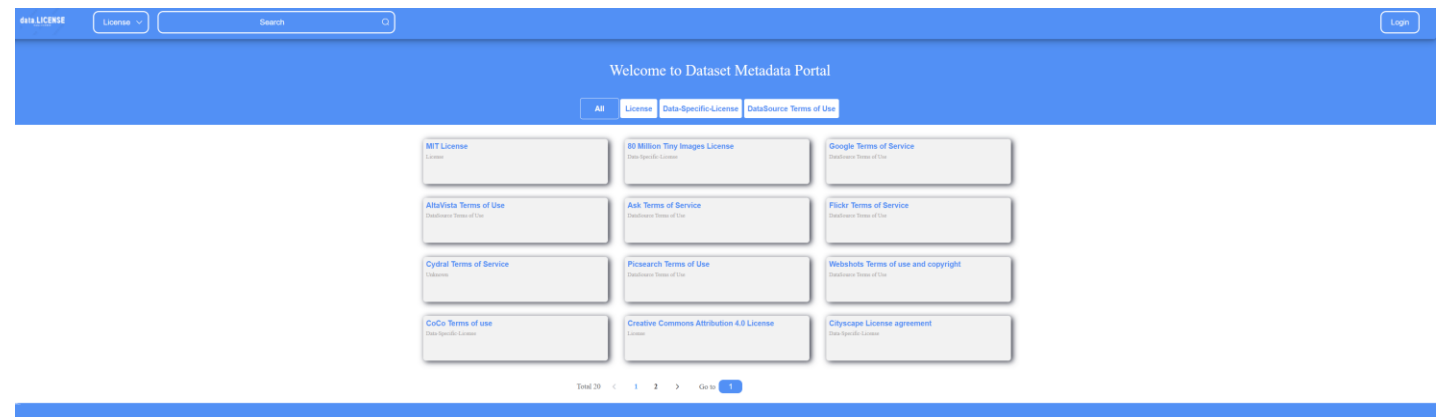
Community building and collaboration

Develop tools and automation

Develop/enhance standards and specifications

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://www.opendataology.com:30800/#/dataSetInfo?id=1>



# OpenDataology - Current progress



Recap



License compliance process for curated datasets



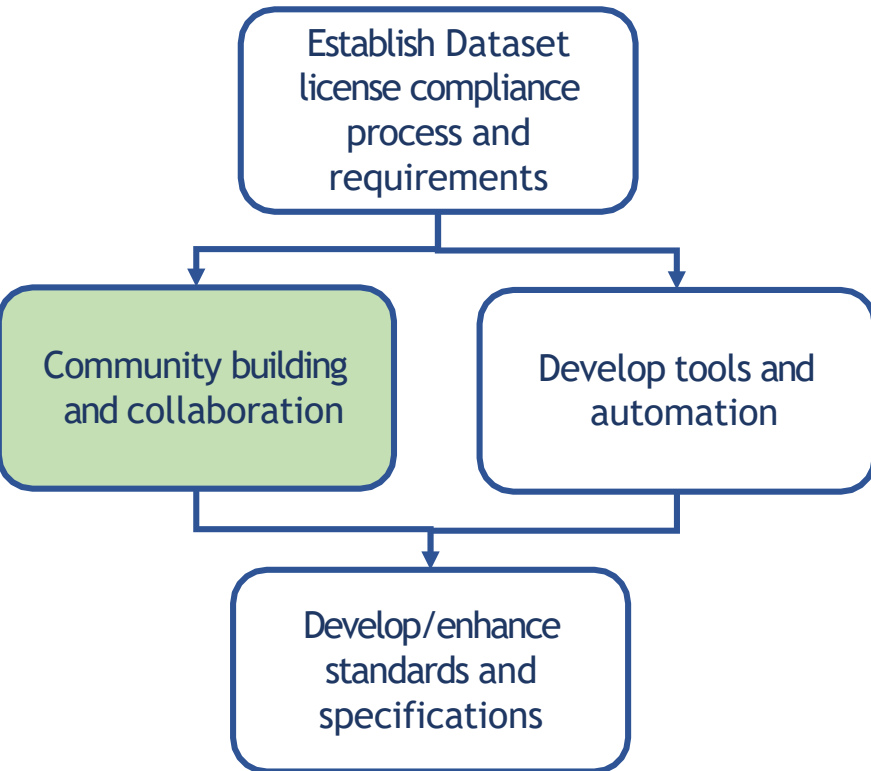
Challenges



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://www.opendataology.com:30800/#/dataSetInfo?id=1>

The screenshot shows a web interface for 'data LICENSE'. It features a search bar with 'License' selected and a search icon. Below the search bar, the results for 'MIT License' are displayed, including fields for Name, type, approved, and identifier. The interface is divided into sections: 'Data', 'Model', and 'Others'. Under 'Data', there is a 'Can' category with a list of items: ModelBenchmark, ModelResearch, ModelPublish, ModelInternal, ModelOutputCom, ModelCom, and ModelRev. Under 'Model', there is a 'Cannot' category (red button) and an 'Obligation' category (blue button) with a list of items: ModelBenchmark, ModelResearch, ModelPublish, ModelInternal, ModelOutputCom, ModelCom, and ModelRev. Under 'Others', there is a 'Limitation' category (purple button) with a list of items: ModelInternal.

# OpenDataology - Current progress



Recap



License compliance process for curated datasets



Challenges



Current progress



Road ahead

Establish Dataset license compliance process and requirements

Community building and collaboration

Develop tools and automation

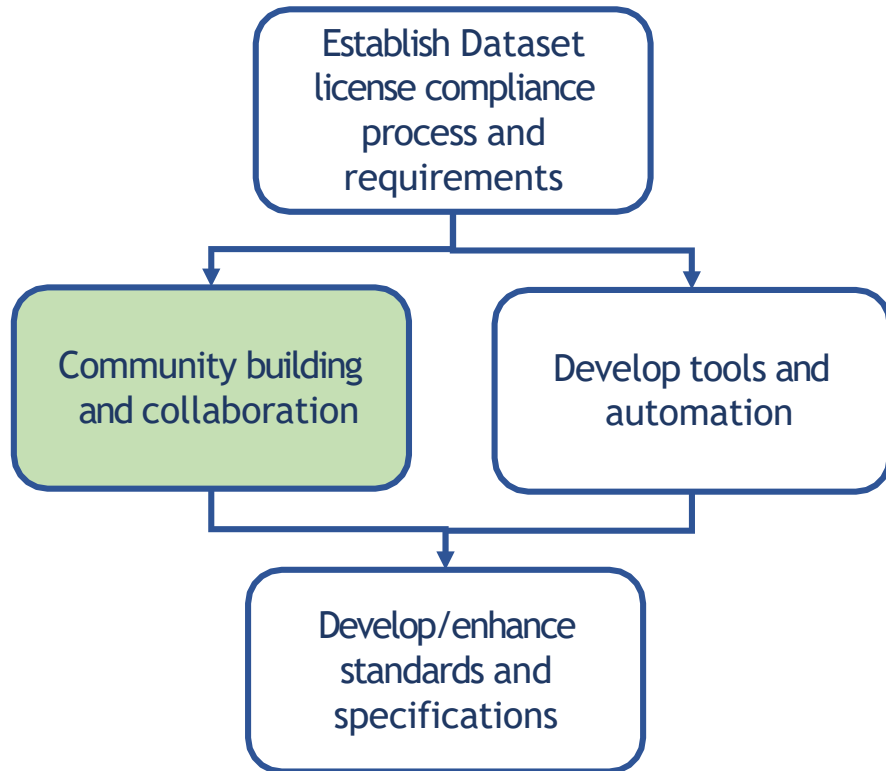
Develop/enhance standards and specifications

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://www.opendataology.com:30800/#/dataSetInfo?id=1>

MetaData					
Name	CIFAR-10	Version	N/A	License ID	1
License Name	<a href="#">MIT License</a>	Licensor	Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton	License From	Present on the official dataset website
License Location	<a href="https://www.cs.toronto.edu/~kriz/cifar.html">https://www.cs.toronto.edu/~kriz/cifar.html</a>	Origin	<a href="https://www.cs.toronto.edu/~kriz/cifar.html">https://www.cs.toronto.edu/~kriz/cifar.html</a>	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."				

