

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, China

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

Publicly available datasets are one of the key drivers for commercial AI software. The use of publicly available datasets is governed by dataset licenses. These dataset licenses outline the rights one is entitled to on a given dataset and the obligations that one must fulfil to enjoy such rights without any license compliance violations. Unlike standardized **Open Source Software (OSS) licenses**, existing dataset licenses are defined in an ad-hoc manner and do not clearly outline the rights and obligations associated with their usage. Further, a public dataset may be hosted in multiple locations and created from **multiple data sources each of which may have different licenses**. Hence, existing approaches on **checking OSS license compliance** cannot be used. In this paper, we propose a new approach to assessing the potential license compliance violations if a given publicly available dataset were to be used for building commercial AI software. We conduct a case study with our approach on 6 commonly used publicly available image datasets. Our results show that there exists potential risks of license violations associated with all the studied datasets if they were used for commercial purposes.

CCS Concepts: • **Software and its engineering** → **Empirical software validation; Requirements analysis**.

Additional Key Words and Phrases: Dataset license, License compliance, AI software requirement

ACM Reference Format:

Gopi Krishnan Rajbahadur, Erika Tuck, Li Zi, Dayi Lin, Boyuan Chen, Zhen Ming (Jack) Jiang, and Daniel M. German. 2022. Can I use this publicly available dataset to build commercial AI software?-A Case Study on Publicly Available Image Datasets . *ACM Trans. Softw. Eng. Methodol.* , (August 2022), 19 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

The adoption and the commercialization of Artificial Intelligence (AI) in software has increased significantly in the past decade [45]. One of the key components of AI software is AI models, which

Authors' addresses: Gopi Krishnan Rajbahadur, gopi.krishnan.rajbahadur1@huawei.com, Centre for Software Excellence, Huawei Canada, ON, Canada; Erika Tuck, etuck@yorku.ca, Lassonde School of Engineering, York University, ON, Canada; Li Zi, lizi4@huawei.com, Huawei China, ON, China; Dayi Lin, dayi.lin@huawei.com, Centre for Software Excellence, Huawei Canada, ON, Canada; Boyuan Chen, boyuan.chen1@huawei.com, Centre for Software Excellence, Huawei Canada, ON, Canada; Zhen Ming (Jack) Jiang, zmjiang@cse.yorku.ca, Lassonde School of Engineering, York University, ON, Canada; Daniel M. German, dmg@uvic.ca, University of Victoria, ON, Canada.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.

1049-331X/2022/8-ART \$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

are trained from large swathes of data. Companies looking to build commercial AI software can acquire datasets through several approaches such as creating and curating data from scratch [13]; acquiring data from external sources such as research publications in the form of publicly available datasets [4]; crawling and curating datasets from publicly available data sources; and, purchasing datasets through dataset vendors [1]. Among these approaches, one of the most popular approaches is to use publicly available datasets to build commercial AI software [3, 23, 35, 40].

It is important to ensure that the usage of publicly available datasets, particularly for commercial purposes, remains legally compliant [4]. The usage of a publicly available dataset is governed by the dataset license associated with the dataset (much like Open Source Software (OSS)). Similar to OSS licenses, dataset licenses outline the *rights* that users are entitled to and the *obligations* that users have to fulfill in order to enjoy these rights. From an AI software development perspective, the rights outlined by a particular dataset license determines whether the dataset can be used in a specific usage context (e.g., for building commercial applications or to redistribute). The obligations under a dataset license can be considered as software requirements (e.g., redistributing the AI software under the same license) that need to be captured and traced throughout the development lifecycle of the AI software, if the dataset is used. Therefore, similar to the usage of OSS in commercial software, to ensure legally compliant commercial usage of a dataset, it is important to adhere to the rights and obligations outlined in the dataset's licence (i.e., dataset license compliance).

