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

December 16, 2021

LF AI & DATA

Antitrust Policy

- Linux Foundation meetings involve participation by industry competitors, and it is the intention of the Linux Foundation to conduct all of its activities in accordance with applicable antitrust and competition laws. It is therefore extremely important that attendees adhere to meeting agendas, and be aware of, and not participate in, any activities that are prohibited under applicable US state, federal or foreign antitrust and competition laws.

- Examples of types of actions that are prohibited at Linux Foundation meetings and in connection with Linux Foundation activities are described in the Linux Foundation Antitrust Policy available at http://www.linuxfoundation.org/antitrust-policy. If you have questions about these matters, please contact your company counsel, or if you are a member of the Linux Foundation, feel free to contact Andrew Updegrove of the firm of Gesmer Undergone LLP, which provides legal counsel to the Linux Foundation.

Recording of Calls

Reminder:

TAC calls are recorded and available for viewing on the TAC Wiki

Reminder: LF AI & Data Useful Links

- Web site: lfaidata.foundation
- Wiki: wiki.lfaidata.foundation
- GitHub: github.com/lfaidata
- Landscape: https://landscape.lfaidata.foundation or https://l.lfaidata.foundation
- Mail Lists: https://lists.lfaidata.foundation
- Slack: https://slack.lfaidata.foundation
- Youtube: https://www.youtube.com/channel/UCfasaeqXJBCAJMNO9HcHfbA
- LF AI Logos: https://github.com/lfaidata/artwork/tree/master/lfaidata
- LF AI Presentation Template: https://drive.google.com/file/d/1eiDNJvXCqSZHT4Zk_-czASlz2GTBRZk2/view?usp=sharing

- Events Page on LF AI Website: https://lfaidata.foundation/events/
- Events Calendar on LF AI Wiki (subscribe available): https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=12091544
- Event Wiki Pages: https://wiki.lfaidata.foundation/display/DL/LF+AI+Data+Foundation+Events

Agenda

- Roll Call (2 mins)
- Approval of Minutes from previous meeting (2 mins)
- Annual Review for Datapractices (20minutes)
- ML Worflow Committee update (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
AT&T	Premier	Voting Member	USA	Anwar Aftab	
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
Zilliz	Premier	Voting Member	China	Jun Gu	Xiaofan Luan
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	

Minutes approval

LF AI & DATA

Approval of December 2nd, 2021 Minutes

Draft minutes from the December 2nd TAC call were previously distributed to the TAC members via the mailing list

Proposed Resolution:

- That the minutes of the December 2nd meeting of the Technical Advisory Council of the LF AI & Data Foundation are hereby approved.

Annual Review for Datapractices

Amber Thomas
12/16/2021

LF AI & DATA

Datapractices

DATAPRACTICES.ORG

Brief Description:

DataPractices is a “Manifesto for Data Practices,” comprised of values and principles to illustrate the most effective, modern, and ethical approach to data teamwork.

Contributed by:

Initially contributed to the Linux Foundation by data.world in March 2019, and added as an Incubation Project in December 2020

Key Links:

Github: https://github.com/datadotworld/data-practices-site

Website: https://datapractices.org/

Artwork: https://artwork.lfaidata.foundation/projects/datapractices/

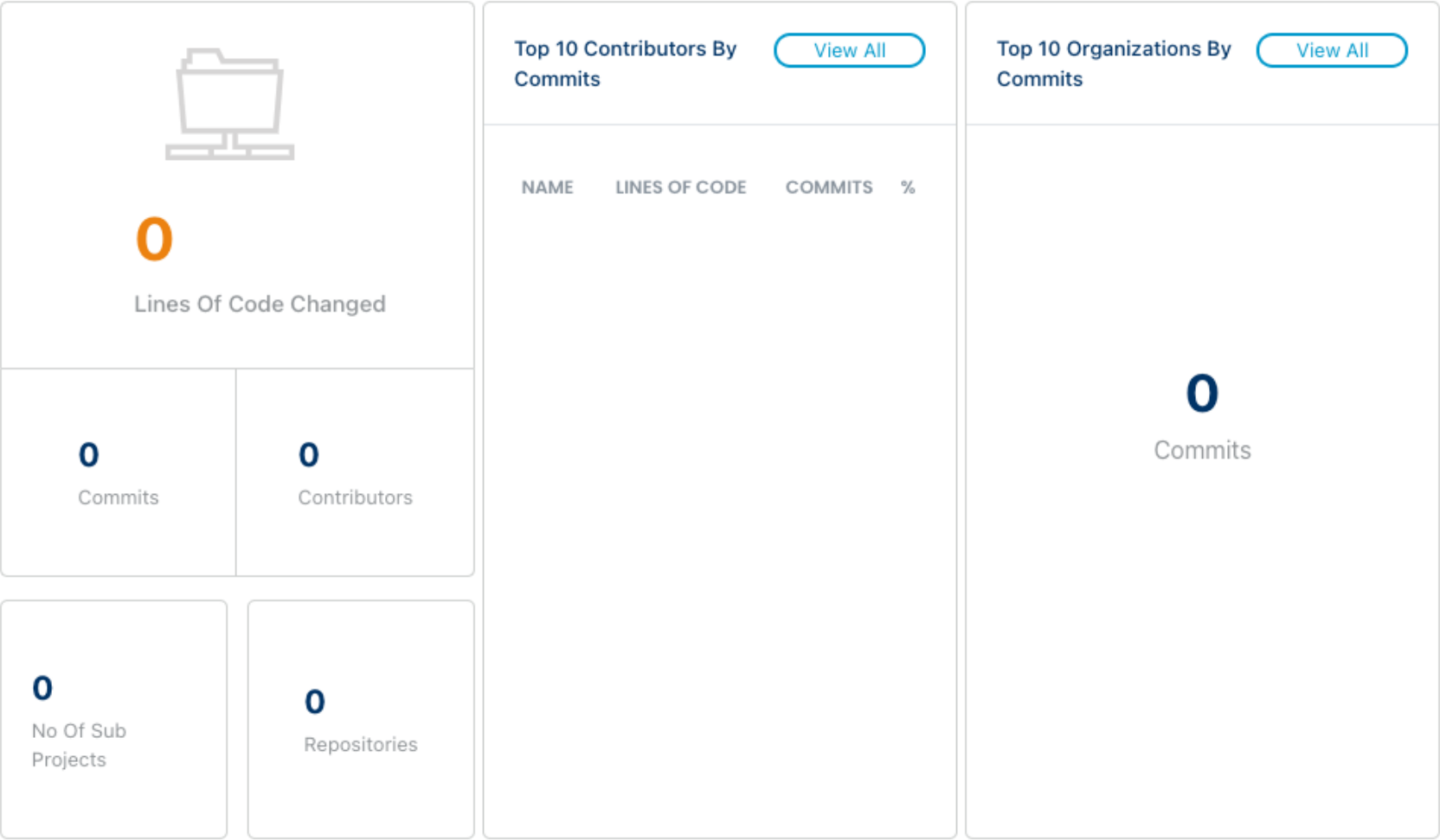
Mailing lists:

- datapractices-announce
- datapractices-technical-discuss
- datapractices-tsc

Organizations contributing

- AirBnB
- Ancestry
- Associated Press
- Ayasdi
- Bayes Impact
- City of Boston
- Charles Schwab
- Comcast
- Continuum
- d3
- Data for Democracy
- Data Syndrome
- data.world
- Domino Data Lab
- Galvanize
- George Washington Univ.
- Harris Data
- Huge
- Jupyter
- Macroscope Media
- Nextdoor
- Pandas
- Polaris
- Shasta Ventures
- Tableau
- UC Berkeley
- Vega

Contributions



Key Achievements in the past year

- New hire to focus more on development of the program moving forward
- Developed a plan for 2022
  - Expansion of content types
    - Slides for workshops and conferences
    - Additional content and background information to better support workshop facilitators
  - Connect content with other programs
    - IBM’s OpenDS4All

Areas the project could use help on

- Finding contributors to update and expand our existing content

Feedback on working with LF AI & Data

- On hold, we didn’t fully utilize your potential involvement

TAC Open Discussion

LF AI & DATA

MLWorkflow Committee: Challenges and proposed solutions for dataset license compliance analysis

(Howard) Huang Zhipeng

LF AI & DATA

Can I use this publicly available dataset to build commercial AI software?