# OpenDataology - Current progress



  
**GitHub**  
Open for collaboration  
and contributions

# OpenDataology - Current progress



Recap



License compliance process for curated datasets



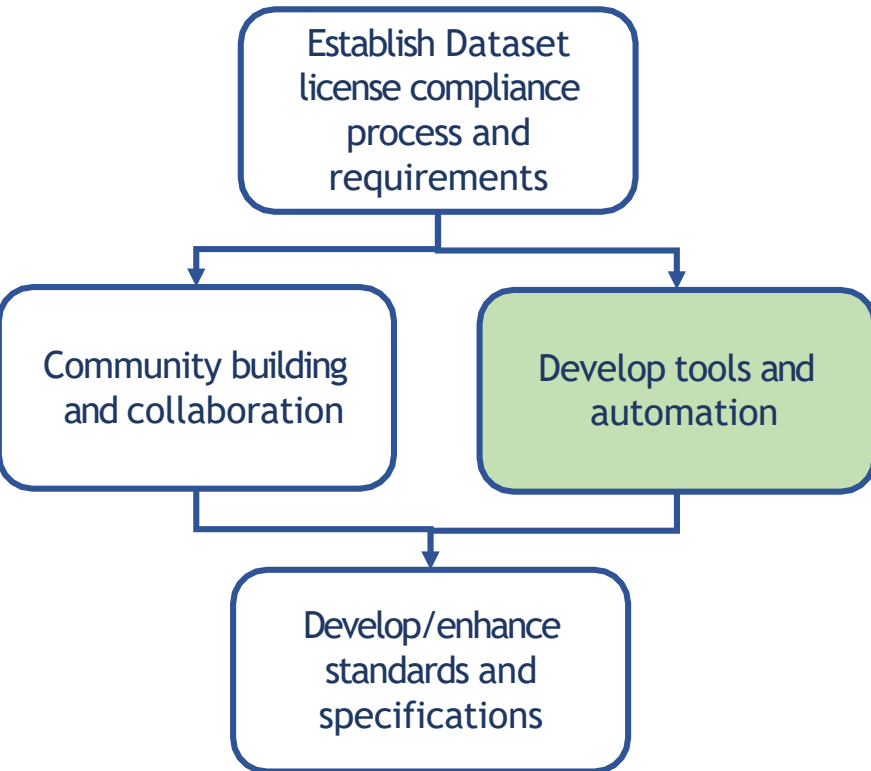
Challenges



Current progress



Road ahead



We developed an automatic license generator tool that helps the creators of datasets to assign license based on the rights and obligations that they wish to impose on the dataset

License-Generator

License-Generator

An easy-to-use license assistant.

Answer all the questions in the 3 steps to get your custom data license.

1 ————— 2

For Data itself For Model

Question : 1.1 Can users themselves make the Data available to other third parties?

Description : Answering "Yes" gives the right to distribute the data, i.e. to make all or part of the Data available to Third Parties under the same terms as those you will choose.

Yes No

Number of all Licenses: 2

<p>MIT License</p> <p>Select</p> <p>data_access_rights: Yes</p> <p>data_modification_rights: Yes</p> <p>data_network_rights: Yes</p> <p>data_represent_rights: Yes</p>	<p>Creative Commons Attribution 4.0 License</p> <p>Select</p> <p>data_access_rights: Yes</p> <p>data_modification_rights: Yes</p> <p>data_network_rights: N/A</p> <p>data_represent_rights: Yes</p>
--	---

# OpenDataology - Current progress



Recap



License compliance process for curated datasets



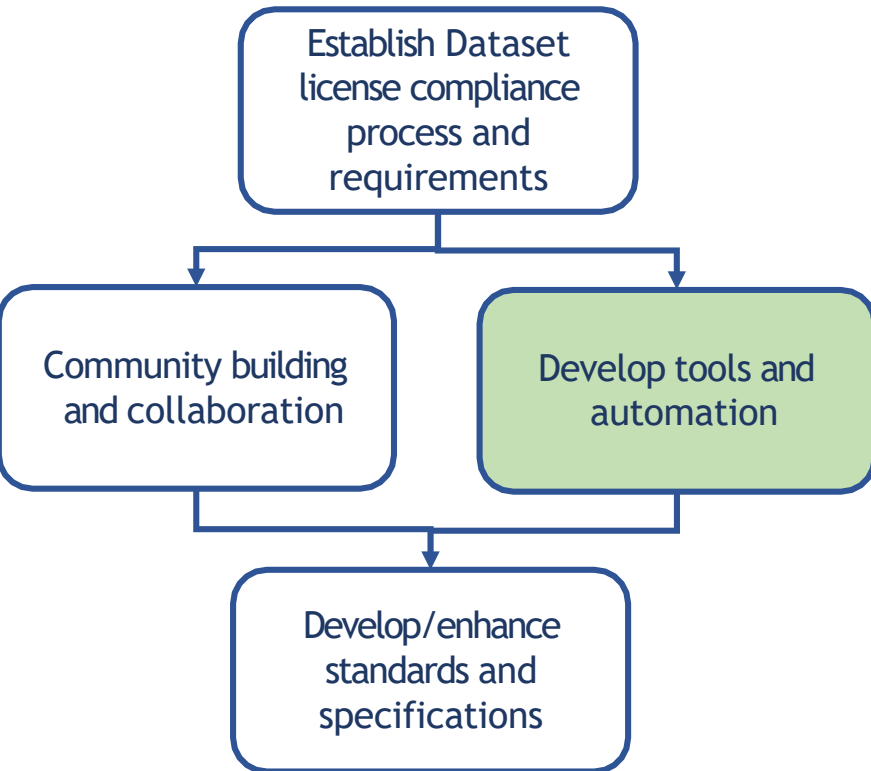
Challenges



Current progress



Road ahead



We developed an automatic license generator tool that helps the creators of datasets to assign license based on the rights and obligations that they wish to impose on the dataset

License-Generator

License-Generator

An easy-to-use license assistant.

Answer all the questions in the 3 steps to get your custom data license.

1 For Data itself 2 For Model

Question : 2.2 Can users perform Research?

Description : "Research" means to access the Data, use the Data to create or improve Models, but without the right to use the Output or resulting Trained Model for any purpose other than evaluating the Model Research under the same terms. Note that the next question pertains to making the results of the Research available (i.e. publishing them).