However ensuring dataset license compliance poses several unique challenges in contrast to OSS license compliance. First, it is difficult to locate and identify the complete and correct license(s) associated with a given dataset. For instance, as Peng et al. [40] state, many retracted datasets continue to be available due to lack of clarity in retraction. Such a case makes it difficult to identify and locate the latest license of a dataset. Second, verifying the validity of the license associated with a given dataset is challenging. Many publicly available datasets are created by combining data from multiple data sources (e.g., websites or other datasets). Each of these data sources may have different licenses (e.g., website's terms of use). However, the creators of publicly available datasets seldom document the different licenses associated with the different data sources that they use. In addition, they do not consider the impact of these different licenses on the license that they assign to the aggregated dataset (which might make the license potentially invalid or wrong). For instance, CIFAR-10 dataset's license grants all rights to users and the only obligation that the license outlines is a requirement for citation. However, CIFAR-10 is created by crawling images from different sources such as Google Images and Flickr which may have licenses that restrict the usage of the images in any commercial context. Therefore, considering only CIFAR-10's license to ensure compliance in such a case might be problematic. Finally, as prior studies note [4, 40], the licenses of publicly available datasets are typically ambiguous and do not clearly outline the rights and obligations that govern the usage of datasets. Such lack of clarity makes it difficult in practice to use the datasets to build commercial AI software without the risk of license violation. For instance, GitHub Copilot [20] is an AI assistant that suggests whole lines of code (or sometimes whole functions). It leverages a large AI model that was trained on billions of lines of source code that are hosted on GitHub. While usage of open source code as a computer program is well-governed by OSS licenses and are common among GitHub users, the right to use code as data, in particular in the context of training AI models for commercial purposes, is not clearly defined in the OSS licenses. Such ambiguity in licenses has led to wide legal debate on the compliance of GitHub Copilot [19, 25, 41].

In this paper, we propose an approach to extract and analyze the compliance of dataset licenses. Our approach can help AI engineers extract the rights, and the ensuing obligations under a license. AI engineers can leverage our analysis results to evaluate any potential risks associated with using such a dataset under their specific usage context. We showcase the usefulness of our approach

through a case study on 6 widely used publicly available image datasets (CIFAR-10, ImageNet, Cityscapes, MS COCO, FFHQ and VGGFace2). We found that although these datasets were used widely in many of prior research [14] and commercial endeavors [3, 23, 35, 40], all of our studied datasets might not be suitable to build commercial AI software due to a **high risk of potential license violations**. Our findings highlight that AI engineers need to be aware of potential license-related risks and problems to avoid legal issues. The main contributions of our paper are:

- (1) This is the first work that proposes an approach to **assisting** practitioners on **assessing the risks of** using publicly available datasets to build commercial AI software.
- (2) As part of our approach, we have proposed several data schema which serves as a standardized way for documenting various dataset license related information that was collected during our analysis. These schema facilitate easy sharing of the analysis results of various datasets (e.g., different AI usage contexts or interpretation of different dataset licenses) and streamline the adoption of our approach. We have provided a replication package with the schema and results of our analysis documented with these schema to facilitate the sharing and the verification of our results¹.
- (3) Through our case study, we present the **potential license compliance violations** that one may run into when using six publicly available datasets in a commercial context. Furthermore, based on our experience on analyzing license compliance for various publicly available datasets, we also provide recommendations for AI engineers who use publicly available datasets to build AI software.

The rest of the paper is organized as follows. Section 2 presents the relevant background and related work of our study. Section 3 presents the approach that we use to identify potential license compliance violations that may occur when using publicly available datasets to build AI software. Section 4 presents the setup and results of our industrial case study. Section 5 provides the feedback from the industry practitioners. Section 6 provides several recommendations for AI engineers in light of our findings. Section 7 documents threats to validity and Section 8 concludes our study.

2 BACKGROUND AND RELATED WORK

In this section, we describe the background details and prior work that is related to our study.

2.1 Legal rights and protections on datasets

In this sub-section, we highlight some of the legal protections that govern and protect the usage of datasets in a commercial context. A dataset and the data points contained in the dataset are typically governed by two important laws: **copyright law and contract law**. Though the actual terms of the protection offered by these laws vary across countries, these laws in general provide several protections that we details below. In addition, with the recent advent of AI, there have been several privacy and fairness related laws that also govern the usage of a dataset, which we detail later in this sub-section.

Copyright law. Copyright law prevents anyone from **copying or reproducing** any part of a work (e.g., an image, video) without explicit permission from the owner of the copyright [36]. Therefore, by default, copyright-protected data cannot be used commercially or distributed unless explicitly allowed by the copyright holder. However, many publicly available datasets are known to contain copyrighted data [4, 40]. Using them to build commercial AI software could potentially result in copyright infringement.