Gopi Krishnan Rajbahadur

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

AI Software development and commercialization is driven by the availability of datasets

QUARTZ

IT'S NOT ABOUT THE ALGORITHM

The data that transformed AI research—and possibly the world

Forbes

What Exactly Is Artificial Intelligence? (Hint: It's All About The Datasets)

UNITE.AI

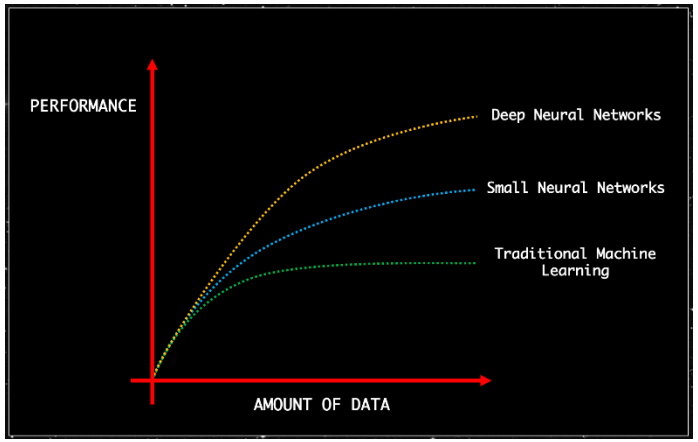
A Cartel of Influential Datasets Is Dominating Machine Learning Research, New Study Suggests

Harvard Business Review

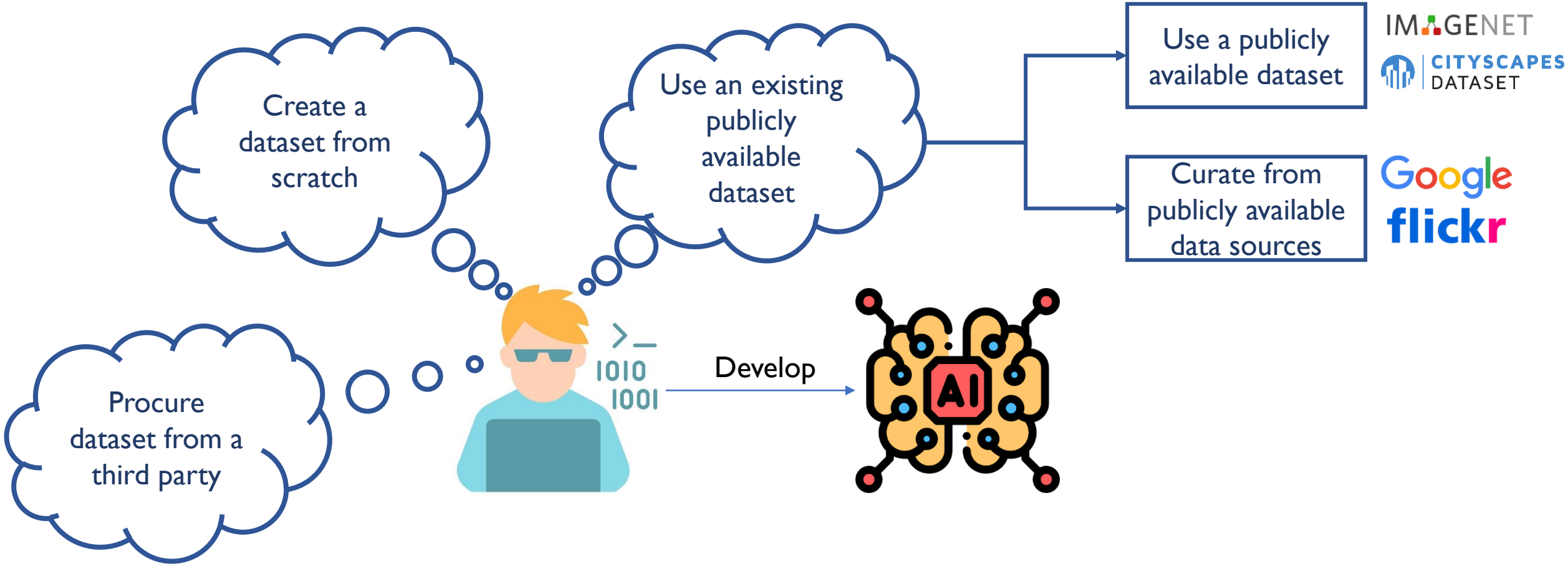
Small Data Can Play a Big Role in AI

RESEARCH AND MARKETS
THE WORLD'S LARGEST MARKET RESEARCH STORE

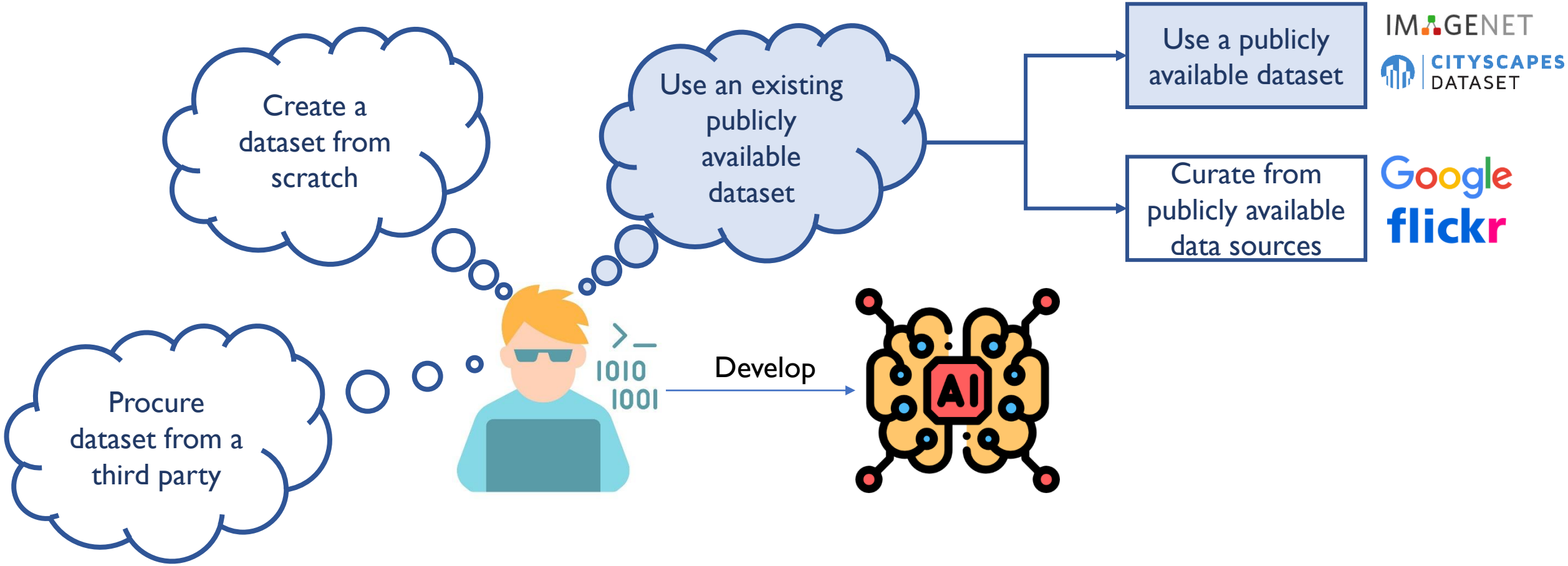
The Global AI Training Dataset Market size is expected to reach $3.1 billion by 2027, rising at a market growth of 17.4% CAGR during the forecast period.