Yes No

Number of all Licenses: 2

MIT License	Select	Creative Commons Attribution 4.0 License	Select
data_access_rights: Yes		data_access_rights: Yes	
data_modification_rights: Yes		data_modification_rights: Yes	
data_network_rights: Yes		data_network_rights: N/A	
data_represent_rights: Yes		data_represent_rights: Yes	



# OpenDataology - Current progress



Recap



License compliance process for curated datasets



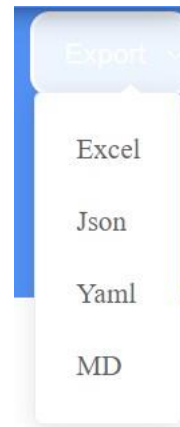
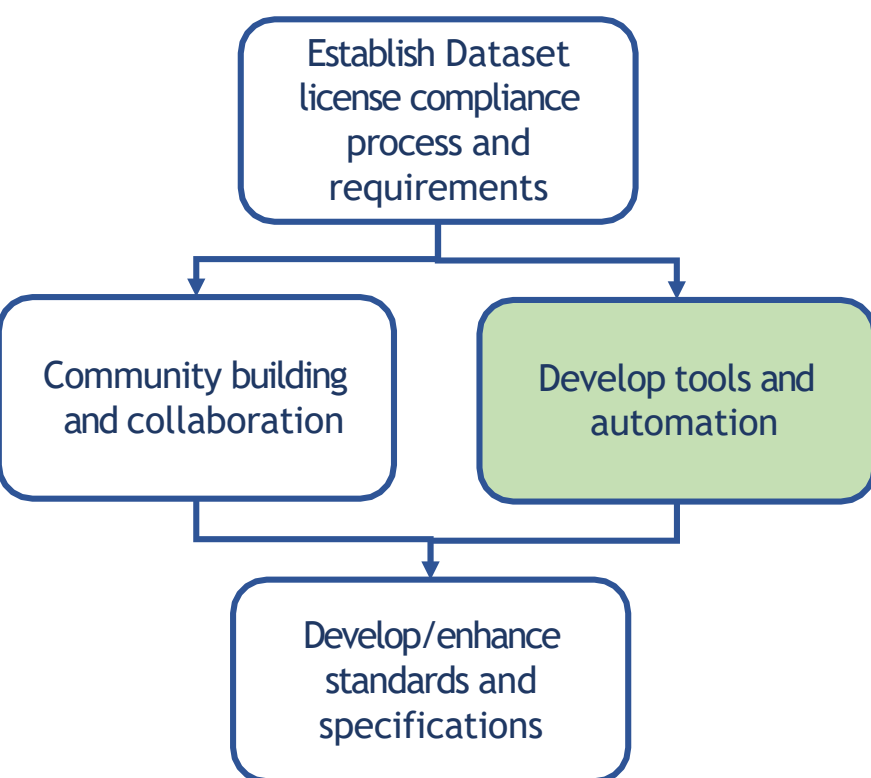
Challenges



Current progress



Road ahead



## Generate machine readable, serializable formats that enable compatibility with SPDX

```
SPDX-like Template
root:
  name: "CIFAR-10"
  versionInfo: "0.1"
  license:
    SPDXID: "CC0-1.0"
    name: "SPDXRef-License1"
  licensors:
    0: "Organization: ExampleCodeInspect ()"
    1: "Person: Jane Doe ()"
  sourceInfo: "Present on the official dataset website"
  homepage: "https://www.cs.toronto.edu/~kriz/cifar.html"
  content:
    0: "Name: imagenet Terms of Access"
    1: "hash: 55EB32BC75A822E4522317F4545A426B"
  originator: "Organization: ExampleCodeInspect (contact@example.com)"
  downloadable: true
  outlet:
```



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

Current core contributors

# OpenDataology - Current progress



Recap



License compliance process for curated datasets



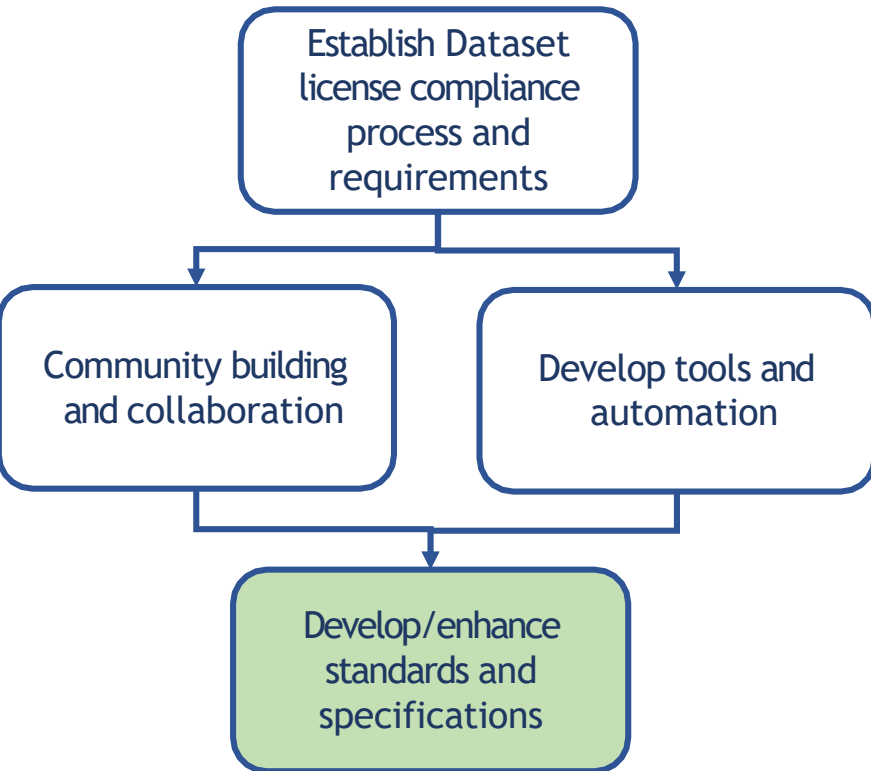
Challenges



Current progress



Road ahead



Dataset-related details	Dataset name	Dataset version	Origin date	Origin
	CIFAR-10	N/A	2009	<a href="https://www.cs.toronto.edu/~kriz/cifar.html">https://www.cs.toronto.edu/~kriz/cifar.html</a>
	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		<a href="https://www.cs.toronto.edu/~kriz/cifar.html">https://www.cs.toronto.edu/~kriz/cifar.html</a>	(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	
<b>Credit/Attribution Notice</b>						
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	Benchmark	Re-search	Publish	Internal Use	Commercialization	
					Output Model	
Rights	✓	✓	✓	✓	✓	Model Reverse Engineer
Obligations	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!

# Outline



OpenDataology project overview



Sandbox requirements



Collaboration with existing LF and LF-AI Projects



Challenges



Road ahead

# OpenDataology - Sandbox requirements



Have an open and documented technical governance.

<https://github.com/OpenDataology/OpenDataology/blob/main/GOVERNANCE.md>



Have an OSI approved license

<https://github.com/OpenDataology/OpenDataology/blob/main/LICENSE>



Have achieved and maintained a Core Infrastructure Initiative Best Practices Passing Badge .

11AUG2022  
<https://bestpractices.coreinfrastructure.org/en/projects/6032>

License CC BY 4.0 SPDX-License-Identifier: CC-BY-4.0

## OpenDataology Governance

The OpenDataology project provides content (standards, data, code and documentation) that helps organizations, especially AI enterprises, use publically available datasets compliantly.

A project of this scope requires input from a wide range of subject matter experts with different backgrounds and affiliations. As such we need a set of principles, roles and operating practices to ensure the results of our contributions are useful, have high quality and are widely consumable.

