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


Gopi Krishnan Rajbahadur

SPDX Ambassdor, Co-Lead and Maintainer, SPDX AI and Dataset Profile

gopikrishnanrajbahadur@gmail.com

@gopirajbahadur



This work would not have been possible without the contributions from Erika Tuck, Li Zi, Zhang Wei, 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

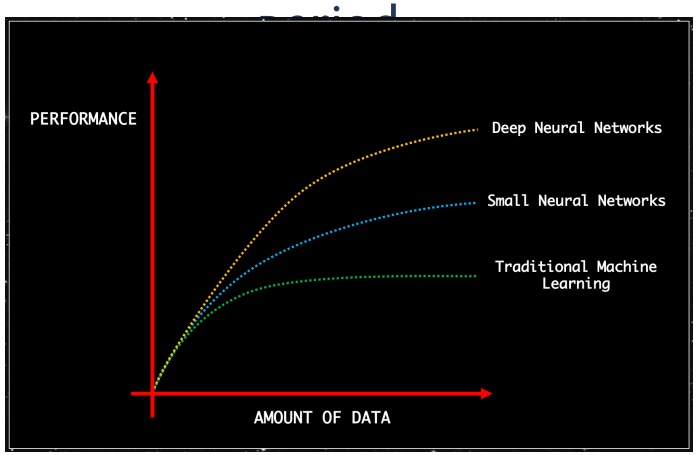
**A Cartel of Influential Datasets Is Dominating Machine Learning Research, New Study Suggests**

Harvard Business Review

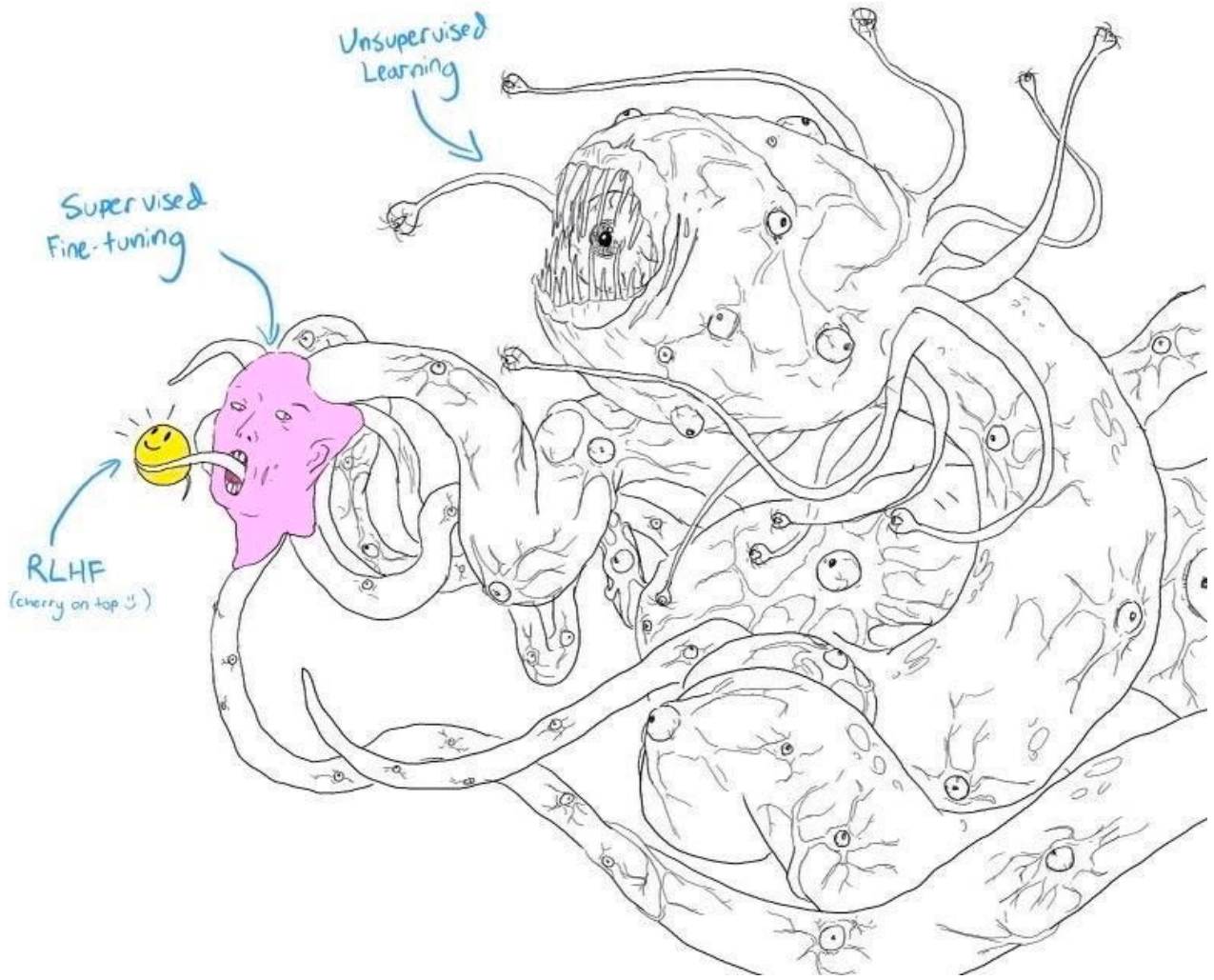
**Small Data Can Play a Big Role in AI**

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



# Large Language Models are Data Hungry beasts




https://lastweekin.ai/p/dont-plan-for-agi-yet

# The pile dataset is used to create a variety of LLMs

**The Pile: An 800GB Dataset of Diverse Text for Language Modeling**


**Leo Gao** **Stella Biderman** **Sid Black** **Laurence Golding**

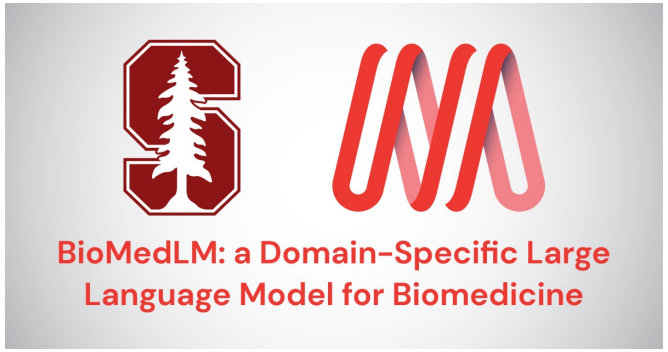
**Travis Hoppe** **Charles Foster** **Jason Phang** **Horace He**