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



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

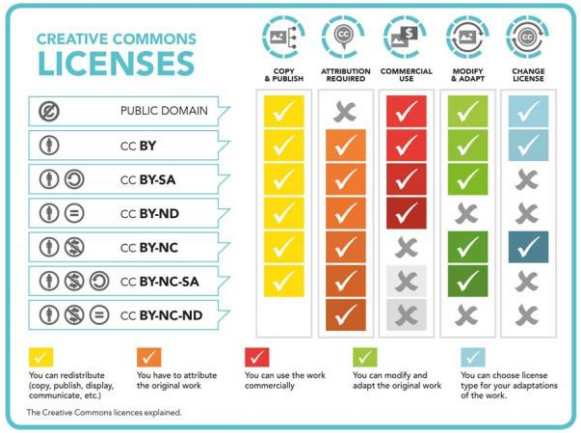


Similar to open-source software, the use of a dataset is completely governed by its license

License

Rights

The rights on the dataset that the users are entitled to



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.

Obligations

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

A key goal of our presentation is to propose an approach to assess the potential license compliance related risks associated with using a publicly available dataset to build commercial 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

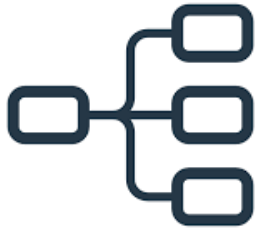
The views presented in this presentation are that of the authors and it does not reflect on the views presented by any corporation or organization.

Unlike OSS, conducting license compatibility analysis for datasets have several challenges

Unclear rights and obligations

Unclear dataset origin

Non-standard license locations



Unclear data sources

Unlike OSS, conducting license compatibility analysis for datasets have several challenges

Unclear rights and obligations

Location not found

Non-standard license locations

Unclear dataset origin

Unclear data sources

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

The rights and obligations associated with a dataset's license is unclear

Unlike OSS, conducting license compatibility analysis for datasets have several challenges

Unclear rights and obligations

Non-standard license locations

Unclear dataset origin

Unknown data sources

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

The licenses are not documented or provided on a standard location

Unlike OSS, conducting license compatibility analysis for datasets have several challenges

Unclear rights and obligations

Non-standard license locations

Unclear dataset origin

Unclear data sources

The CIFAR-10 dataset

PyTorch Keras kaggle GitHub

DeepAI TensorFlow

Dataset being available in multiple platforms makes it hard to identify dataset's provenance and its license

Unlike OSS, conducting license compatibility analysis for datasets have several challenges