## OpenDataology

openssf best practices passing License MIT

### License

OpenDataology is licensed under MIT

Copyright 2022 OpenDataology

## OpenDataology

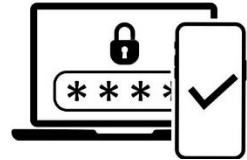
openssf best practices passing License MIT

<https://bestpractices.coreinfrastructure.org/en/projects/6032>

# OpenDataology - Sandbox requirements



Have its own GitHub organization that includes the following ([template](#) here)



Enablement of two-factor authentication.

<https://github.com/organizations/OpenDataology/settings/security>



A LICENSE file to every repo.

<https://github.com/OpenDataology/OpenDataology/blob/main/LICENSE>



A README file welcoming new community members to the project & explaining why the project is useful & how to get started.

<https://github.com/OpenDataology/OpenDataology/blob/main/README.md>

# OpenDataology - Sandbox requirements

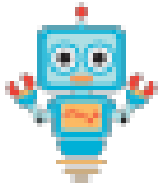


Have its own GitHub organization that includes the following ([template](#) here)



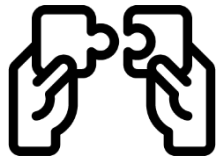
Invite The Linux Foundation (@thelinuxfoundation as a co-owner of the GitHub org.)

[https://github.com/orgs/OpenDataology/people/pending\\_invitations](https://github.com/orgs/OpenDataology/people/pending_invitations)



Enablement of the GitHub DCO app:  
<https://github.com/apps/dco>

<https://github.com/OpenDataology/OpenDataology/blob/main/LICENSE>



A CONTRIBUTING file explaining to other developers and your community of users how to contribute to the project. The file should explain what types of contributions are needed and how the process works.

<https://github.com/OpenDataology/OpenDataology/blob/main/CONTRIBUTING.md>

# OpenDataology - Sandbox requirements



Have its own GitHub organization that includes the following ([template](#) here)



A CODEOWNERS or COMMITTERS file to define individuals or teams that are responsible for code in a repository; document current project owners and current and emeritus committers.

<https://github.com/OpenDataology/OpenDataology/blob/main/CONTRIBUTING.md>



A CODE\_OF\_CONDUCT file that sets the ground rules for participants - [template](#) here

[https://github.com/OpenDataology/OpenDataology/blob/main/CODE\\_OF\\_CONDUCT.md](https://github.com/OpenDataology/OpenDataology/blob/main/CODE_OF_CONDUCT.md)



A GOVERNANCE file that documents the project's technical governance.

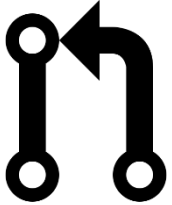
<https://github.com/OpenDataology/OpenDataology/blob/main/GOVERNANCE.md>



A SUPPORT file to let users and developers know about ways to get help with your project

<https://github.com/OpenDataology/OpenDataology/blob/main/SUPPORT.md>

# OpenDataology - Sandbox requirements



Submit a completed Project Contribution Proposal via a GitHub pull request to <https://github.com/lfai/proposing-projects/tree/master/proposals> .

<https://github.com/lfai/proposing-projects/pull/52>

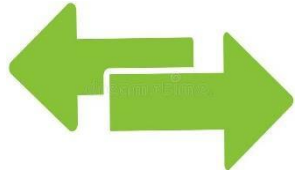


Identify who on the project will be handling security issues (could be a team).

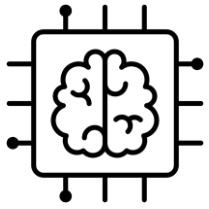
<https://github.com/OpenDataology/OpenDataology/blob/main/SUPPORT.md>



# OpenDataology - Sandbox requirements



Be deemed by the TAC and Governing Board to add value to the artificial intelligence, data and analytics space and to fall within the mission and scope of LF AI & Data.



Be deemed by the TAC and Governing Board to add value to the artificial intelligence, data and analytics space and to fall within the mission and scope of LF AI & Data.



Receive the affirmative vote of the TAC.



To be decided in the TAC meeting

# Outline



OpenDataology project overview



Sandbox requirements



Collaboration with existing LF and LF-AI Projects



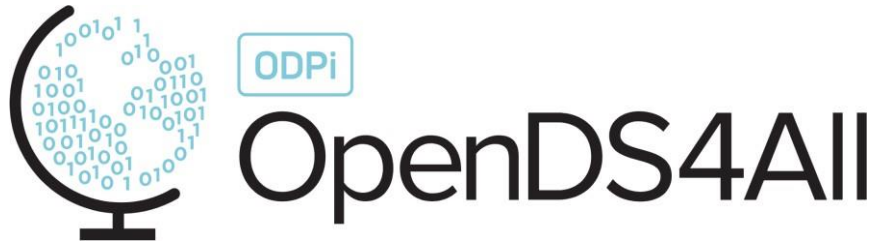
Challenges



Road ahead

# OpenDataology - Collaboration with existing LF-AI and LF projects

We are either actively collaborating or initiating collaboration with these LF-AI and LF projects



# OpenDataology - Collaboration with existing LF-AI and LF projects

We are either actively collaborating or initiating collaboration with these LF-AI and LF projects

Interested in OpenDS4All and hope to get support Inbox x



**zicheng qu** <quzicheng315@gmail.com>  
to opens4all-technical-discuss

Tue, May 24, 2:40 PM (9 days ago) ☆ ↶ ⋮

Dear OpenDS4All team,

I am Zev Qu, I learn the introduction of OpenDS4All and the core idea, and it attracts me very much.

We are currently working on a project called OpenDataology about best practices for dataset metadata and license compliance. The project is now about entering the sandbox phase of LF-AI & Data, and our primary goal is to make each data license clear and understandable in both human and machine-readable ways, so that potential license compliance risks can be identified for users when they adopt publicly available datasets or even datasets from multiple data sources.

We also have deep cooperation with SPDX international standards and its community. They are now launching AI BOM, a profile that can express AI software attributes in software, models, data, etc. It can help to achieve machine-readable and compliance analysis automation by using AI BOM schema.

I think OpenDS4All has great project inspiration and material around overview foundation data practices and the whole pipeline, and there will be many opportunities for cooperation in the future. Our project link is <https://github.com/OpenDataology/OpenDataology>.

I really hope OpenDataology can get your support and feel free to give us any feedback. Looking forward to your reply.

Best regards,  
OpenDataology team



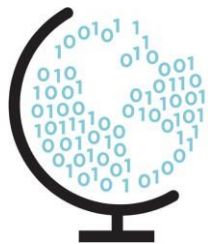
**Andre de Waal**  
to me, Andre

Thu, May 26, 2:46 AM (7 days ago) ☆ ↶ ⋮

Dear Zev,

No problem, let's reconnect again towards the end of the year and see how your project is progressing.

Best wishes,  
Andre



ODPi


# OpenDS4All

# OpenDataology - Collaboration with existing LF-AI and LF projects

We are either actively collaborating or initiating collaboration with these LF-AI and LF projects



**OpenBytes**

 **Clement Li** 上午 11:07 ✓ ✕ 🗨️ 🗺️ 📄 📌 ⋮


Hello, Edward. I'm Clement Li, I used to join openbytes and data-licensing channel, and of course I got lots of interesting ideas here. In recent months we have been preparing a project on dataset compliance best practice at LF-AI & Data. Now the project is about to enter the sandbox, our goal is to make each of data licenses clear and easy to understand, and most important thing is that we will cooperate deeply with SPDX international standards and its community. SPDX is an open standard for communicating software bill of material information, and this year SPDX provide a profile named AI BOM which shows the properties of AI Software, Data and Models, further more it will help people to build dataset lineage information for AI Datasets like they do on software.