However, it is important to note that in **certain cases and countries** it might be acceptable to use copyright-protected data for various purposes including commercial purposes without explicit

¹<https://github.com/TheTuckStopsHere/Replication-package>

permission from the copyright holders. For instance, in the United States, as the recent lawsuit *Authors Guild v. Google* [50] suggests, when there is **no material damage** done to the copyright holders, under the **Fair Use doctrine** [39] it is acceptable to use copyrighted data for commercial purposes, even without explicit permission from the copyright holders. However, in other countries the same may not be true. In the United Kingdom and Canada, for instance, copyright-protected data can only be used for non-commercial purposes without explicit permission from the copyright holders under the fair dealing exception to copyright infringement [37]. Similarly, in the EU, under the **Text and Data Mining Law** [49], copyright-protected material can be used without explicit permission from the copyright holders only for non-commercial purposes. In other parts of the world, different copyright laws may restrict the usage of copyright-protected materials in a commercial context. In summary, though there are certain exceptions that allow one to use copyright-protected data in a commercial context, it usually varies on a case-by-case basis. Therefore, using publicly available datasets with copyright-protected data to build commercial AI software could result in **potential copyright infringement**.

Contract law. Contract law states that owner of the copyright of a work (e.g., images, videos) may grant a **license** which outlines the rights that others can enjoy and the obligations that they must fulfil to enjoy those rights. When the terms of the license are not respected, i.e., if a right that is not granted by the license is exercised on the data or if the obligations are not fulfilled, then the potential violation may amount to breach of contract or contract violation. Depending on the country where the dataset was created or is intended to be used, either copyright law or contract law takes precedence.

In addition to the aforementioned protections, many countries around the world have regulations that dictate how private information in datasets can be fairly leveraged [8, 9, 38]. Though all privacy, fairness, copyright and license related issues can cause compliance issues, in this paper we only focus on identifying the potential license compliance related issues that arise from using publicly available datasets. The potential risks that we assess does not necessarily constitute as legal risks. We simply propose an approach to identify the potential risks.

2.2 Objectives and challenges associated with analyzing dataset license compliance

The key objective for a company to analyze the license of a publicly available dataset is to understand and determine if the given dataset can be used **in a specific usage context** (e.g., model training, redistribution) commercially while ensuring license compliance. License compliance is a form of regulatory compliance. Ensuring regulatory compliance in AI software is pivotal across the globe [5, 24, 29, 55]. When an AI software uses a publicly available dataset, it is implicit that the owners of the AI software enter into an agreement with the copyright holders of the dataset. Such an agreement entails that the AI software uses the dataset only per the rights provided by the license and fulfils the obligations outlined in the license, which makes the obligations a software requirement [24]. If the AI software fails to do so, it will be in breach of the dataset's license, which becomes an issue of regulatory compliance. As several prior studies note, organizations that are found **non-compliant of laws and regulations face serious consequences** [5, 24, 29, 55]. Therefore, compliance is both a functional and non-functional requirement of an AI software [5].

However, as Benjamin et al. [4] outline, **licenses associated with publicly available datasets do not clearly state the rights and obligations associated with the usage of datasets**. Such a problem poses a challenge to AI engineers trying to encode obligations as requirements. Ensuring license compliance under such ambiguous licenses is challenging as we explain in Section 1. In cases such as those as Breaux et al. [5] point out, when requirements, in particular legally enforceable requirements, are ambiguous, it is important for AI engineers to do their *due diligence* or take *due care* to make sure their systems are compliant. In other words, AI engineers should make **reasonable**

efforts to make sure that AI software is not found in breach of licenses. To make such a claim, AI engineers should employ traceability from the rights and obligations in the license to software requirements [5, 29]. In addition, it is imperative for AI engineers to justify that their interpretation of the rights and obligations is consistent with how they implemented those requirements in the constructed AI software [5]. Therefore, a rigorous approach to ensure license compliance, like the one we propose in our paper that ascertains the rights and obligations associated with a publicly available dataset, is vital.

2.3 Related work

Assessing license compliance violations. Though there is no prior work that studies the license compliance violations that happen when publicly available datasets are used to build commercial AI software, there is a rich body of work that studies license compliance violations when reusing OSS. Several prior works focus on how to assess license inconsistencies (i.e., violating the license of the OSS component that is being reused) that occur in OSS [53, 54]. For instance, several studies focus on designing automated tools to assess licence compliance [16, 17, 27, 51, 56]. Despite there being several techniques to ensure OSS license compliance, they cannot be directly used to assess license compliance violations that happen when using publicly available datasets. As discussed previously, unlike OSS, publicly available datasets have unclear licenses and license validity issues. Furthermore, as German and Di Penta [16] show, to check license compliance it is important to identify a software component's provenance and lineage (and they outline methods to do so). However, identifying a dataset's provenance and lineage is not as straightforward. Therefore, an approach to assessing dataset license compliance violations will be quite different from that of OSS. This is the main focus of our paper.