Unclear rights and obligations

Location not found

Non-standard license locations

Unclear dataset origin

Unclear data sources

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

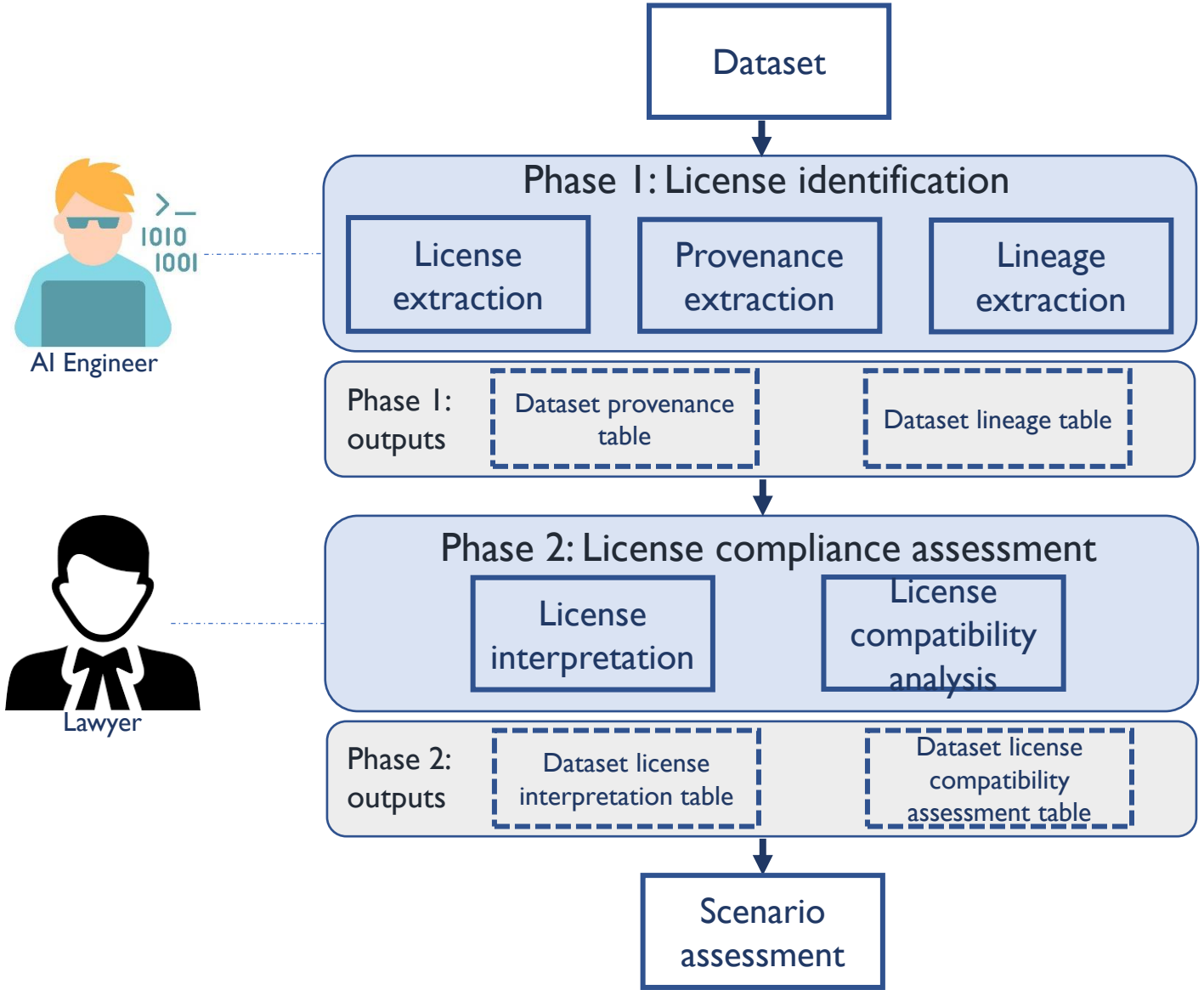
The CIFAR-10 dataset

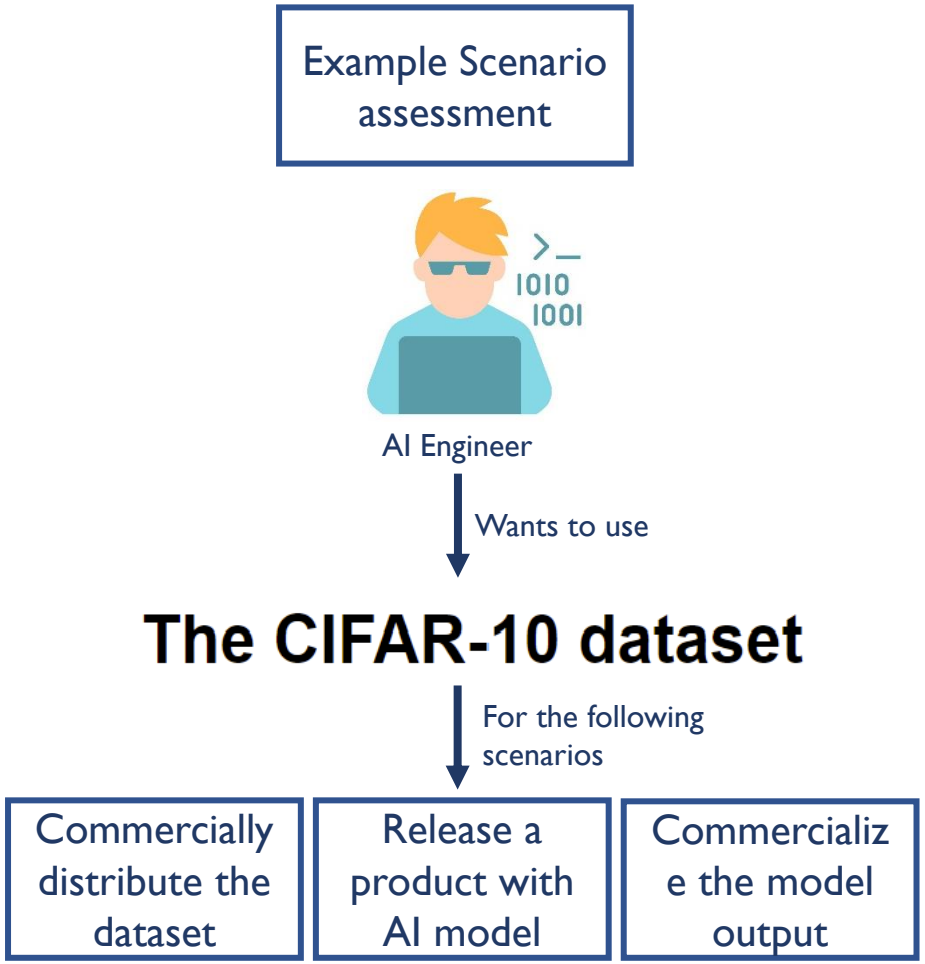
These data sources are not specified

The license of these data sources are not taken into consideration when considering the CIFAR-10 dataset's license

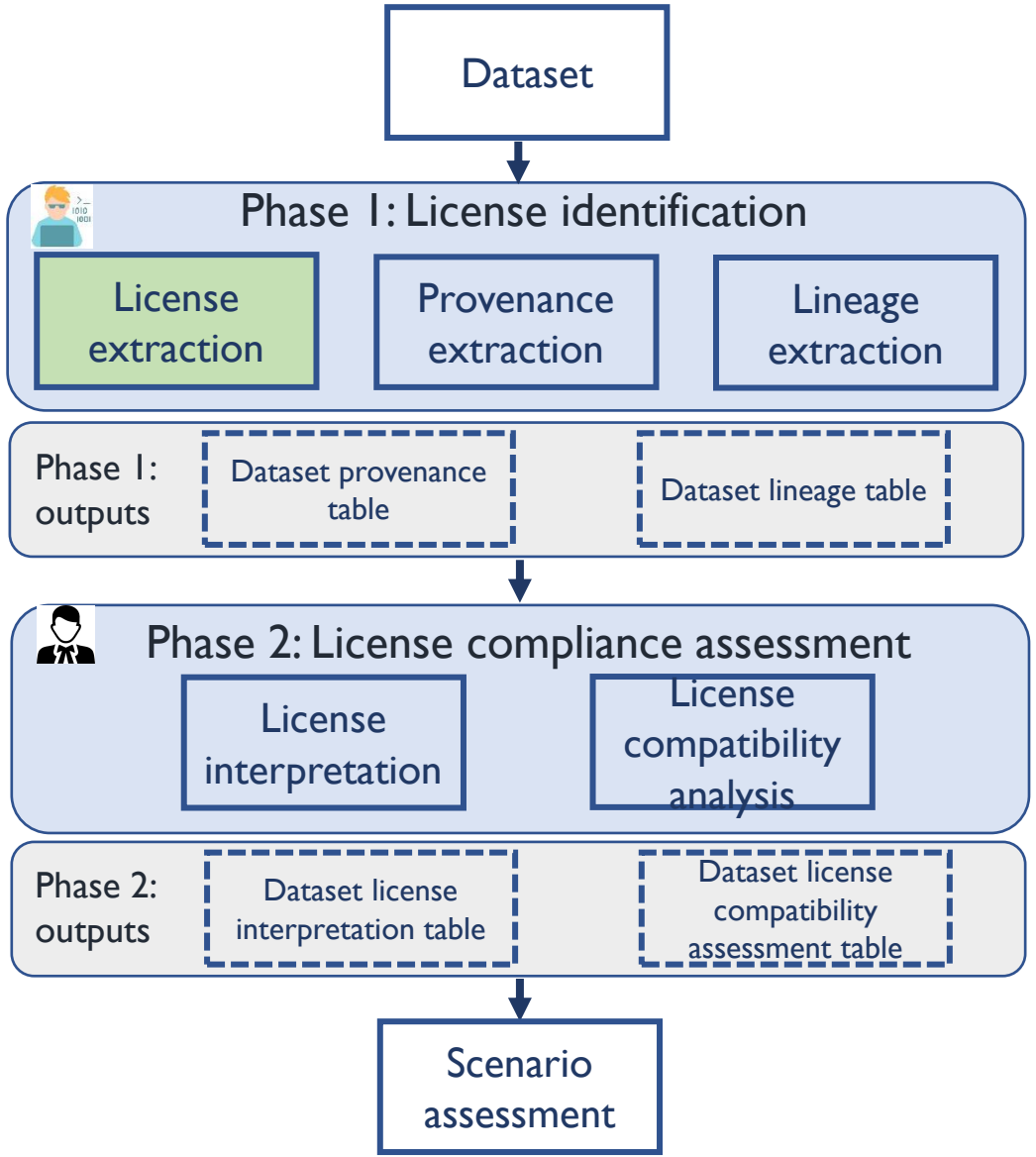
Data sources and their license not being mentioned makes it hard to ascertain the rights and obligation of the license associated with the given dataset