I think OpenBytes is an amazing project about AI and Data, about format, metadata and license, there will be many opportunities for cooperation in the future. Our github project is here <https://github.com/OpenDataology/OpenDataology> , I really hope that you can support this project and feel free to give any feedback to us. Thanks.

**OpenDataology/OpenDataology**  
Practice of AI dataset metadata and license compliance

Stars	Last updated
3	14 hours ago

添加人 GitHub

 **Edward Cui** 上午 11:09

Thanks for reaching out! Congrats on the progress. And I will look into it for sure.

😊 1 🗨️

# OpenDataology - Collaboration with existing LF-AI and LF projects

We have been working with SPDX community to create AIBOM for AI software and datasets that will have fields that enables dataset license compliance



SPDX AI team minutes, May 25, 2022

Attendees:

- Gopi Krishnan Rajbahadur
- Zev (Zicheng) Qu
- Kate Stewart
- Clement
- L Jean Camp

Regrets:

Derek

Agenda

- Data Sheets (Gopi)
- Overview of Mindspore (Zev)

Notes

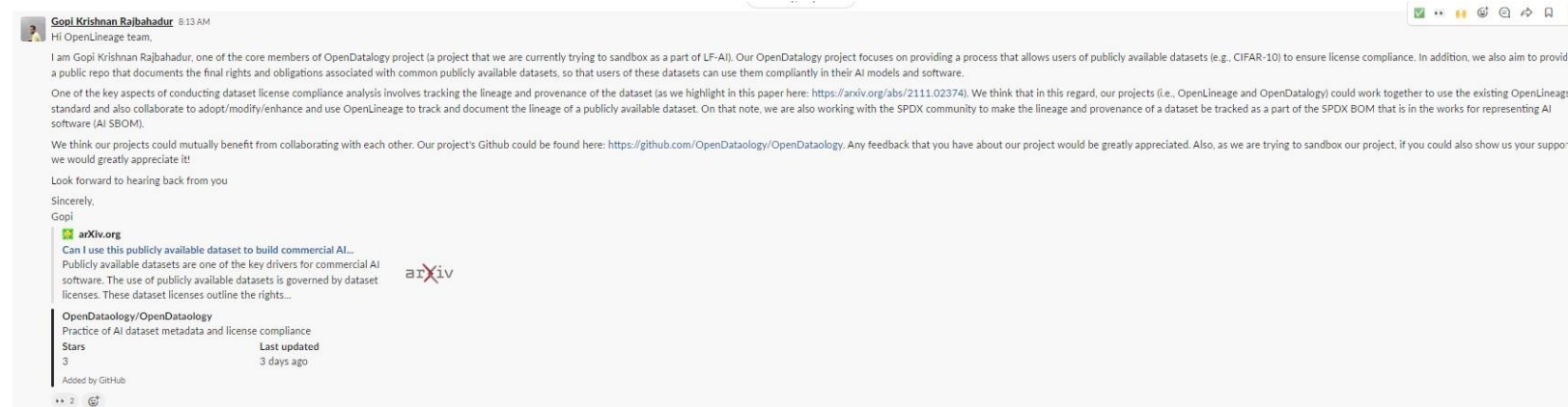
\* Zev - Mindspore SIG - interested in compliance. Mindspore has >300 model, and over 100 datasets. Looking to practice reference AI BOM in there.

\* AI SBOM mailing list - Gopi requesting it be created; TODO: Kate to create, and minutes to go up github.

# OpenDataology - Collaboration with existing LF-AI and LF projects

We are either actively collaborating or initiating collaboration with these LF-AI and LF projects

Open Lineage



# Outline



OpenDataology project overview



Sandbox requirements



Collaboration with existing LF and LF-AI Projects



Challenges



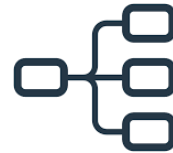
Road ahead



# 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



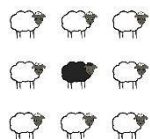
Unclear license locations



Identifying the minimum licensable data unit



Multiple license interactions and their effects are unclear



Multiple copies/variants of dataset hosted in different places

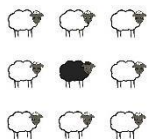
# Challenges in ensuring dataset license compliance



Unclear licensing range



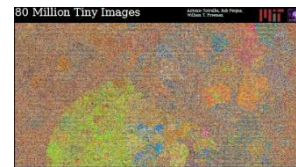
Unclear license locations



Multiple copies/variants of dataset hosted in different places



200X ??



2008

## The CIFAR-10 dataset

• [Learning Multiple Layers of Features from Tiny Images](#), Alex Krizhevsky, 2009.

2009

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

2022



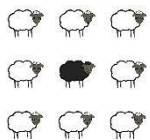
# Challenges in ensuring dataset license compliance



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



License is provided in the GitHub page

IMAGENET

License is provided along with the website

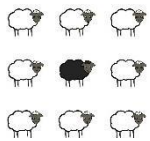
# Challenges in ensuring dataset license compliance



Unclear licensing range



Unclear license locations



Multiple copies/variants  
of dataset hosted in  
different places

11AUG2022

## The CIFAR-10 dataset

 PyTorch

 Keras

 kaggle

 GitHub

 DeepAI

  
TensorFlow

# Challenges in ensuring dataset license compliance



All the data sources are not specified



Identifying the minimum licensable data unit



## The CIFAR-10 dataset



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

# Challenges in ensuring dataset license compliance



All the data sources are not specified



Identifying the minimum licensable data unit

11AUG2022



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



Rights and obligations  
are unclear



Multiple license  
interactions and their  
effects are unclear

## The CIFAR-10 dataset

### IMAGENET

Please cite it if you intend to use this dataset.

- [Learning Multiple Layers of Features from Tiny Images](#), Alex Krizhevsky, 2009.

[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



Rights and obligations are unclear



Multiple license interactions and their effects are unclear



## 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!



### 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



OpenDataology project overview



Sandbox requirements



Collaboration with existing LF and LF-AI Projects



Challenges



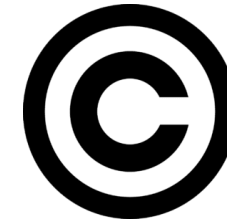
Road ahead

# OpenDataology - Look ahead

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



License compliance



Copyright compliance



Privacy compliance



Ethics compliance

Establish Dataset license compliance process and requirements

Community building and collaboration


Develop tools and automation

Develop/enhance standards and specifications

# OpenDataology - Look 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



2025-Q2  
Ethics compliance



Establish Dataset  
license compliance  
process and  
requirements

Community building  
and collaboration

Develop tools and  
automation

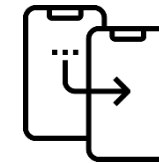
Develop/enhance  
standards and  
specifications