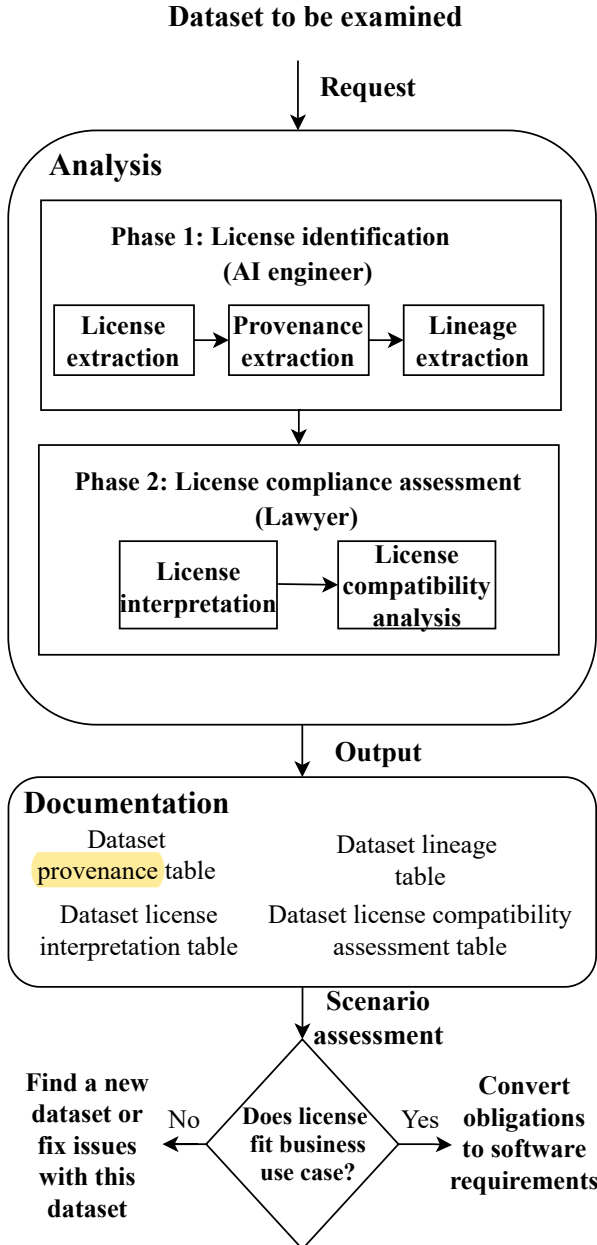
Software regulatory compliance. Similar to complying with the terms of a license, software also needs to comply with certain regulatory requirements. For instance, software built to handle information about patients in the United States needs to comply with HIPAA [15]. The rules outlined by such regulations are requirements (similar to rules outlined by a license) that must be followed by a software to which the given regulation is applicable [5]. There are several studies that focus on how to ensure that the requirements outlined by a regulation are followed by the software under consideration [5, 6, 29, 55]. The first step in checking if the software complies with the regulation is to identify the rights and obligations outlined by the regulation. Zeni et al. [55] propose a tool called GausiT to semi-automatically identify the rights and obligations of a regulation. Breaux et al. [5] provide an alternate method to extract the rights and obligations from a regulation to enable checking if the software is indeed compliant with the regulation. They showcase the usefulness of their approach by checking if a software is compliant with HIPAA. Unfortunately, we cannot use the methods outlined in these studies as these methods presuppose that a regulation clearly outlines the rights and obligations and that the regulations are valid. However, in our case, the licenses associated with publicly available datasets are typically unclear and have concerns regarding their validity.