**Anish Thite** **Noa Nabeshima** **Shawn Presser** **Connor Leahy**

EleutherAI
contact@eleuther.ai


GPT Neo
1.3B

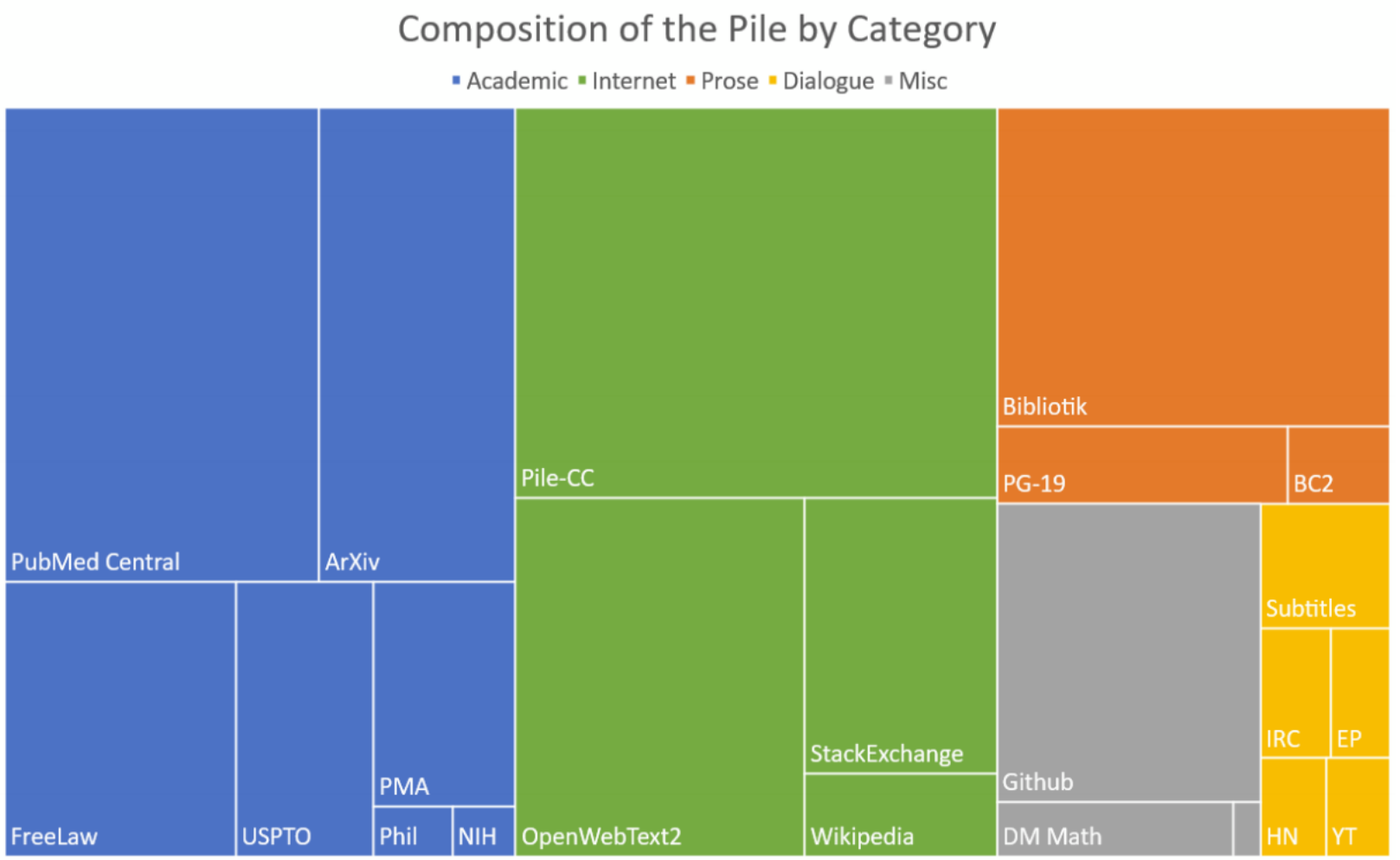
AIX

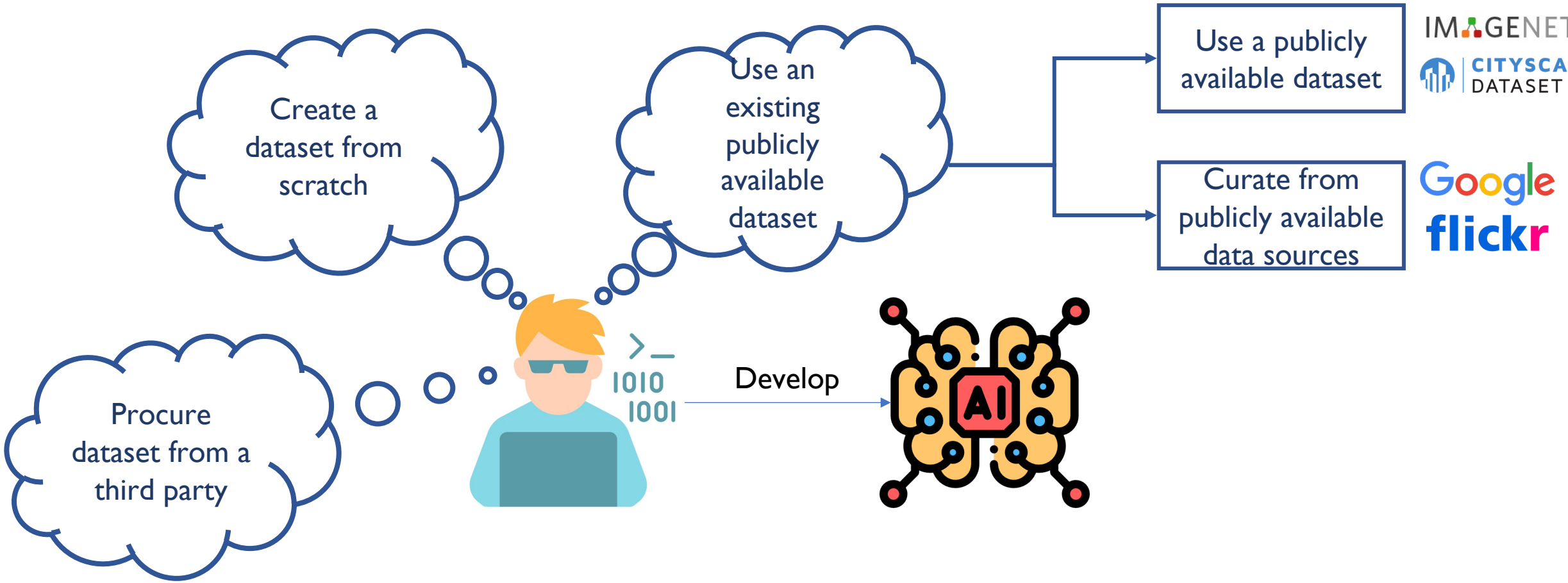

THUDM/**Chinese-Transformer-XL**

1 Contributor · 7 Issues · 209 Stars · 37 Forks

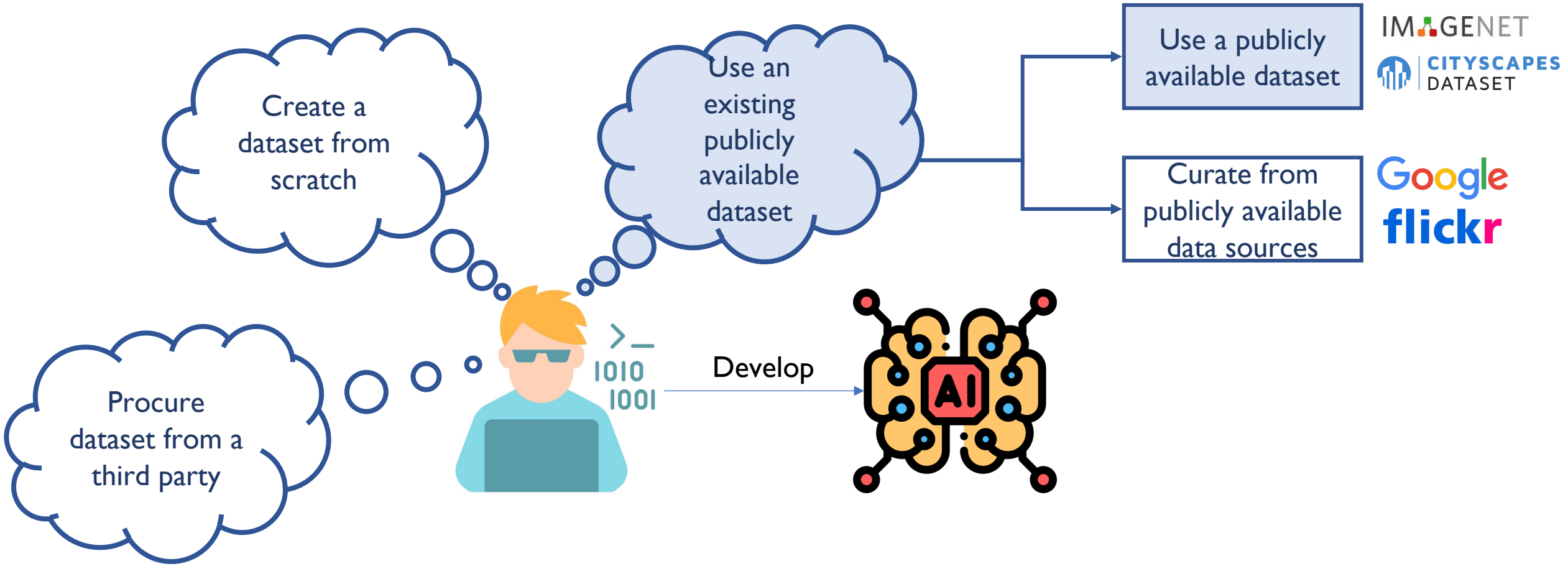
# Each of the dataset may contain a different license associated with them



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

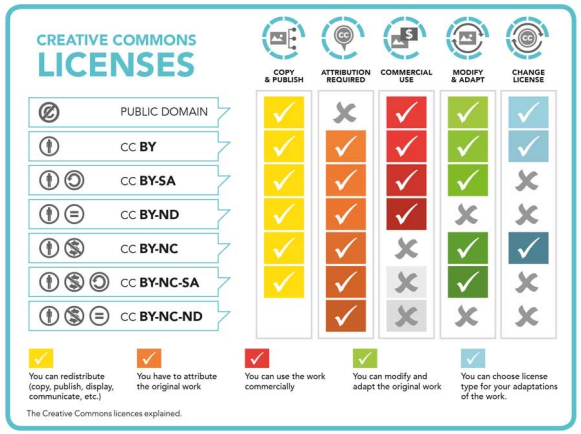


# There are several ways of acquiring the data required to build AI software



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


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.

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

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.

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

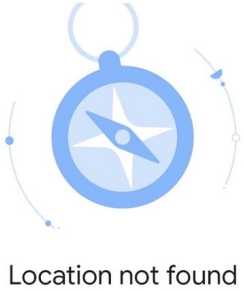
We loosely define the term dataset license. Unlike OSS, most datasets don't have a definitive license rather they outline terms of use, agreements. For the purposes of this talk, we call them license

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

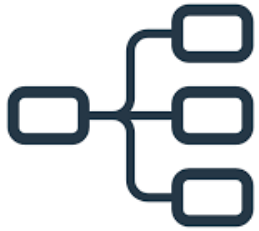
# Unlike OSS, conducting license compatibility analysis for datasets have several challenges

Unclear rights and obligations

Unclear dataset origin



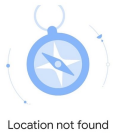

Non-standard license locations



Unclear data sources

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

Unclear rights and obligations




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.

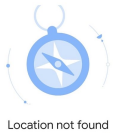
No clear mention if the dataset can be used for commercial purposes

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

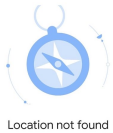
IMAGENET

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**

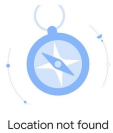
PyTorch Keras kaggle GitHub

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




Non-standard license locations

Unclear dataset origin

Unclear data sources

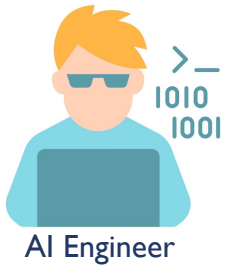

**The CIFAR-10 dataset**

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




AI Engineer

Lawyer

Dataset

**Phase 1: License identification**

- License extraction
- Provenance extraction
- Lineage extraction

Phase 1: outputs

- Dataset provenance table
- Dataset lineage table

**Phase 2: License compliance assessment**

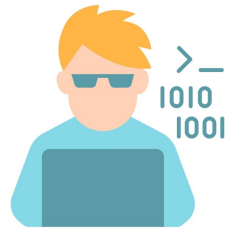
- License interpretation
- License compatibility analysis

Phase 2: outputs

- Dataset license interpretation table
- Dataset license compatibility assessment table