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

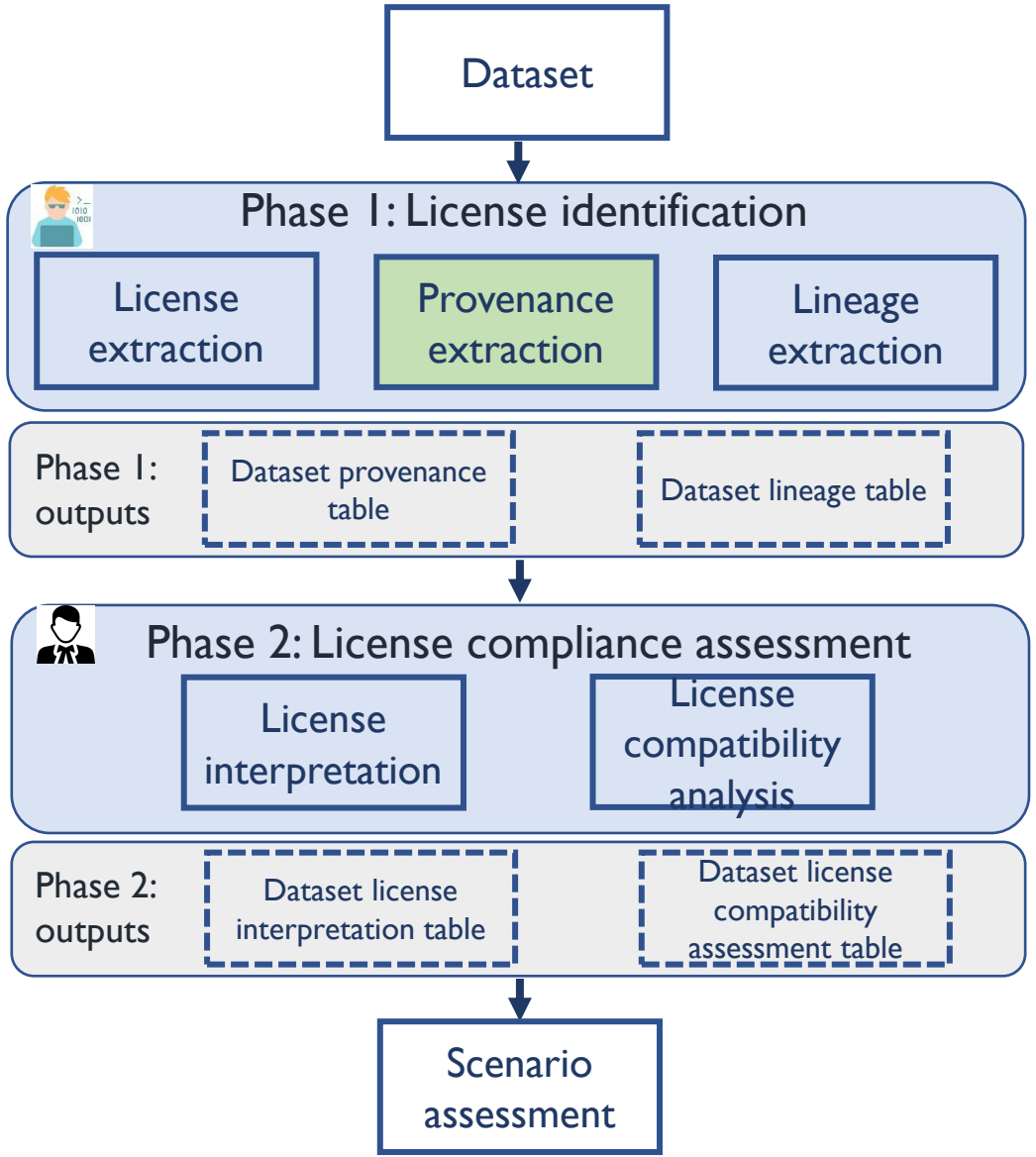


CIFAR-10 License (available on official website)

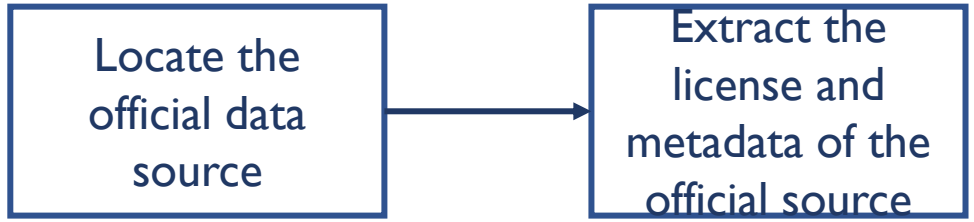
Please cite it if you intend to use this dataset.

- 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

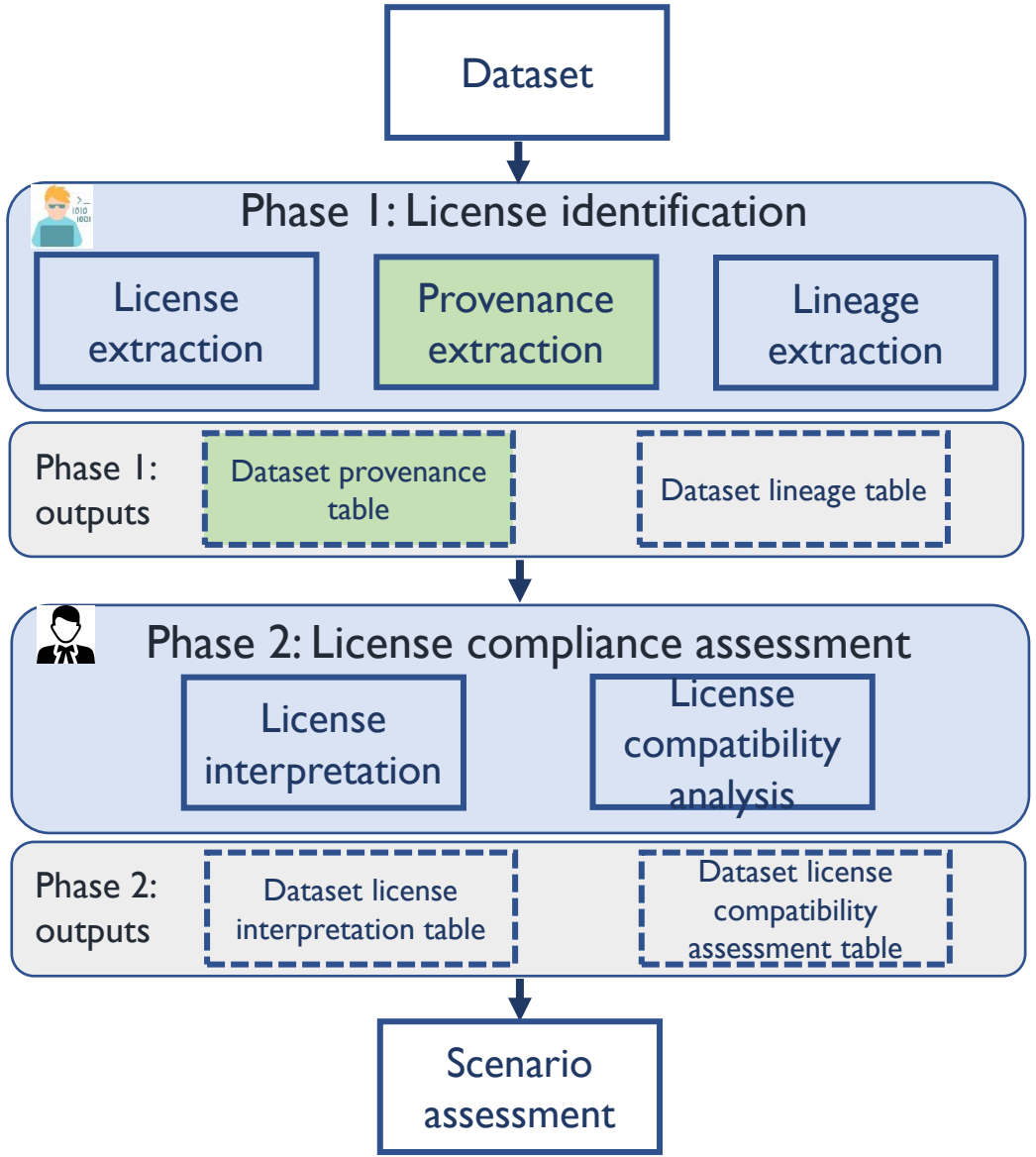


Provenance extraction sub-steps



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

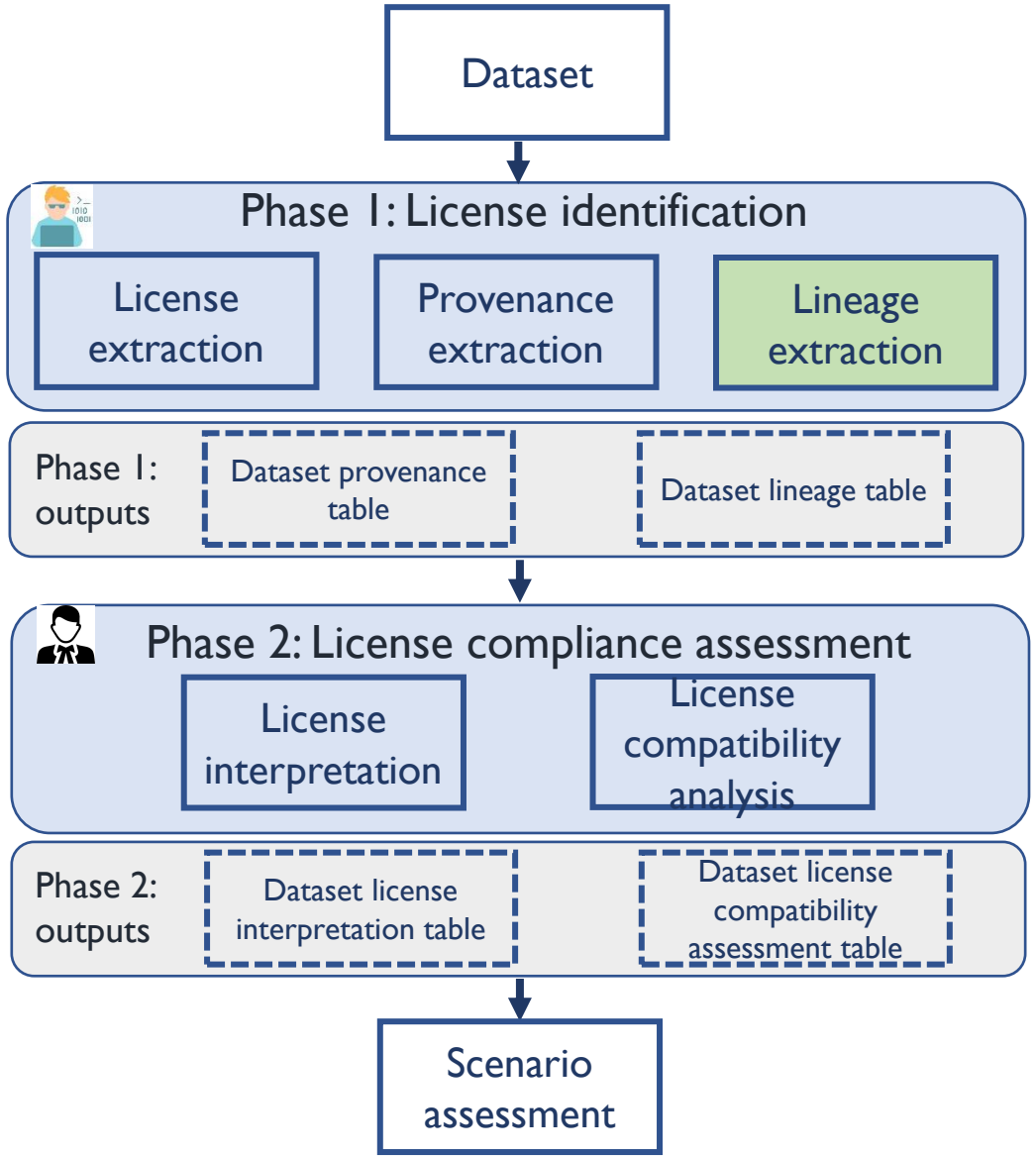
Our approach to assess the potential risks of using publicly available datasets in commercial AI software