# OpenDataology - Look ahead

We aim to develop various tools and automation procedures such as



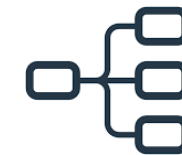
Automated license generator



Ensuring compliance through data clone detection



Automated provenance extraction



Automated lineage extraction



License Compliance process automation

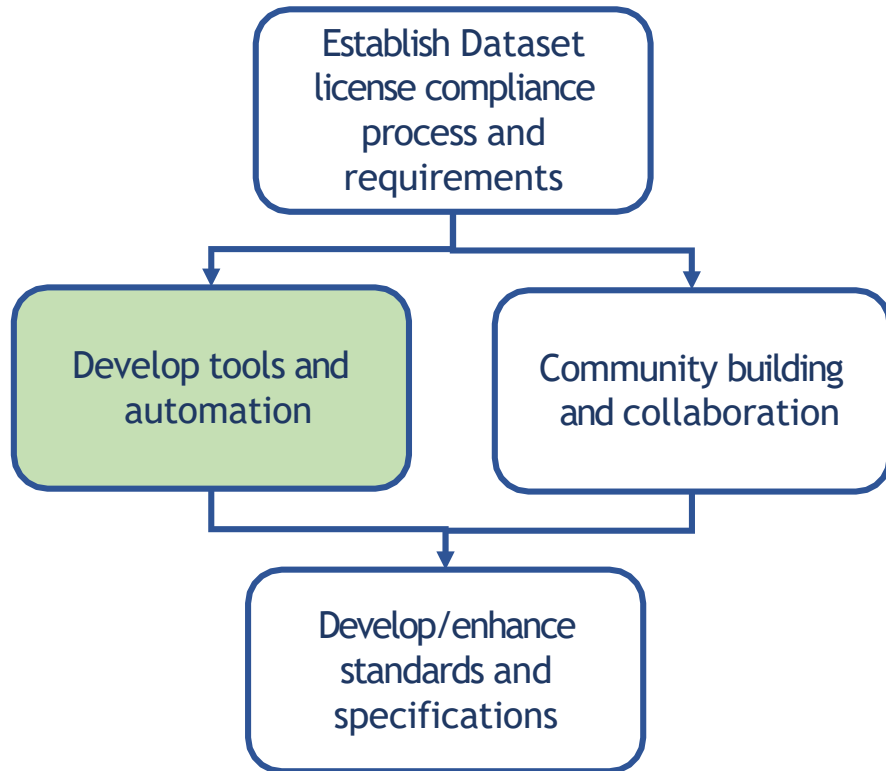
Establish Dataset license compliance process and requirements

Develop tools and automation

Community building and collaboration

Develop/enhance standards and specifications

# OpenDataology - Look 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

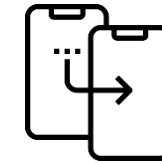
Tools that helps users extract and document the provenance and lineage details of datasets automatically using NLP on relevant documents and websites



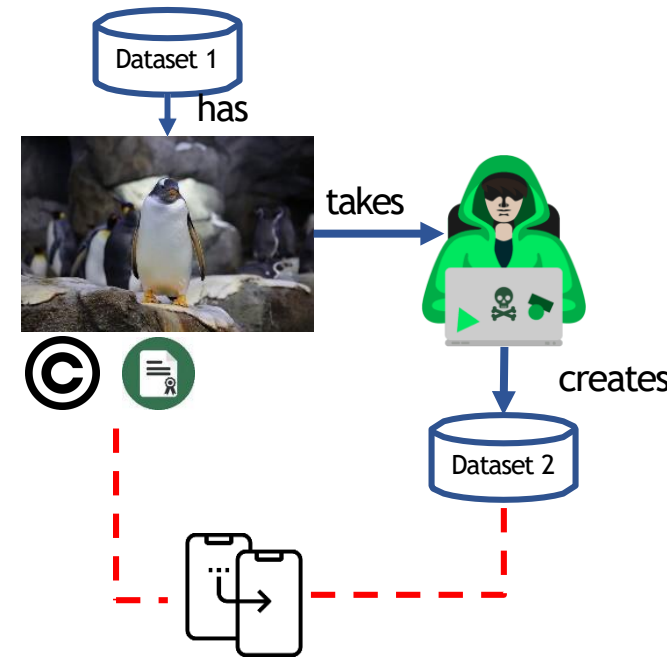
Automated lineage extraction

# OpenDataology - Look ahead

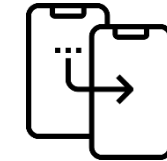
We aim to develop various tools and automation procedures such as