Scenario assessment

Example Scenario assessment




AI Engineer

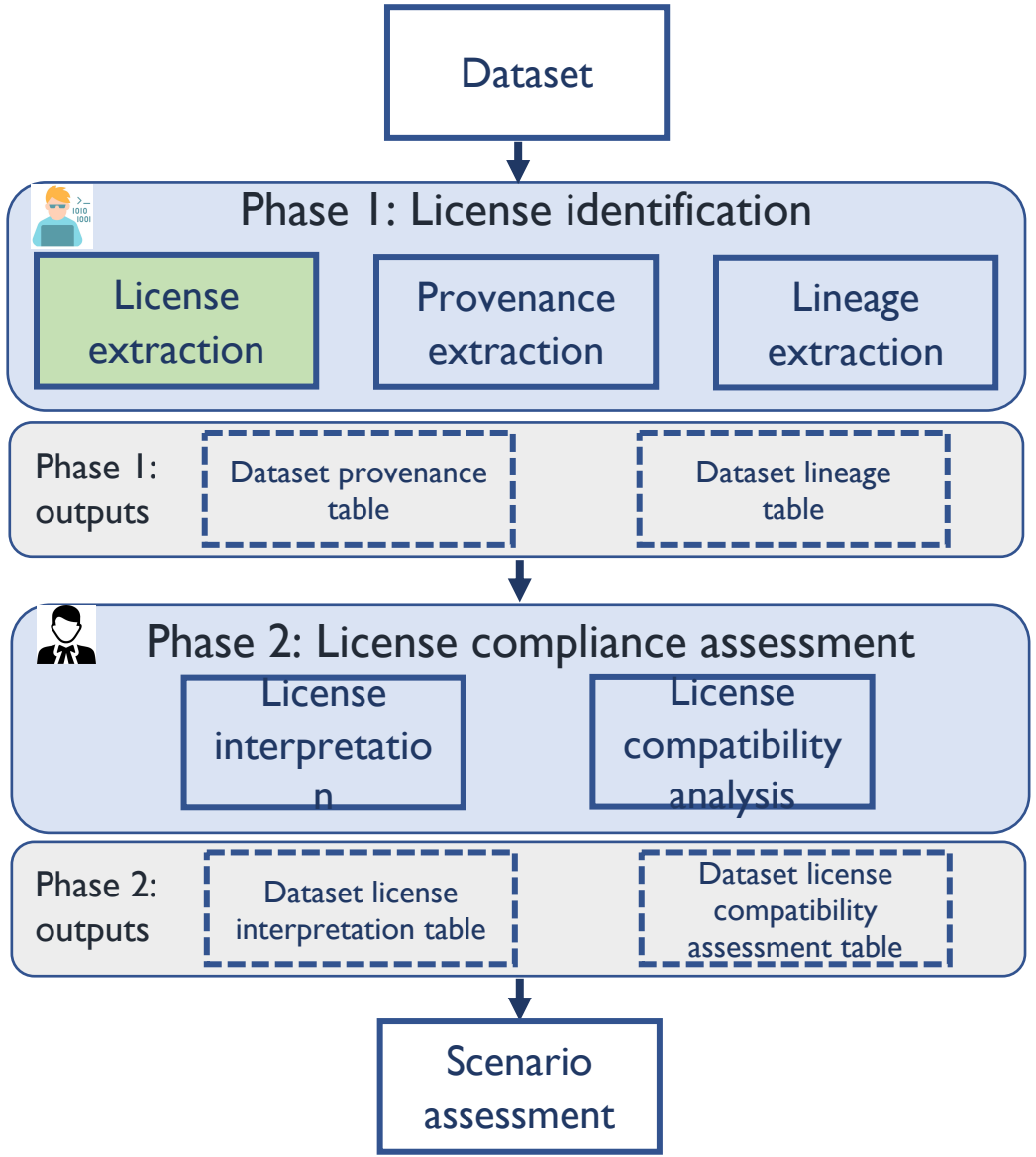
Wants to use

**The CIFAR-10 dataset**

For the following scenarios

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

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

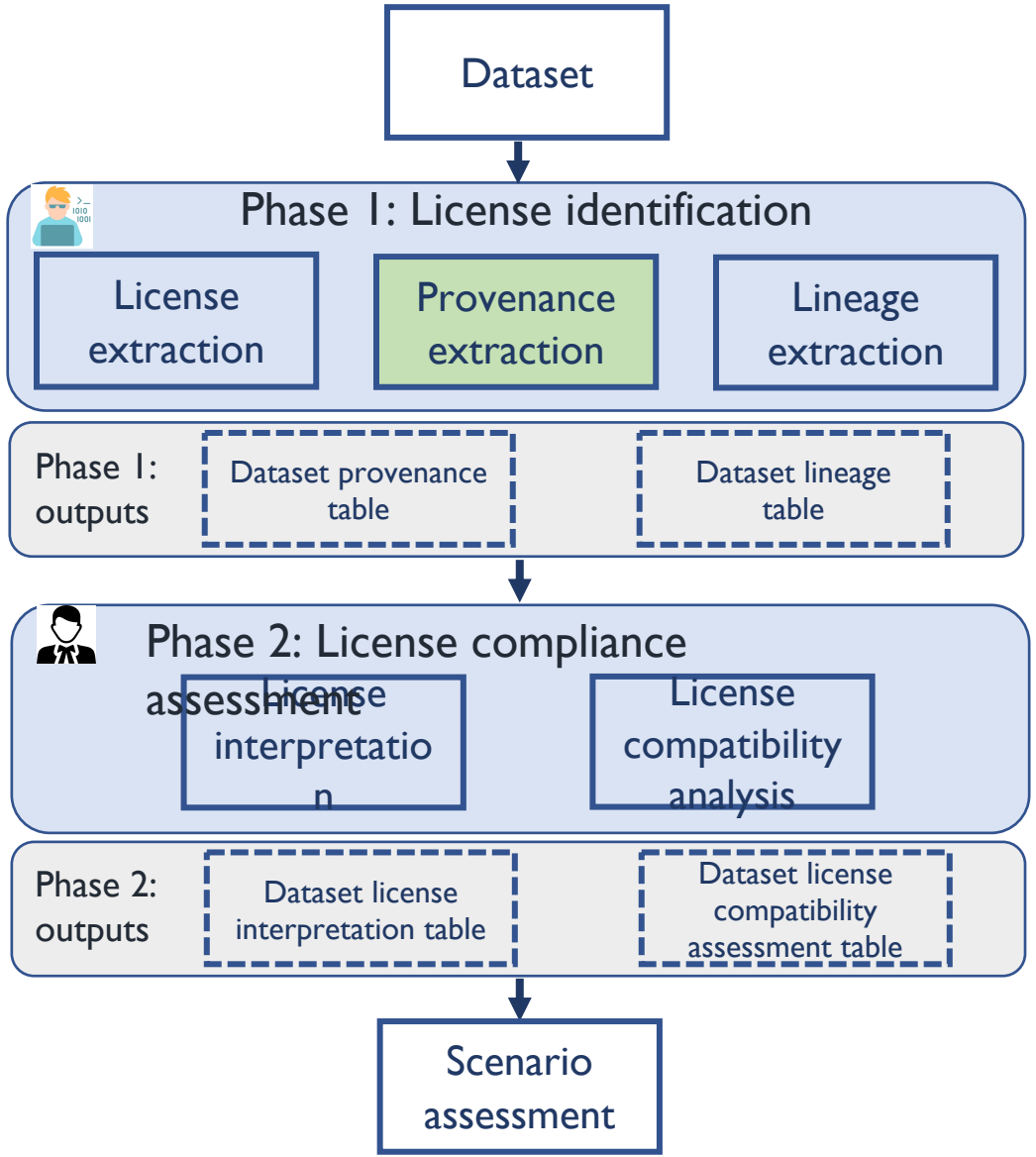



## CIFAR-10 License (available on official website)

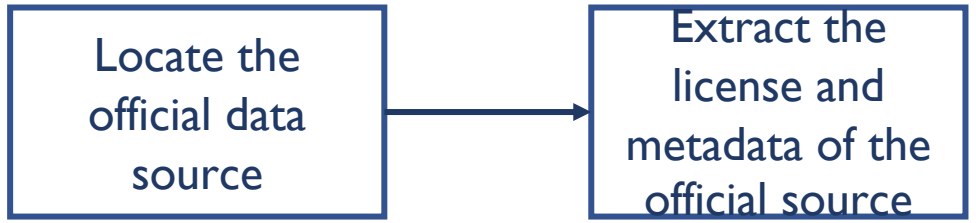
Please cite it if you intend to use this dataset.