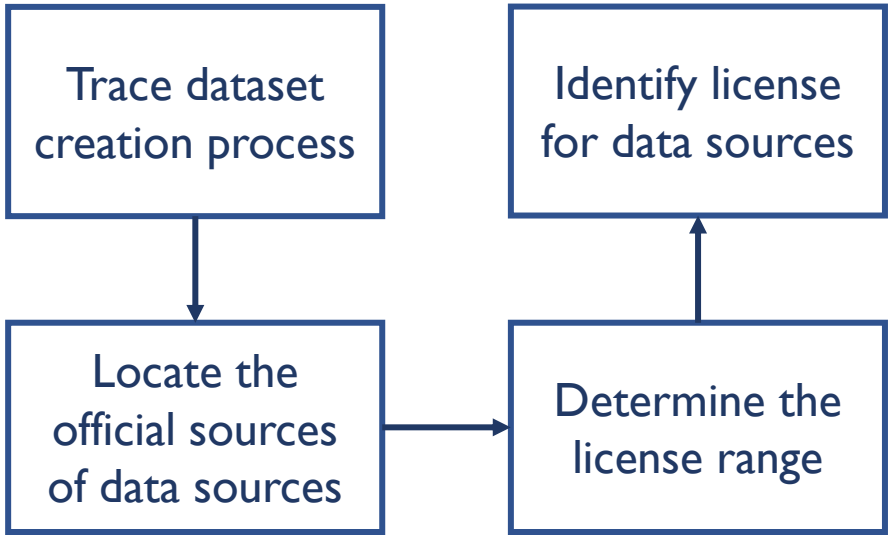
CIFAR-10's dataset provenance table

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

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

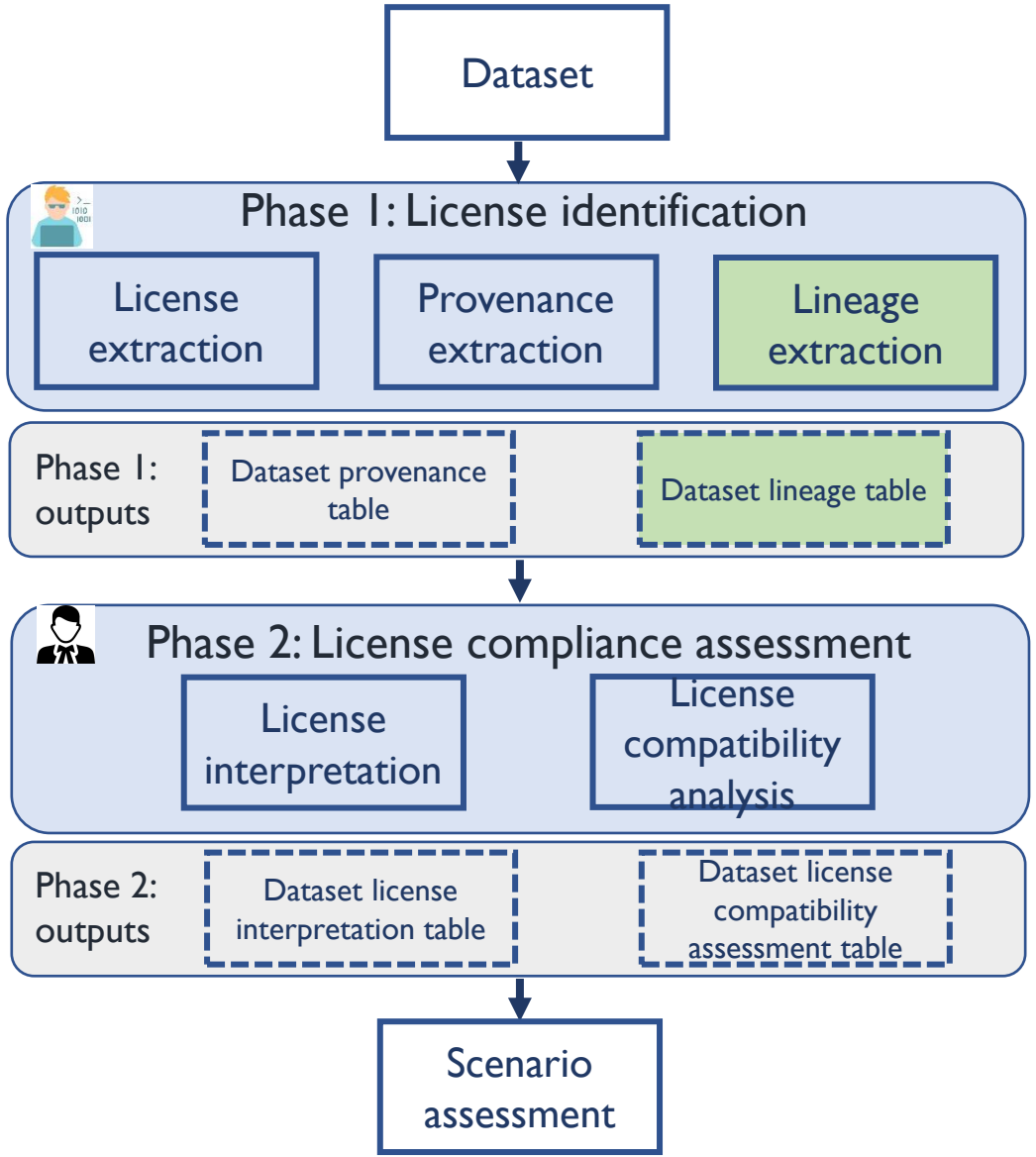


Lineage extraction sub-steps

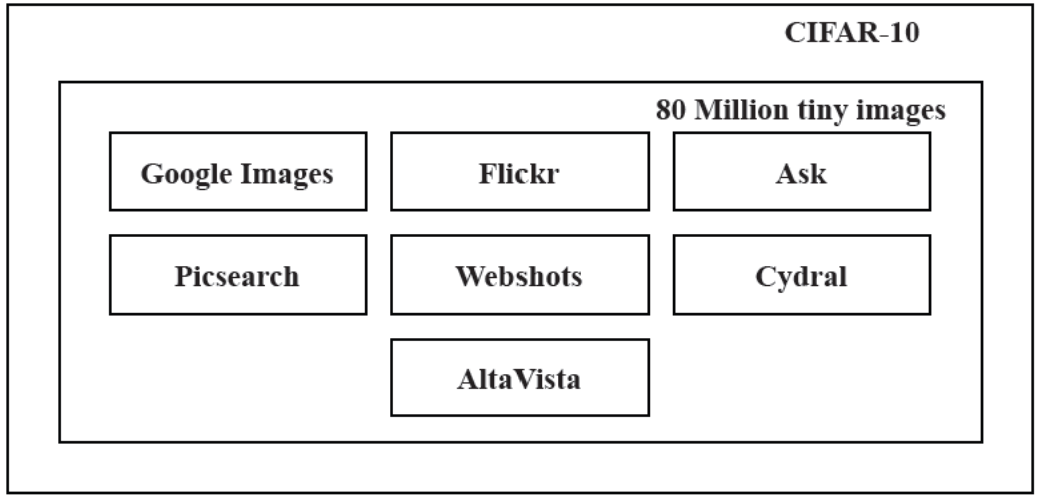


Lineage extraction step helps us mitigate unknown-data problem