Licenses for publicly available datasets. There are a few dataset specific licenses that are available, and these licenses can largely be grouped into copyleft and permissive licenses. The copyleft dataset licenses typically require the AI software using the dataset to be licensed under the same license under which the dataset was provided (or they do not clearly mention what license the software built using these datasets should have). Some examples of copyleft licenses are Community Data License Agreement (CLDA) sharing 1.0 and Creative commons (CC)-By-SA (Share Alike), CC-BY-NC-SA. In contrast, permissive license allow the AI software that uses these datasets to be licensed under any license one wishes. Example of permissive licenses are CC-BY (public domain), CDLA permissive. However, as Benjamin et al. [4] note even these licenses are not standardized and

their rights and obligations are typically ambiguous. In addition to showcasing the shortcomings of standard dataset licenses, they provide a new license called the **Montreal** Data License (MDL), which is more suitable for datasets. Different from their study [4], we focus on assessing potential license compliance violations if certain datasets are used in a commercial context.

3 APPROACH

Fig. 1. An overview of our approach.



In this section, we describe our approach for identifying whether using a publicly available dataset to build a commercial AI software would result in a license compliance violation. It starts with an AI Engineer initiating the request for license compliance analysis for a particular dataset. The license compliance analysis approach consists of two phases: Phase 1, called *License identification*, involves gathering the information we need about a dataset's license, provenance and lineage. Phase 2, called *License compliance assessment*, involves interpreting the output of Phase 1 and then completing a risk assessment based on the findings. Phase 1 can be done by an AI engineer; however, **Phase 2 must be done by a lawyer**, as interpreting the information contained in the collected licensing information requires legal expertise [16, 17, 22]. Over the course of our approach, various information (e.g., a dataset's provenance, lineage) are documented in a database. Once the analysis is complete, the AI Engineer uses final dataset license compatibility assessment table (the final table that is produced from our approach) to ascertain if the given dataset can be used commercially for the usage scenario they are interested in.

It is important to note that when an AI Engineer initiates a request to analyze the license compliance risks associated with a dataset, we first check if we already have conducted the analysis by looking up the results from the previous analyses. Only if we haven't analyzed the given dataset before, will we go through our proposed approach. Else, we simply return the stored license compatibility assessment table from the last time we assessed the given dataset.

In the remainder of this Section, we explain both phases and the steps and sub-steps contained in each phase. As a running example, we analyze the license compliance of CIFAR-10 dataset [26, 30].

3.1 Phase 1: License Identification

In this phase, an AI engineer identifies all the license details that need to be interpreted for potential license compliance violations. For the running example, the License Identification phase was individually carried out by the first three authors of the study to ensure correctness. The results of all the steps were considered final only after all three authors agree.

(Step 1) License extraction. Unlike OSS source code, dataset licenses are not made available in a standard fashion. Therefore, we first search for the license on the website from which the dataset was downloaded. If we are unable to find a license, **we then check if a license is provided as a separate file as part of the dataset**. If not, we reach out to the owners of the dataset asking about the dataset's license.

Running example. In this paper, we download the CIFAR-10 dataset from its official website. Also on the CIFAR-10 website, we find the following request from the dataset creators: "Please cite [31] if you intend to use this dataset," alongside a link to the paper. We extract this as the dataset's license.

(Step 2) Provenance extraction. The provenance of a dataset represents the true source of origin of the dataset [21]. A dataset could be first created by some researchers and could later be distributed through different platforms (sometimes even with a different license than the one assigned by the creators). In addition, it is important to verify if the dataset we obtained is indeed the same dataset created by the creators of the dataset. Therefore, it is pivotal to ascertain the provenance of a dataset to verify if the license that we extracted in the previous step is indeed the correct license of the dataset.

(Sub-step 1) Locate official dataset source. The first task is to locate the **official source of the dataset** (e.g., official website, research paper, or technical report) by querying search engines with appropriate search terms.

Running example. We have already located the official source of CIFAR-10, which we verify through the presence of the accompanying publication.

(Sub-step 2) Extract license and metadata of official source. We then extract the license and certain metadata about the dataset by collecting the details that we outline in the provenance schema, including the contents of the license. (See Table 1, which shows the recorded provenance of CIFAR-10.). For extracting the license, we recommend official websites as the first place to look for the license. Sometimes this requires creating an account. If a license is not found on the official website, then one can download the dataset to see whether the license is packaged with the dataset files. Once we extract a dataset's provenance, if the license that we extracted in the previous step contradicts the license extracted in this step, we use the license that we extracted through provenance as the dataset's license.