- 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

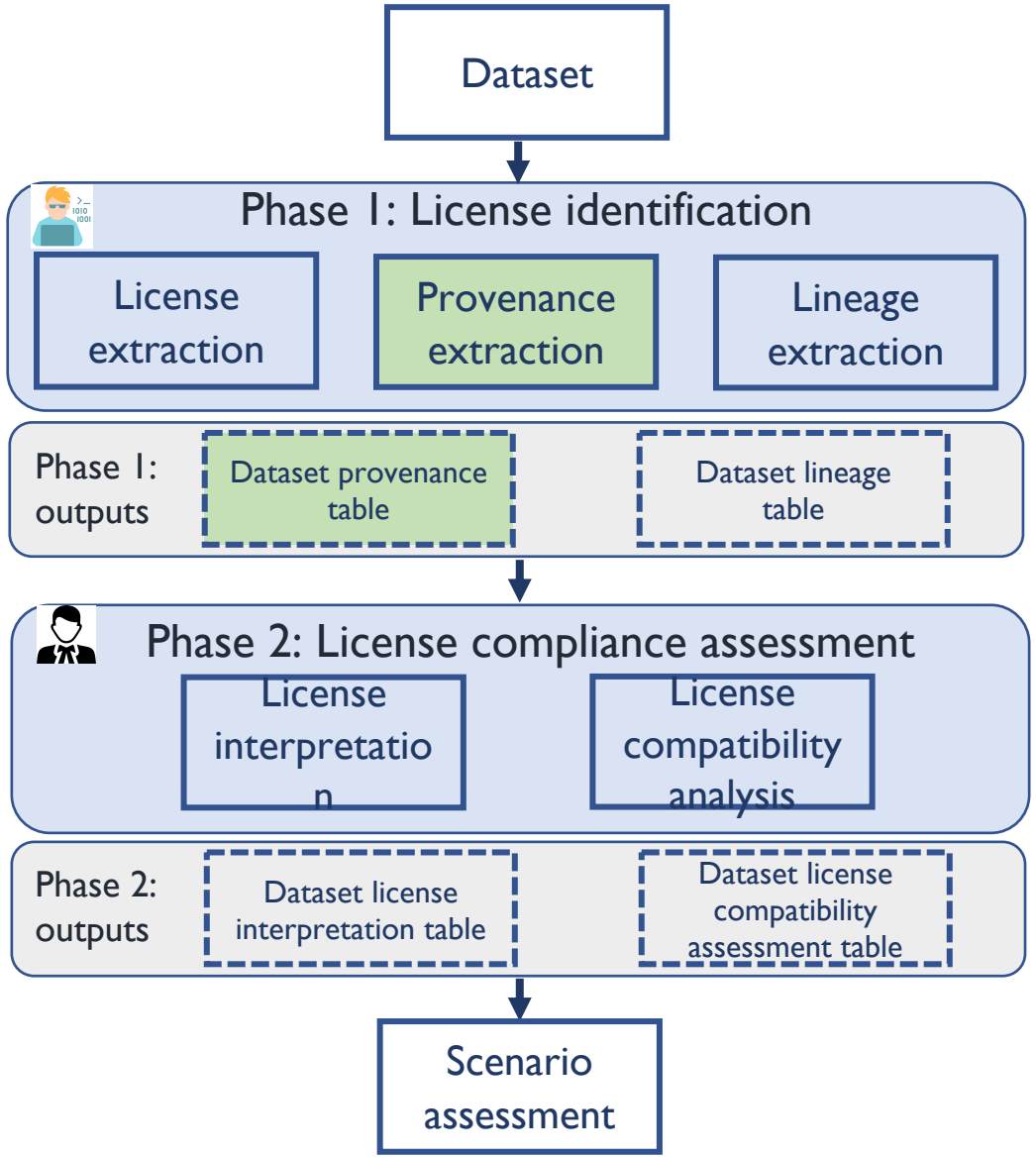



## Provenance extraction sub-steps




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

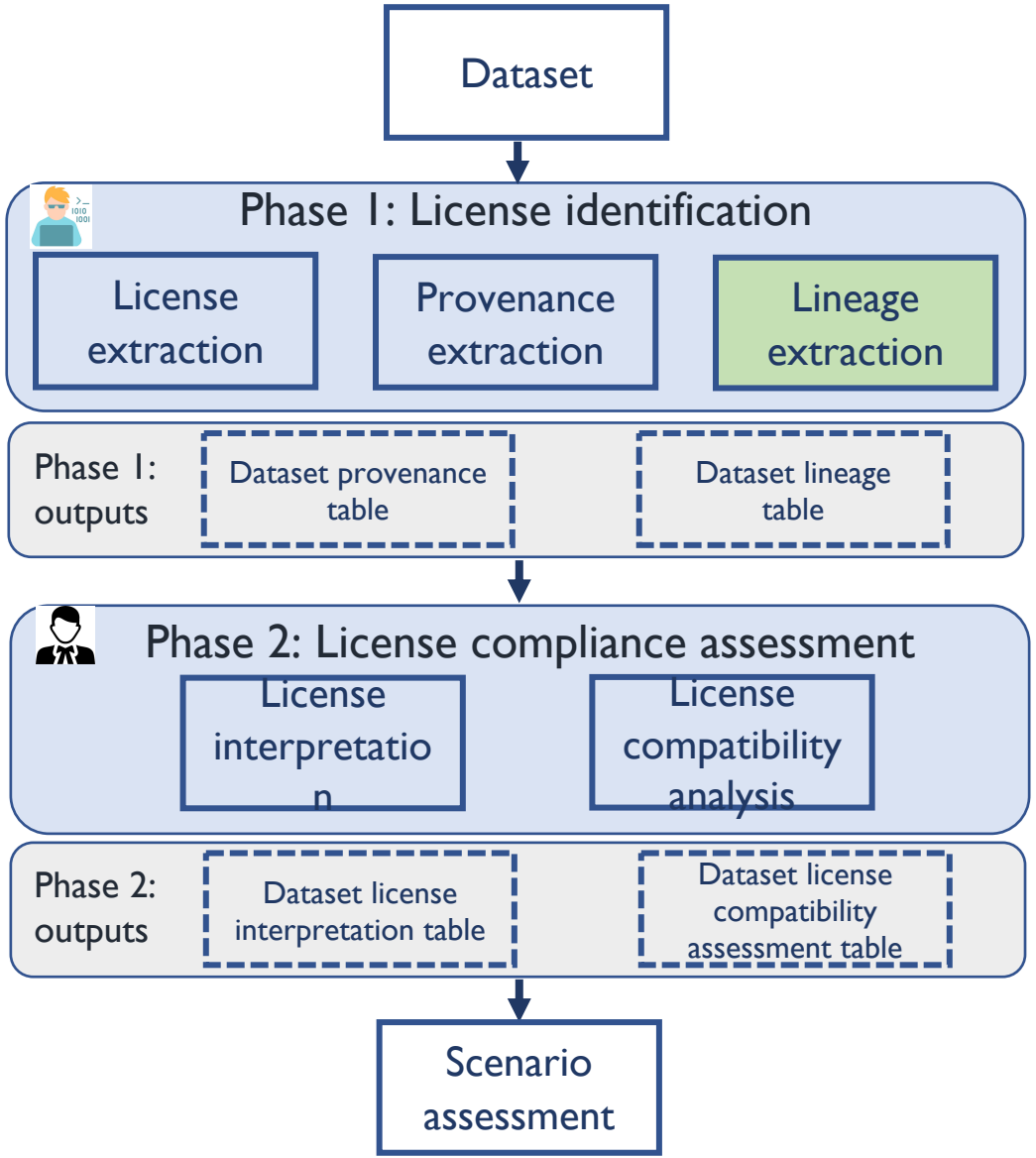
# Our approach to assess the potential risks of using publicly available datasets in commercial AI software