Ensuring compliance through data clone detection



Data clone



Define clone types

Create detection strategies

- Type 1
- Type 2
- Type 3
- Type 4

Establish Dataset license compliance process and requirements

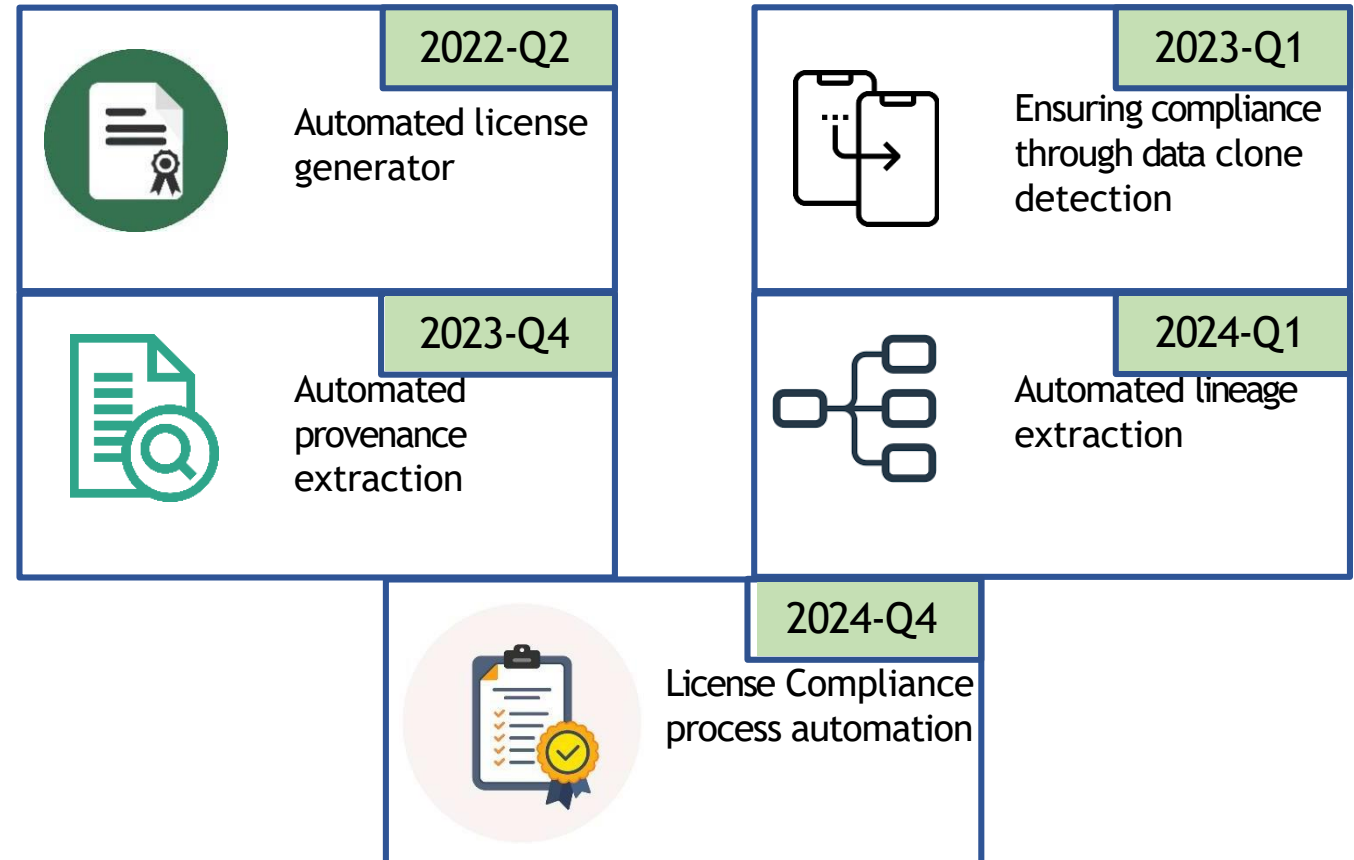
Develop tools and automation

Community building and collaboration

Develop/enhance standards and specifications

# OpenDataology - Look ahead

We aim to develop various tools and automation procedures such as



Establish Dataset  
license compliance  
process and  
requirements

Develop tools and  
automation

Community building  
and collaboration

Develop/enhance  
standards and  
specifications

# OpenDataology - Look 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



Establish a Slack channel for discussion

Establish Dataset license compliance process and requirements

Develop tools and automation

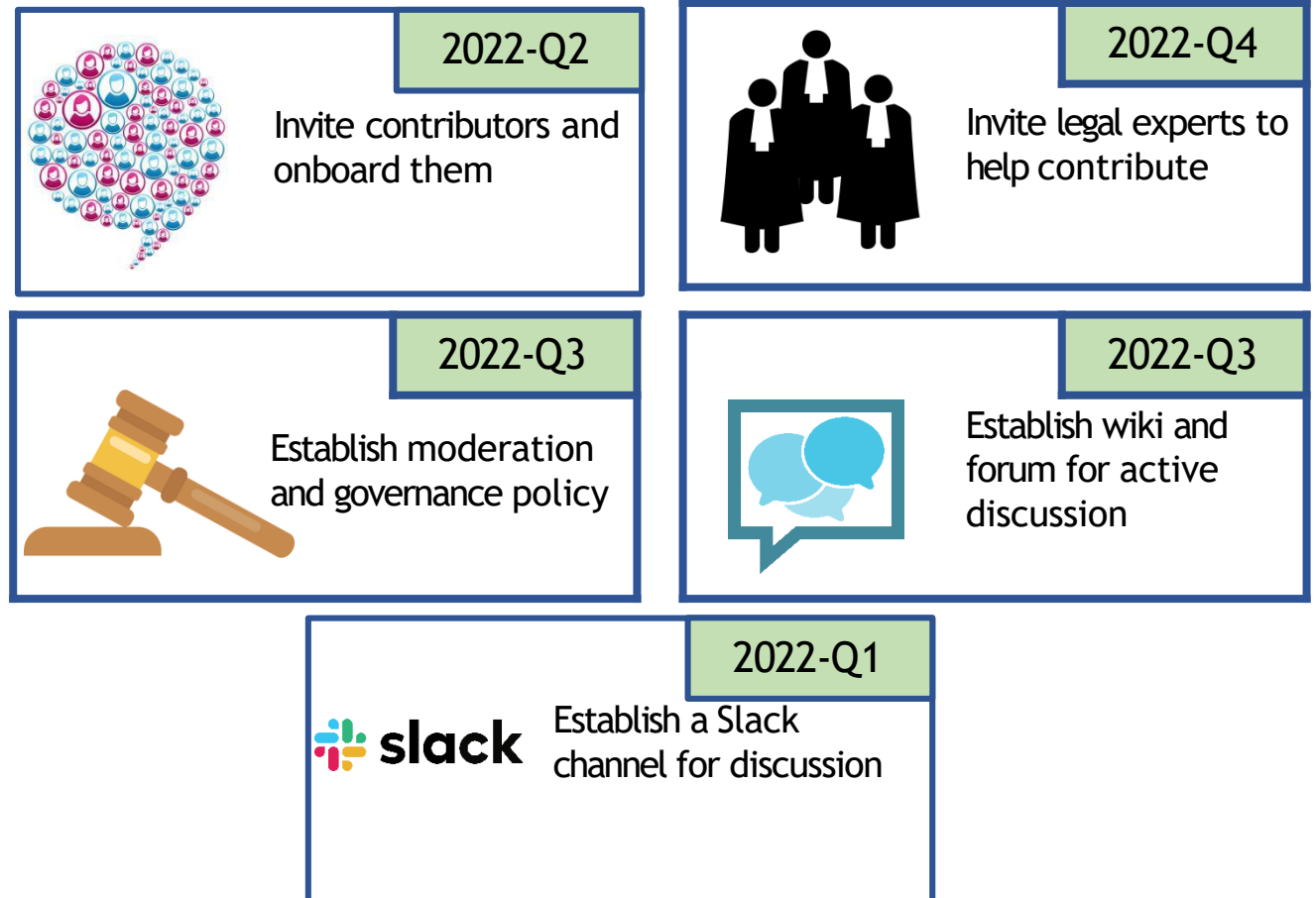
Community building and collaboration

Develop/enhance standards and specifications



# OpenDataology - Look ahead

We aim to develop various tools and automation procedures such as



Establish Dataset license compliance process and requirements

Develop tools and automation

Community building and collaboration

Develop/enhance standards and specifications

# OpenDataology - Look ahead

We aim to develop various tools and automation procedures such as



Enhance existing standards



Create new standards

Establish Dataset  
license compliance  
process and  
requirements