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

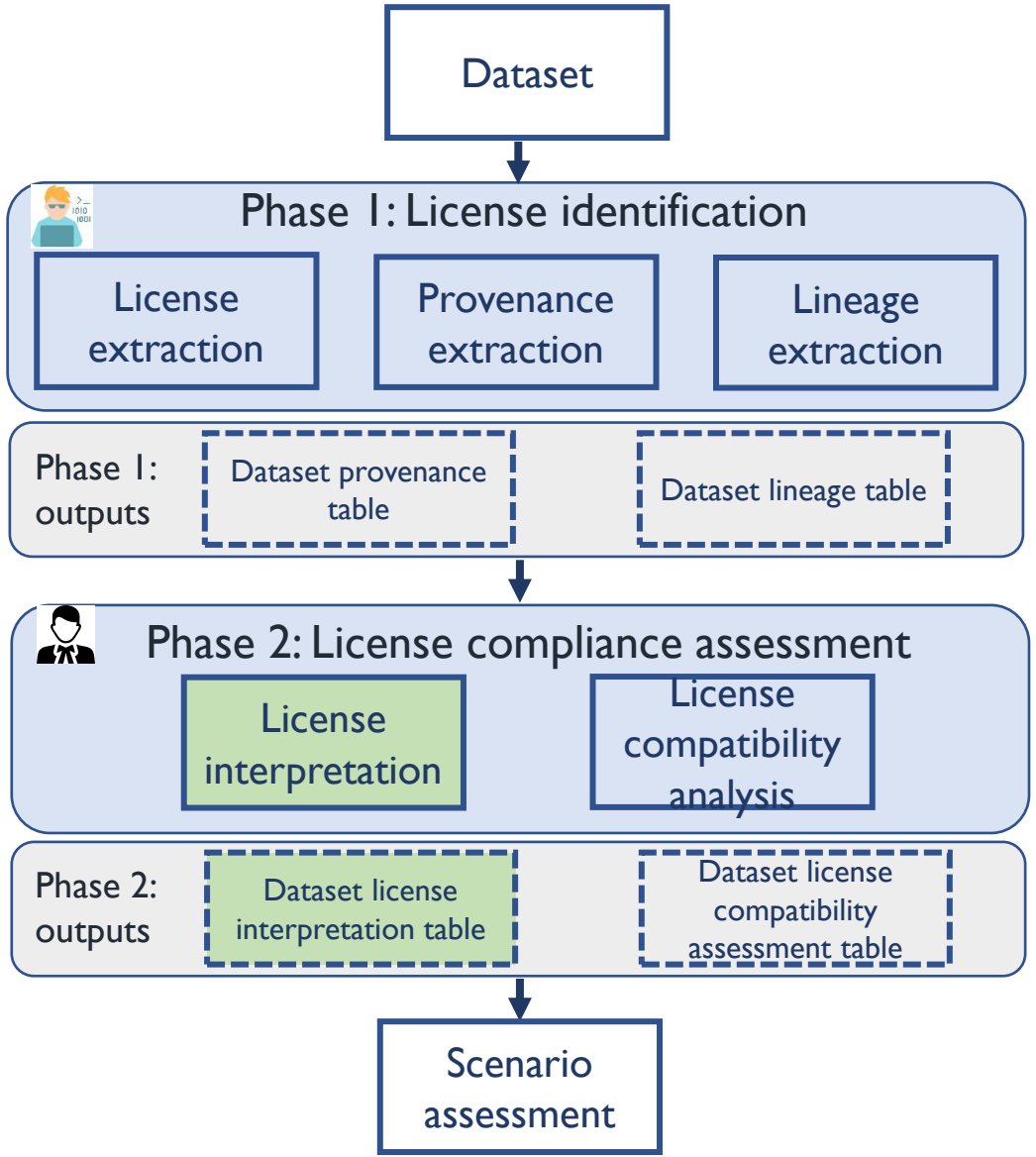


CIFAR-10's dataset lineage table



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



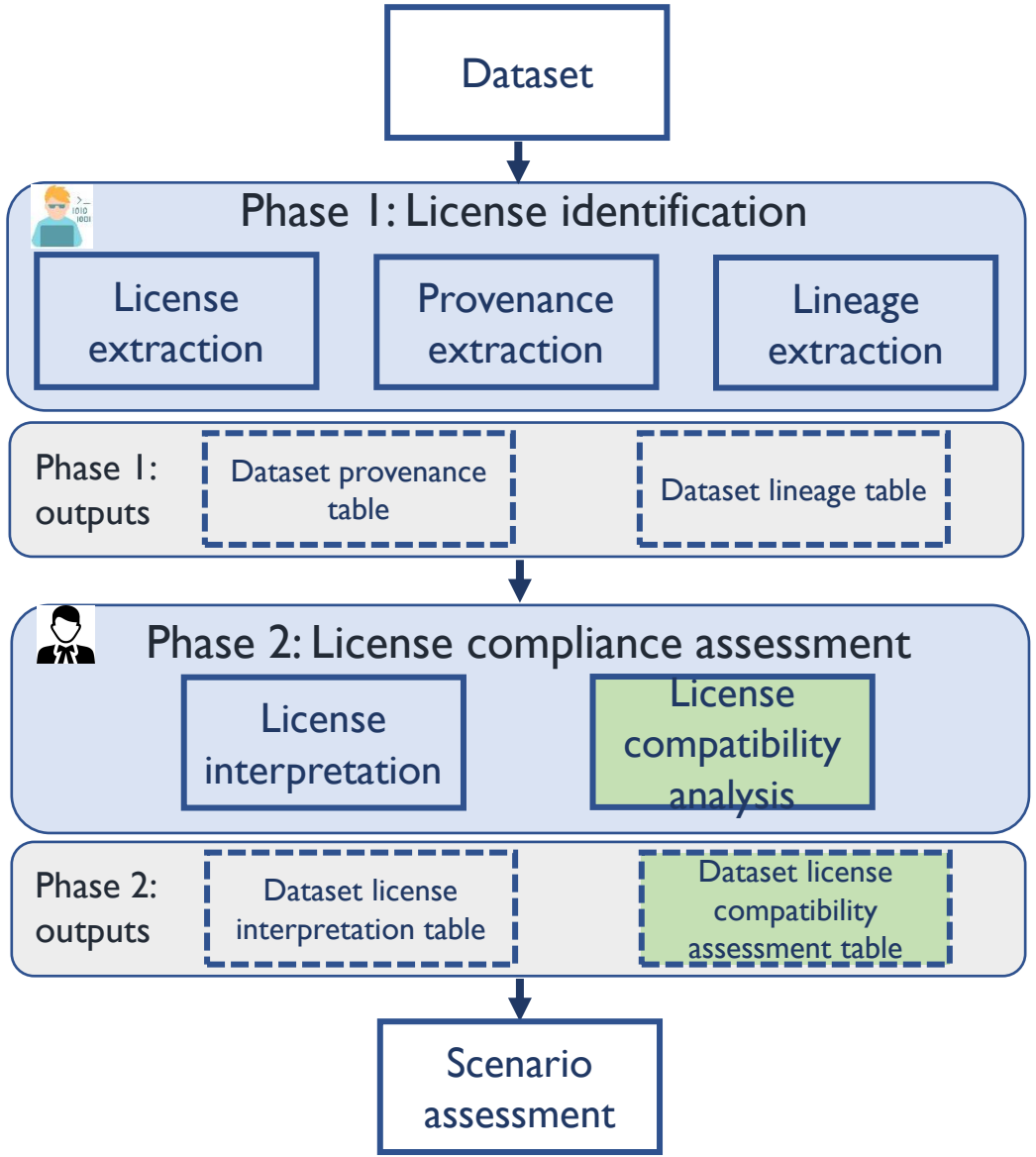
CIFAR-10's dataset license interpretation table (Based on enhanced Montreal Data License)