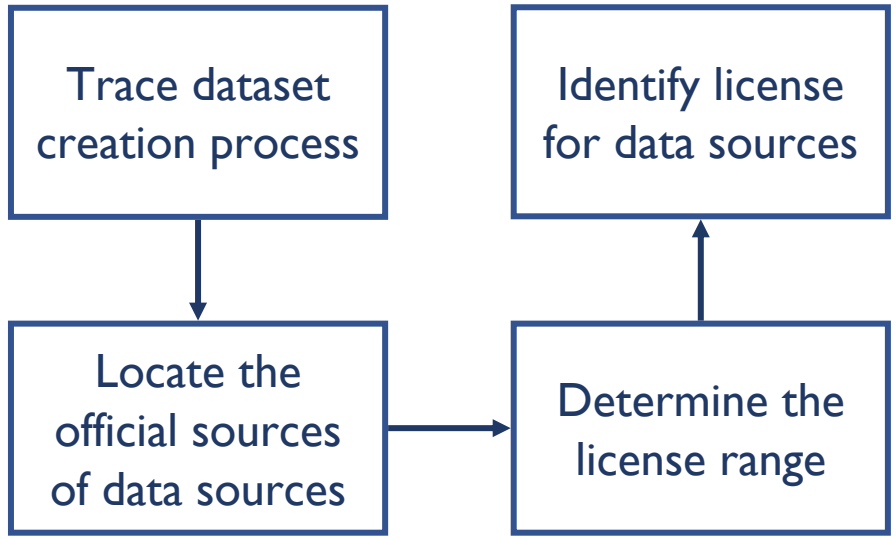

## CIFAR-10's dataset provenance table

| | | | | |
|---|---|---|---|---|
| **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