Running example. While extracting CIFAR-10's provenance, we learn that the authors of the website that hosts the dataset [30] and the authors of the associated publication are the same. As such, we simply collect the dataset's provenance details as shown in Table 1. Note that detailed explanation of our schema can be found in our replication package. In addition, since the license extracted in this and the previous step are consistent, we know the license that we extracted in the previous step is the dataset's license.

(Step 3) Lineage extraction. A dataset's lineage tracks the data sources from which a given dataset was created [2]. Many publicly available datasets, including computer vision and NLP datasets, are created by **collecting data from different sources**, which are generally other datasets or popular websites that host or index data points such as images. Not surprisingly, these different data sources typically have a license that is different from that of the dataset but are no less important to verifying the validity of the license of the dataset under examination. It is important to note that if a dataset is completely created from scratch (i.e., not extracted from any data source), then we do not have any lineage details for that dataset.

(Sub-step 1) Trace dataset creation process. We want to know how a dataset was created, which requires that we trace all of its data sources. The information we need to begin this sub-step is stored in the "Description of the data collection process" field that we recorded when we extracted the dataset's provenance. Then, if we learn that this data source in turn has other data sources, we also record those data sources recursively.

Running example. We previously learned through provenance extraction that CIFAR-10 is a subset of another dataset called 80 Million Tiny Images. Reading the research paper associated with 80 Million Tiny Images [48], we learn that this dataset has **7 data sources: Google, Flickr, Ask, Altavista, Picsearch, Webshots and Cydral**. Figure 2 is a visual representation of CIFAR-10's data sources.

(Sub-step 2) Locate official sources of data sources. We verify information about the data sources by locating their official sources (e.g., official dataset website, search engine and so on) similar to sub-step 1 of Provenance extraction.

Running example. First, we locate the official dataset website for 80 Million Tiny Images [47] through Google search. After learning that the dataset has been withdrawn, we then locate the archived versions of the website available on the Internet Archive [34]. Following that, we visit the official websites for each of the 7 data sources listed above.

(Sub-step 3) Determine the license range. A key challenge in determining the correct license of a data source is that a dataset might have been created some time ago. Since then, the license of the data source might have evolved. To mitigate this problem, we identify the license range associated with a given dataset. License range refers to the time period when the data contained in the dataset was likely collected. This, of course, is merely an estimate, but it is important because if a dataset's contents come from a data source (e.g., Google Images), then we need to know which past versions of the data source's license are relevant to the dataset (e.g., Google's Terms of Service from 2005,

and/or 2007, and/or 2012, and so on). Only the licenses that were effective during the license range of the dataset are of interest to us.

To determine the end of the license range, we use the *Origin date* field that we collect during the provenance extraction stage, which identifies the year when the dataset first originated. A variety of chronological information about a dataset can be used as a proxy for the origin date, including when the dataset’s research paper was first published, when the dataset was first made available as part of a programming challenge, and so on. We consider the year preceding the origin date to be the beginning of the license range. Data sources such as websites and search engines inherit their license ranges from the dataset to which they contribute content. However, the license range for data sources that are datasets themselves is based on the dataset’s origin date.

Running example. During provenance extraction, we learn that publication year for CIFAR-10 is 2009 [31], which we determine to be the *Origin date*. So we set the license range to be 2008 - 2009. Its top-level data source, 80 Million Tiny Images, was created in 2006 [47]. So its license range is 2005 - 2006. Then, the 7 data sources of 80 Million Tiny Images inherit the same licensing range as their parent dataset, 80 Million Tiny Images.

(Sub-step 4) *Identify license for data sources.* Using a Web archive [33], we collect the earliest available license of each data source during the license range. We do this recursively until we identify the license associated with all the data sources that contributed to the creation of the dataset under consideration. Along with the licenses, for each data source, we collect the provenance details that we outline in Table 1. We use a web archive because, the license as it appeared when the dataset was first released is the one that applies.

Running example. No license captures were available during the license range for CIFAR-10 and for two of its data sources (Ask and Webshots), so we simply store the earliest captures we can find. The license content of 80 Million Tiny Images and Cydral is completely unavailable, and so, we store *N/A*. For the remaining data sources, we record the earliest license captures from the correct license range and its provenance details. We do not show the provenance details recorded for each data source due to space constraints.