License metadata	Licensor	License name	Dataset name	Dataset version
	Alex Krizhevsky	Custom license	CIFAR-10	N/A
	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

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)

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

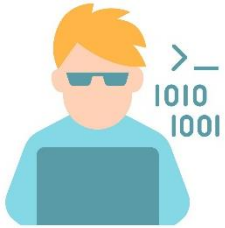
Our approach to assess the potential risks of using publicly available datasets in commercial AI software

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	✓	✓ (✗)	✓ (✗)	✓ (✗)
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

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

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

Our potential risk assessment results on studied publicly available datasets

	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)	✓	⊘	⊘

Recommendations

Employ caution while using publicly available datasets to build commercial AI software

To assess license compliance of datasets, use our systematic approach and clearly document all the results to demonstrate due diligence

Share knowledge regarding the risks associated with using a given publicly available dataset commercially

Request to community

We would like to create standards by working with LF-AI and its associated communities to create open standards to document various license compliance related information (e.g., provenance, lineage, rights and obligations associated with dataset licenses).

We would also like to work with LF-AI and its associated communities to standardize the framework to assess the potential risks associated with dataset license compliance issues.

We would also like to work with LF-AI and its associated communities to create tools and techniques to support and automate the aforementioned framework and enforce the standards.

Unlike OSS, conducting license compatibility analysis for datasets have several challenges

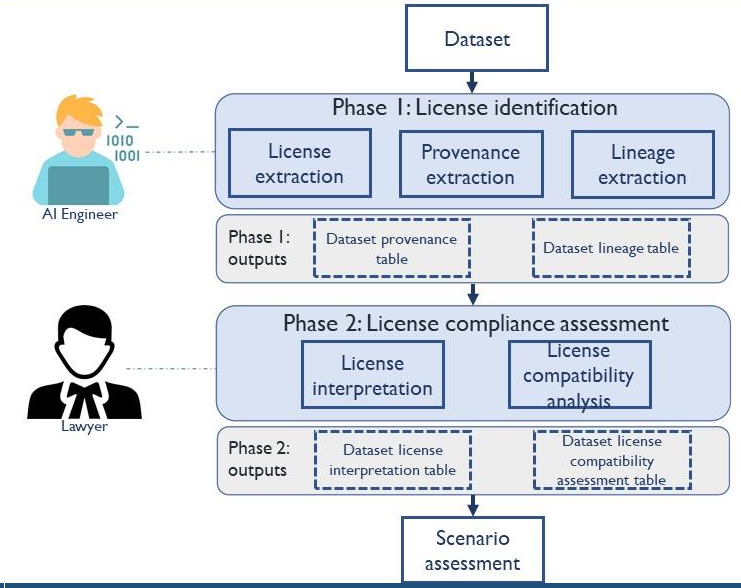
Unclear rights and obligations

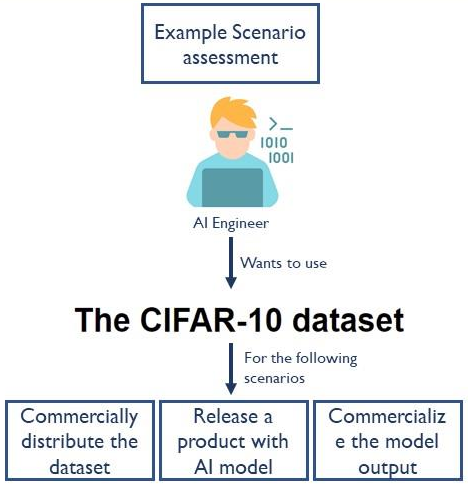
Unclear dataset origin

Non-standard license locations

Unclear data sources

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





Our potential risk assessment results on studied publicly available datasets

	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)	✓	✗	✗

Request to community

We would like to create standards by working with LF-AI and its associated communities to create open standards to document various license compliance related information (e.g., provenance, lineage, rights and obligations associated with dataset licenses).

We would also like to work with LF-AI and its associated communities to standardize the framework to assess the potential risks associated with dataset license compliance issues.

We would also like to work with LF-AI and its associated communities to create tools and techniques to support and automate the aforementioned framework and enforce the standards.

Unlike OSS, conducting license compatibility analysis for datasets have several challenges

Unclear rights and obligations

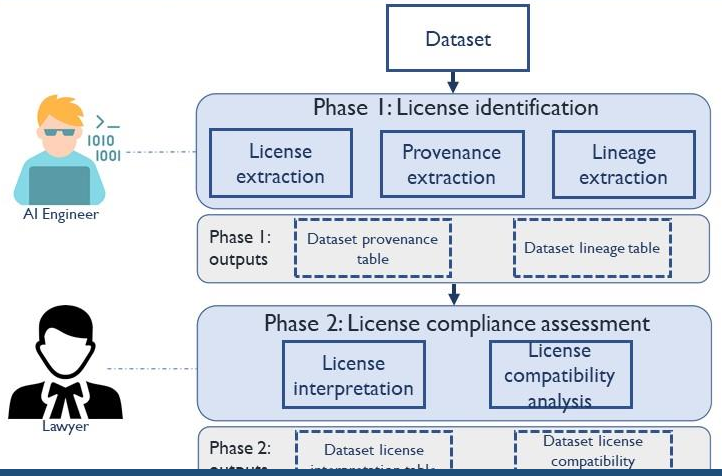
Unclear dataset origin

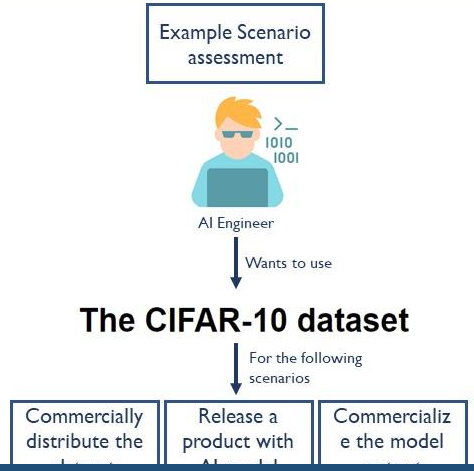
Non-standard license locations

Location not found

Unclear data sources

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





Gopi Krishnan Rajbahadur

gopikrishnanrajbahadur@gmail.com

@gopirajbahadur

	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)	✓	⊘	⊘

We would like to create standards by working with LF-AI and its associated communities to create open standards to document various license compliance related information (e.g., provenance, lineage, rights and obligations associated with dataset licenses).

We would also like to work with LF-AI and its associated communities to standardize the framework to assess the potential risks associated with dataset license compliance issues.

We would also like to work with LF-AI and its associated communities to create tools and techniques to support and automate the aforementioned framework and enforce the standards.

Upcoming TAC Meetings

LF AI & DATA

Upcoming TAC Meetings (Tentative)

- December 30, 2021: Canceled for the holiday
- January 13, 2021: Meetings resume, ART graduation proposal

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

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

LF AI & DATA