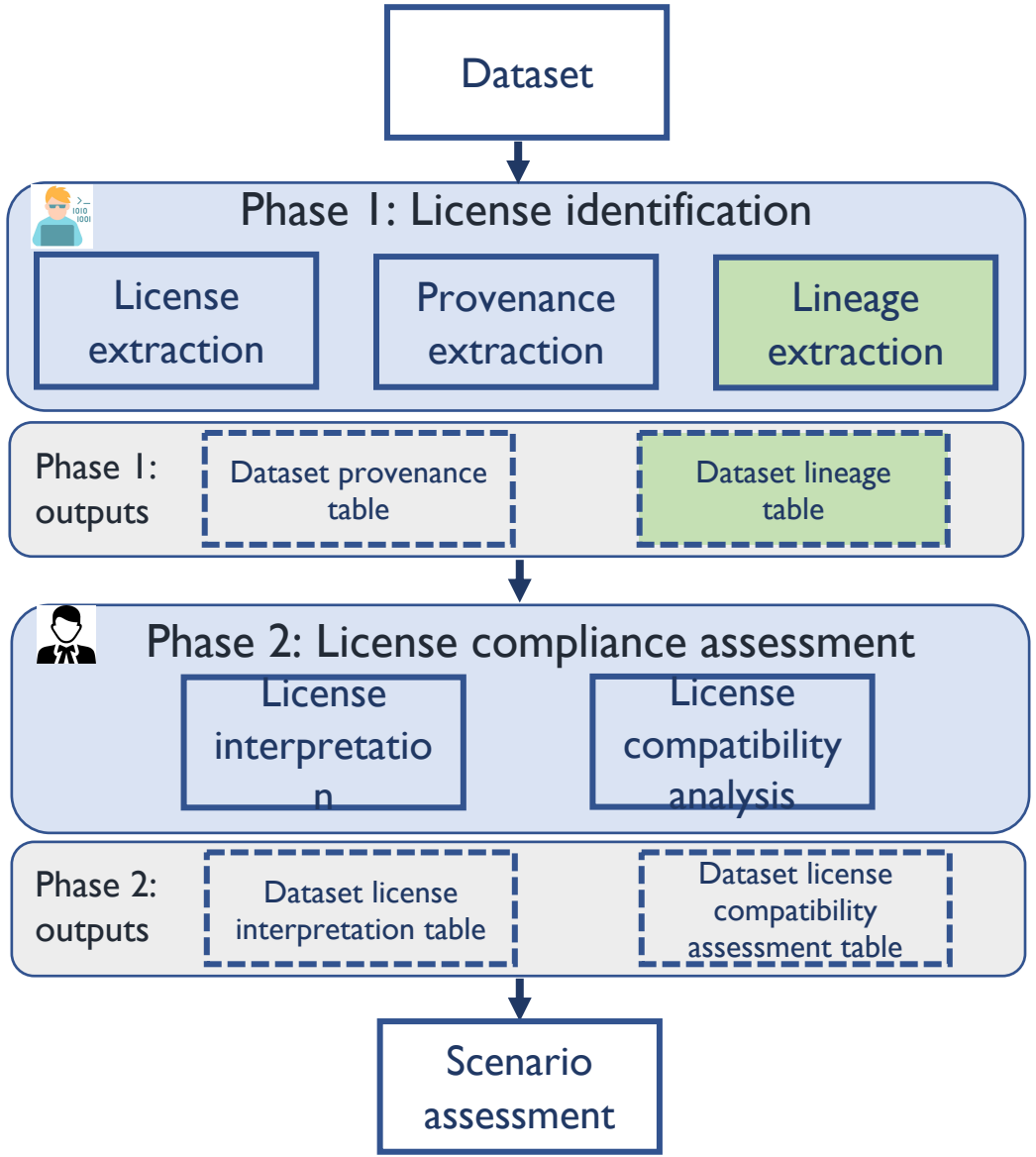



## Lineage extraction sub-steps

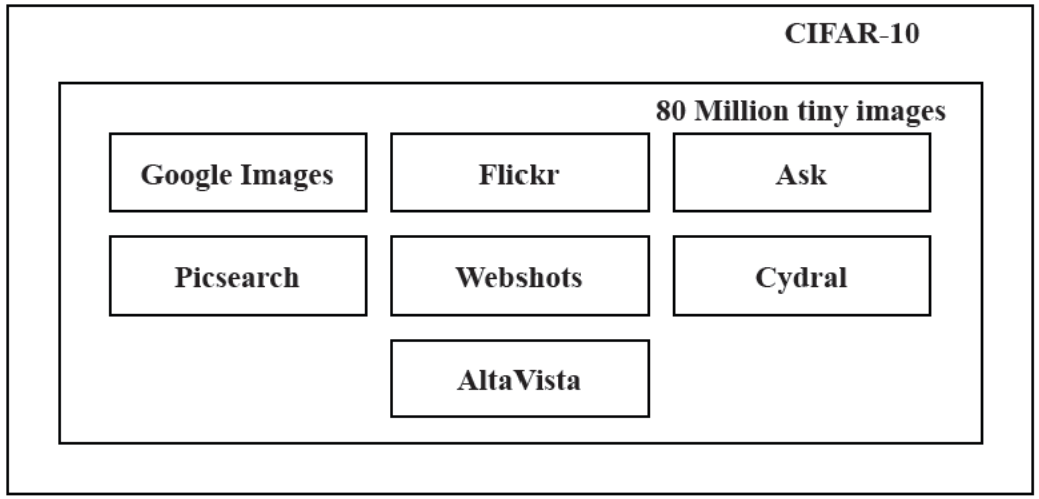



Lineage extraction step helps us mitigate unknown-data problem

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

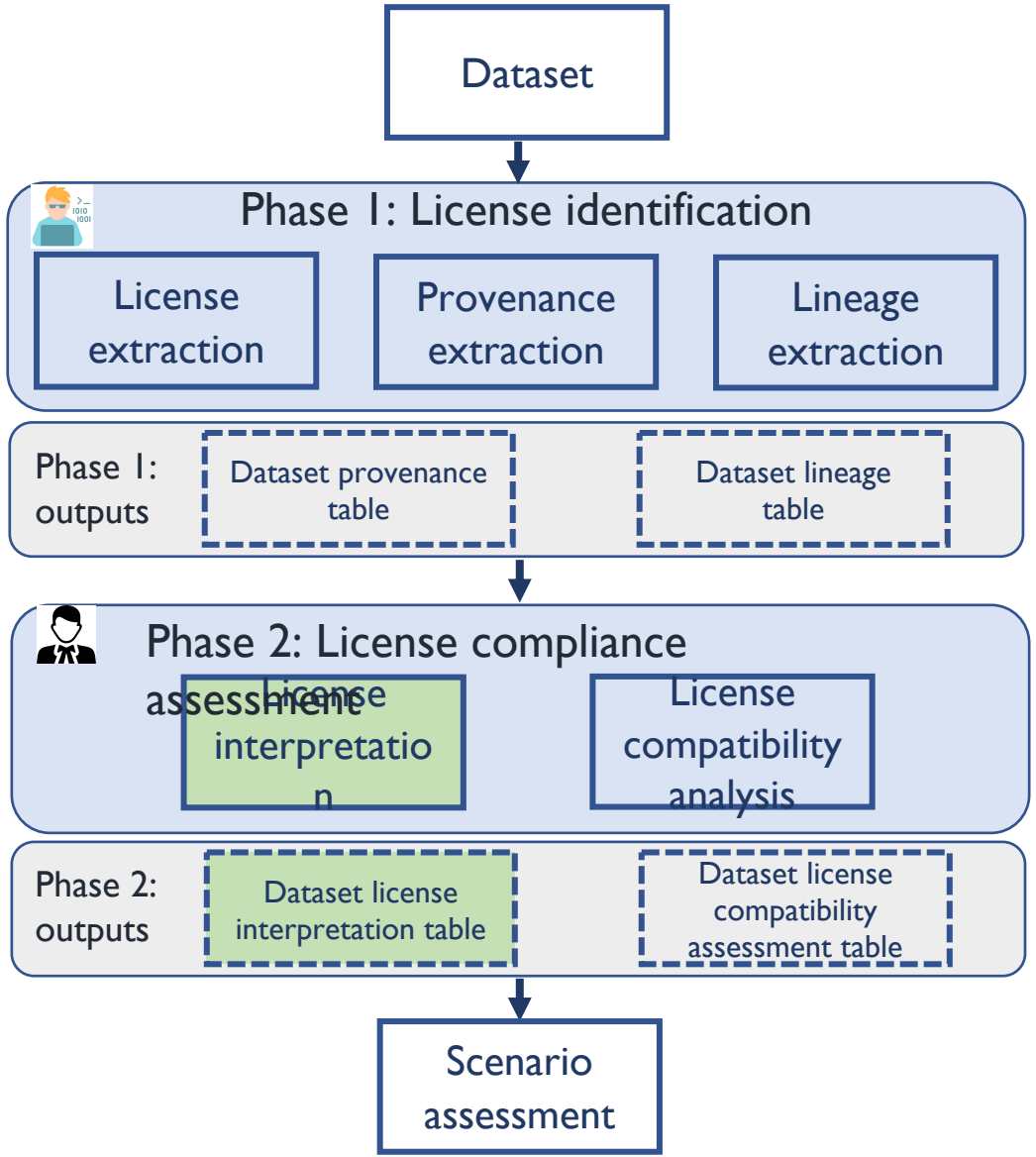



## CIFAR-10's dataset lineage table




Provenance details are recorded for each of the data source

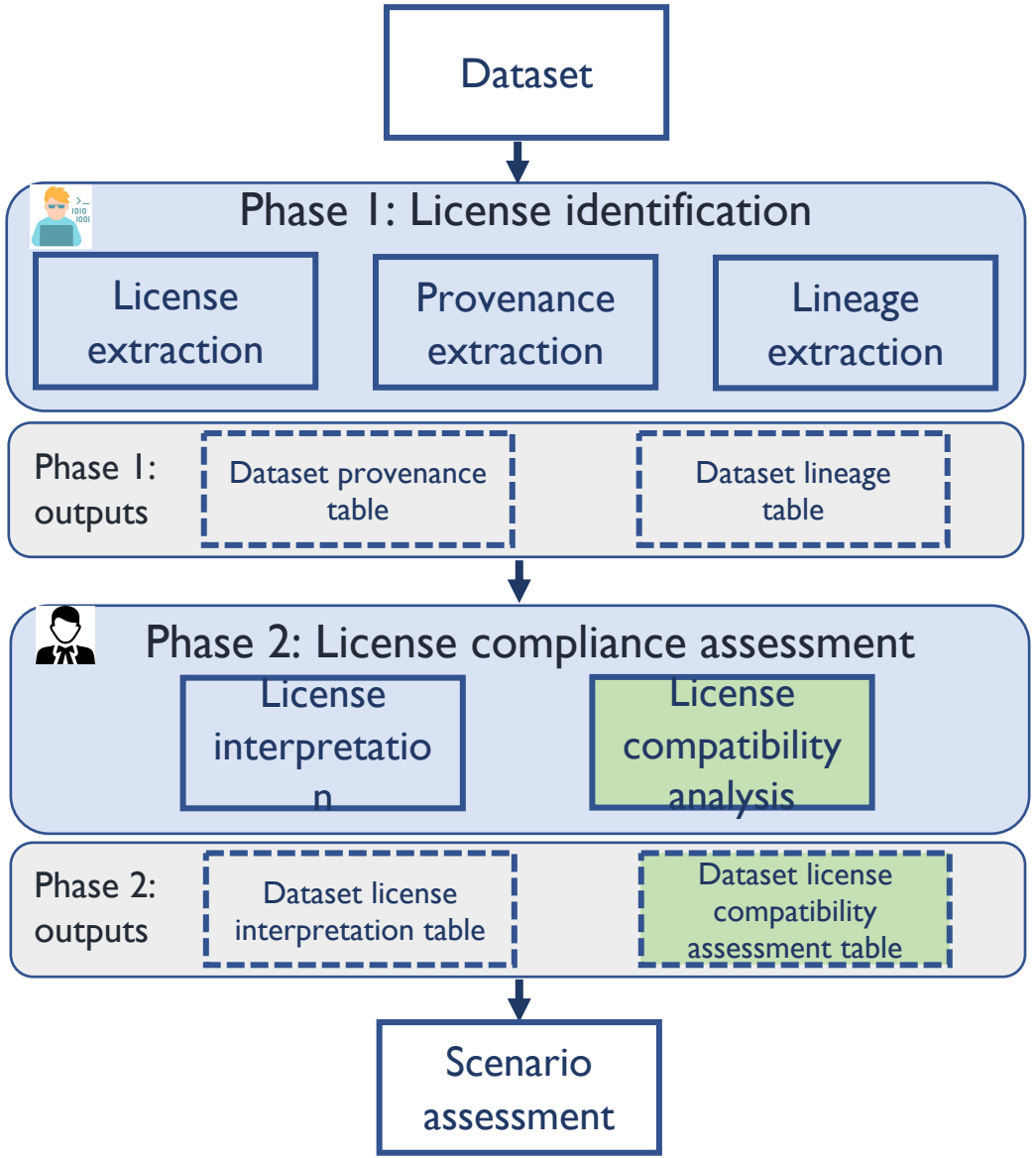
# Our approach to assess the potential risks of using publicly available datasets in commercial AI software




## 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** | | **Model Reverse Engineer** |
| | | | | | **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



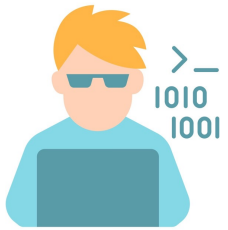

## CIFAR-10's dataset license compatibility table (Based on analyzing the license of all data sources)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **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** | **Bench-mark** | **Re-search** | **Publish** | **In-ternal Use** | **Commercialization** | | **Model Reverse Engineer** |
| | | | | | **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** | | **Model Reverse Engineer** | |
| | | | | | **Out-put** | **Model** | | |
| **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 SPDX 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 through OpenDataology and its associated communities to standardize the framework to assess the potential risks associated with dataset license compliance issues.