Table 1. Recorded provenance details for CIFAR-10

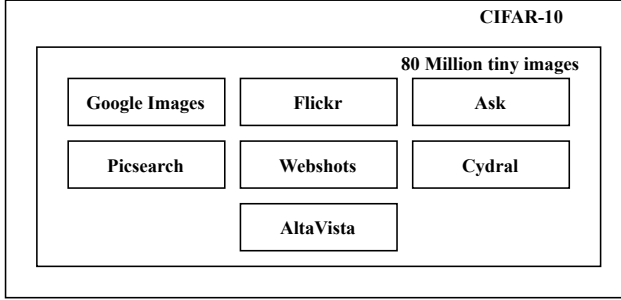
Dataset-related details	Dataset name	Dataset version	Origin date	Origin
	CIFAR-10	N/A	2009	https://www.cs.toronto.edu/~kriz/cifar.html
	Description of dataset		Description of data collection process	
	The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. 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

官方原始数据集的license信息，不包含每个image的license

3.2 Phase 2: License compliance assessment.

Once an AI engineer identifies the license associated with a dataset and the licenses of the data sources that were used to create the dataset, a lawyer decomposes each license to identify the rights

Fig. 2. Recorded lineage details for CIFAR-10 and its data sources (for each data source, we collect the provenance details that we mention in Table 1)



and obligations it outlines. As several prior studies in license and regulatory compliance [16, 17, 22] illustrate, software and legal expertise is required for this type of task. Therefore, in our paper we have consulted a lawyer to help us interpret these extracted licenses and complete Phase 2.

(Step 1) License interpretation. The first step of Phase 2 requires a lawyer to read the license associated with the dataset and the licenses of the data sources from which the data were collected. For each license, the lawyer then extracts the rights and obligations outlined in it.

In order to ensure that the extracted rights and obligations are documented in a standard format that facilitates understanding, we document them using a modified version of the Montreal Data License (MDL) proposed by Benjamin et al. [4]. The reason is because MDL standardizes and documents what rights are granted to the dataset itself (e.g., modify, distribute) and what rights are granted to the dataset in conjunction with AI software (e.g., train, research). These rights can be directly linked to the usage scenarios. For example, if the dataset license grants rights to distribute the dataset, then the usage scenario of distributing the dataset will be permissible. The MDL was originally proposed to create licenses for sharing new datasets. However, for this work, we re-purpose the MDL format to document the legal interpretations associated with datasets (i.e., to decompose licenses into their rights and obligations) and identify the usage scenarios that are acceptable. We further enhance MDL in the following three aspects:

- (1) The MDL does not record the obligations that the user of a dataset needs to fulfil when enjoying a specific right. Therefore, we enhance the MDL by capturing both rights and obligations for all fields in Rights to dataset and Rights to dataset in conjunction with AI models.
- (2) When recording Rights to dataset in conjunction with AI models, the MDL does not capture if the license allows reverse engineering the trained AI model to reconstruct the data. As such, we add a new field “Model reverse engineer” to record this.
- (3) We add fields to record license validity period and the liability or warranty provided by the license.

We call our modified MDL documentation schema the Enhanced MDL. We provide the schema and a description of each field in Table 2 in our replication package.

Running example. The main results of our decomposition can be seen in Table 2. The fields marked in blue represent our enhancement. In summary, a user has rights to access the CIFAR-10 data for every use case outlined in the schema. To do so, the user is required to cite the associated technical report, as requested on the official dataset website.

(Optional sub-step) Documentation and schema update. In the event that one needs to ascertain if a new usage scenario other than the ones that are dictated by the rights that our enhanced MDL schema captures, then they should enhance our schema. They can do so by adding a column to capture the rights and obligations that would pertain to their usage scenario.

Running example. In case we wanted to understand if we could build adversarial models with the given dataset, we would enhance our schema to include rights and obligations required to use the dataset to build an adversarial model.

(Step 2) License compatibility analysis. Based on the well formatted information of the Enhanced MDL, a risk assessment is performed to decide whether the dataset should be used in the commercial context for which it is being considered. The risk assessment, which involves interpreting the compatibility between multiple legal licenses, advises practitioners on what rights they can enjoy commercially and what obligations they have to satisfy to do so. These rights in turn inform the practitioners what usage scenarios they can use the dataset for, and what are the obligations that they need to fulfil to do so.