Develop tools and  
automation

Community building  
and collaboration

Develop/enhance  
standards and  
specifications

# OpenDataology - Look ahead

We aim to develop various tools and automation procedures such as

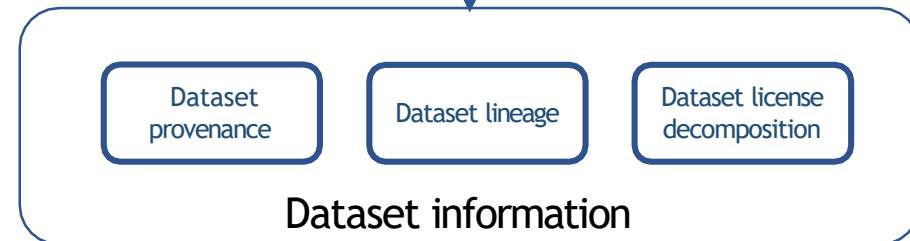


Enhance existing standards

## What makes up an SPDX Document?



Enhance with



Establish Dataset license compliance process and requirements

Develop tools and automation

Community building and collaboration

Develop/enhance standards and specifications

# OpenDataology - Look ahead

We aim to develop various tools and automation procedures such as



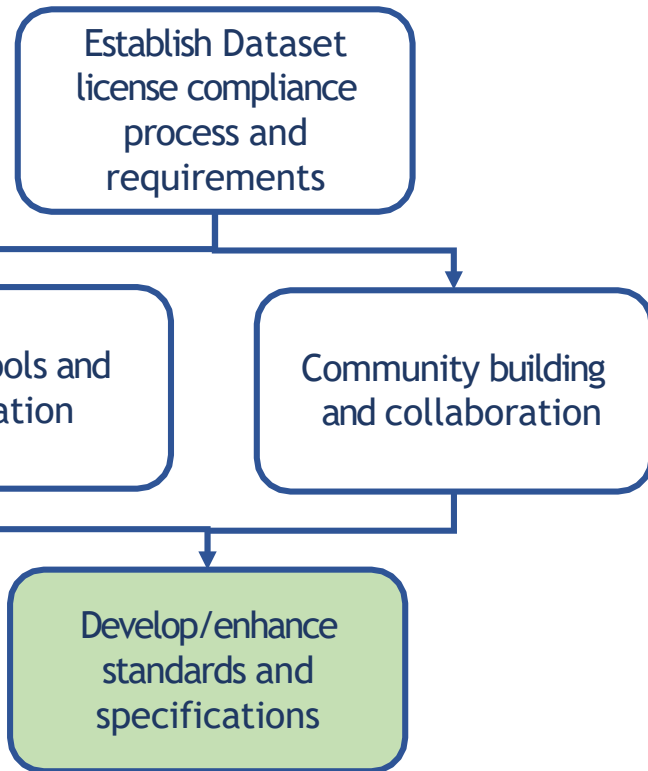
2022-Q4

Enhance existing standards

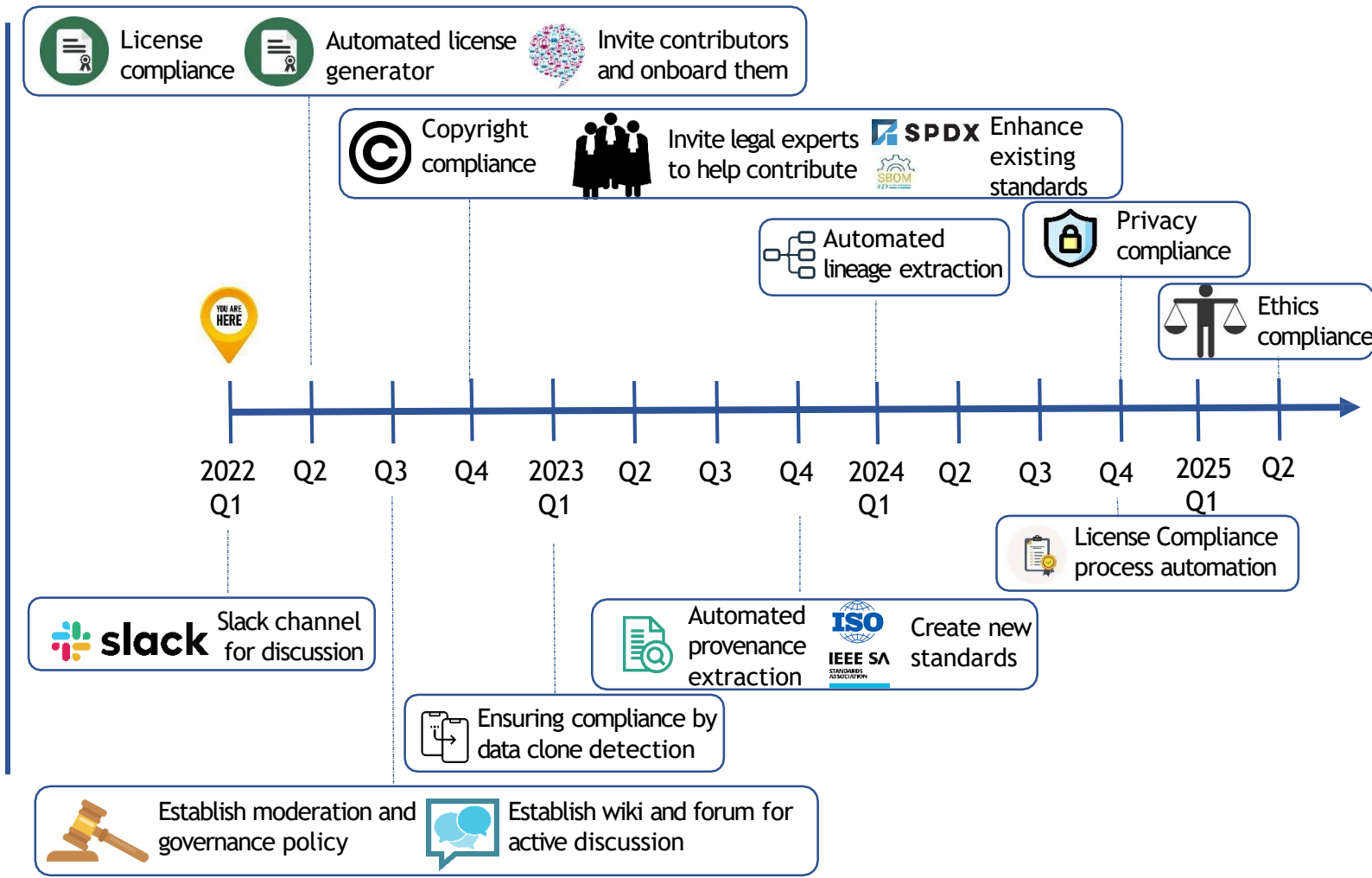
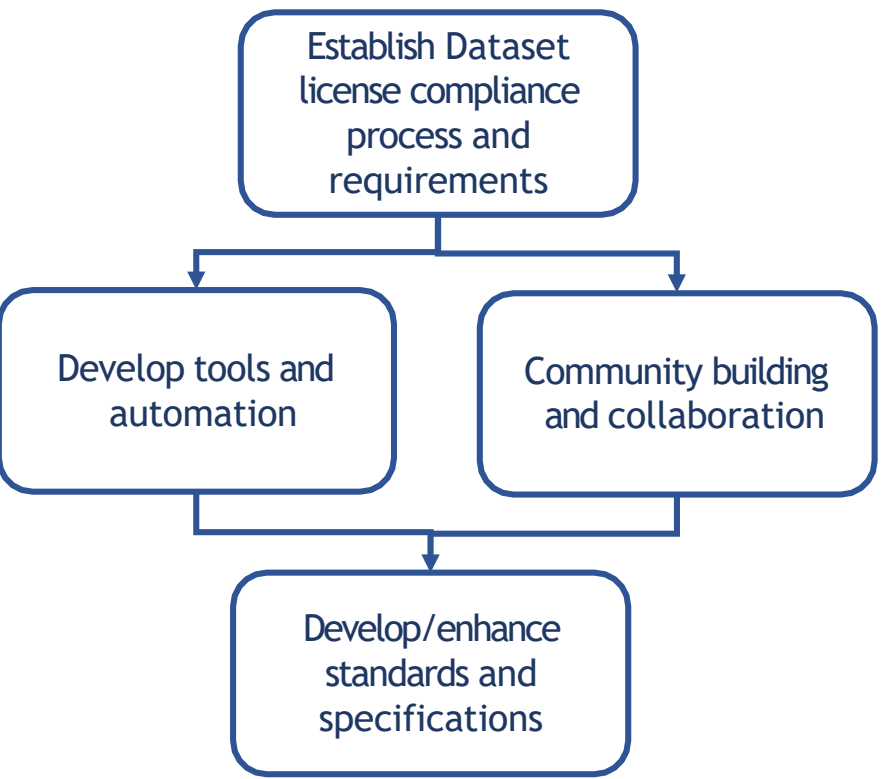


2023-Q4

Create new standards



# OpenDataology - Look ahead



# TAC Vote on Project Proposal: OpenDataology

## **Proposed Resolution:**

The TAC approves the OpenDataology Project as an incubation project at the sandbox level of the LF AI Foundation

# Upcoming TAC Meetings

 **DLF** AI & DATA

# Upcoming TAC Meetings

- › August 25, 2022 – Feathr – new incubation project
- › September 8, 2022 – Open

Please note we will be restarting project reviews in the September timeframe. 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

 **OLF** 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](mailto:legal@linuxfoundation.org) with any questions about The Linux Foundation's policies or the notices set forth on this slide.