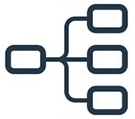
We would also like to work with LF-AI and OpenDataology and its associated communities to create tools and techniques to support and automate the aforementioned framework and enforce the standards.

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

Unclear rights and obligations

Unclear dataset origin

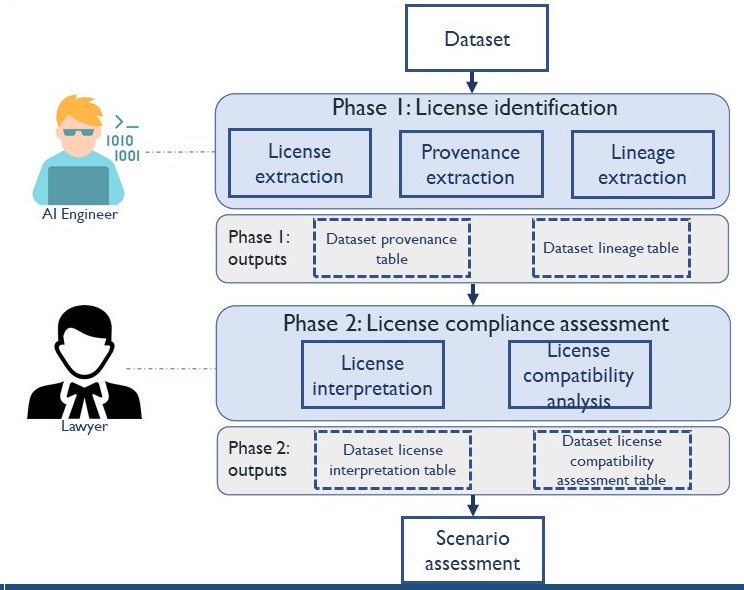

Non-standard license locations

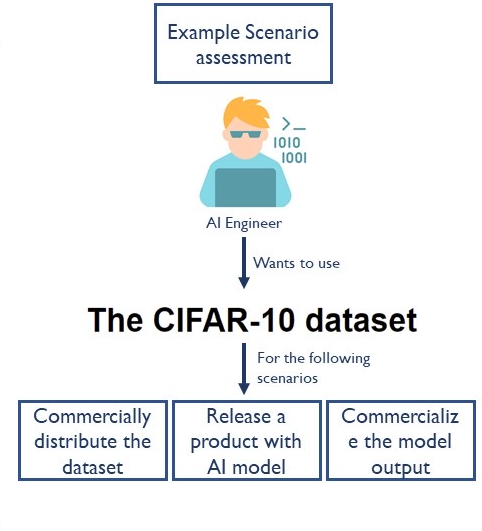


Unclear data sources



# 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.

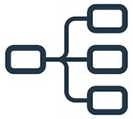
We would also like to work with LF-AI and its associated communities to create tools and techniques to support and automate the aforementioned framework and enforce the standards.

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

Unclear rights and obligations

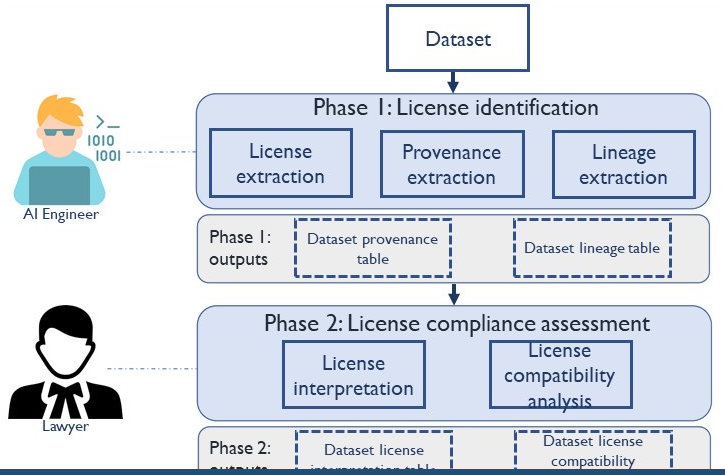
Unclear dataset origin

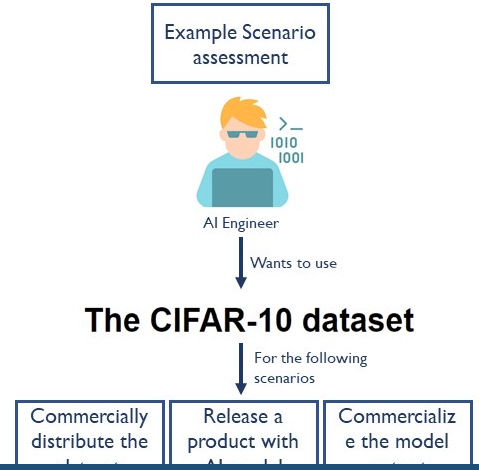

Non-standard license locations



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.

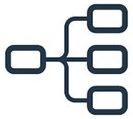
We would also like to work with LF-AI and its associated communities to create tools and techniques to support and automate the aforementioned framework and enforce the standards.

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

Unclear rights and obligations

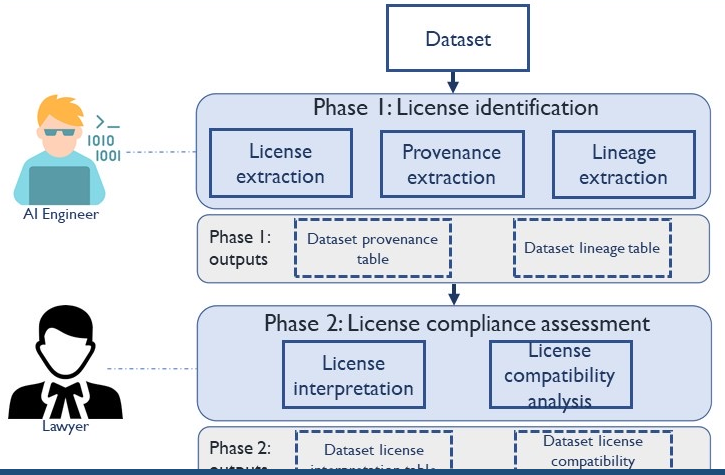
Unclear dataset origin

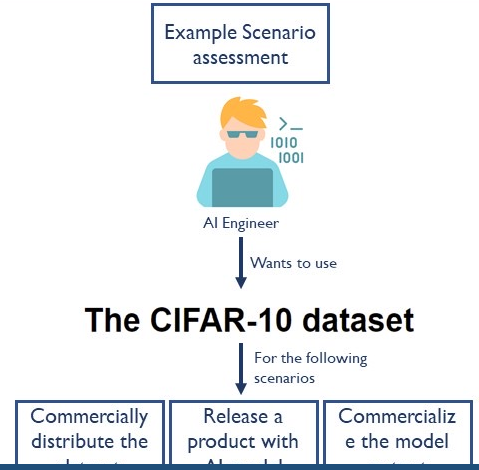

Non-standard license locations



Unclear data sources

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





## 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.