Before assessing the usage scenarios that are permitted, we need to ascertain the final rights and obligations that are associated with the given dataset after taking the licenses of the data sources into consideration. To reconcile the rights and obligations of using a dataset, we need to match use cases across the different interpreted licenses we have collected. We employ the following heuristics:

- (1) If a dataset license and all its data source licenses agree that a use case is allowed, then the action is permitted. This permission is reflected in the verified license.
- (2) If the license of any data source restricts a use case then, even if the main dataset license allows it, the verified license records this discrepancy by marking the action as not permitted.

Essentially, we aim to verify if the rights and obligations of the data sources allow for the dataset license to be valid. For instance, a very permissive dataset license cannot be valid if one of the data sources has a more restrictive license, since the license of the dataset would no longer respect the license under which it obtained the data. Our approach ends with this step after producing the licence compatibility analysis table.

Running example. For CIFAR-10, there are large discrepancies between what the dataset license allows and what the data source licenses allow. Table 2 shows the final rights that have changed after the license compatibility analysis in red. Tagging under Rights to Data is restricted by the license of several data sources of CIFAR-10 including Google and Flickr's licenses and, as such, is not permitted in the verified license (as indicated by the red cross in Table 2). In summary, we find that even though CIFAR-10's license allows all rights to the dataset (even commercial) as long as the paper is cited, the licenses of its data sources are more restrictive, creating potential risks of license compliance violation if the dataset is used for any commercial purposes including training an AI model, or modifying or distributing the dataset.

Scenario assessment. Using the rights allowed in the previous step as given by the final license compatibility assessment analysis table, we determine if the given dataset can be used for the required commercial usage scenario (e.g., to embed a model trained on the given dataset in a commercial product). If the rights permit the usage scenario, we convert the obligations to AI software requirements and make it a part of the AI software development life cycle so that it can be continuously traced, documented and tested.

Running example. If we want to check if the usage scenario of commercializing the model constructed using CIFAR-10 is allowed, we check if the model commercialization rights are present and what are its associated obligations. From Table 2 we find that model commercialization rights

are not present, hence the usage scenario of commercializing the model trained on CIFAR-10 will not be permissible.

Table 2. License interpretation of CIFAR-10 using the enhanced MDL schema. The fields in blue are our enhancements. The values in red are the changed rights after our license compatibility analysis step.

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	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

data source 的 license 不允许这些使用场景

4 CASE STUDY DETAILS

In this section, we explain the details of our case study. Through our case study, we assess the potential license compliance violations that we may incur if we use publicly available datasets to build commercial AI software.

4.1 Case Study Setup

First, we collected a list of popular publicly available datasets from several source like Papers with Code [52] and prior study by Benjamin et al. [4]. Then we chose six common datasets from the list that we obtained based on their popularity and likelihood to be used commercially. The datasets that we study are CIFAR-10 [31], ImageNet [43], Cityscapes [10], FFHQ [28], VGGFace2 [7] and MS COCO [32]. For instance, one of our studied dataset ImageNet is hailed as the dataset that changed machine learning [18]. Similarly, CIFAR-10 and Cityscapes datasets are listed as one of the top 10 most used dataset by Papers with Code [52]. Finally, we include FFHQ and VGGFace2 datasets in our study as several model stores use these datasets to pre-train their AI models [3, 23, 35]. These datasets are commonly used for tasks like object detection, image recognition, and semantic segmentation. In addition, all of our studied datasets are image datasets. We choose to focus on image datasets in our study as there are several AI applications and it would be extremely hard to generalize across multiple applications. The most common commercial usage scenarios associated with the datasets are: commercially distributing the datasets (DD), Release product with embedded AI models (RPEAI) trained on these datasets and Commercialize the model output (CAI).

We first study the characteristics of the licenses associated with the studied datasets. Then we assess the potential license compliance violations that may occur when using these publicly available datasets by using the approach that we outline in Section 3.

Table 3. Results of applying our approach on the studied datasets.

Dataset	Commercial usage scenario					
	DD		RPEAI		CAI	
	Yes (obligations)	No	Yes (obligations)	No	Yes (obligations)	No
CIFAR-10		✗		✗		✗
ImageNet		✗		✗		✗
Cityscapes		✗		✗		✗
FFHQ	C+D			✗		✗
VGGFaces2	A+E+D			✗		✗
MS COCO*		✗		✗		✗

A - Provide a link to license CC-BY-NC 4.0; B - Provide a link to the license CC-BY 4.0; C - Provide a link to license CC-BY-NC-SA 4.0; D - Remove infringing content as soon as possible when an infringement is detected. E- Indicate changes;
 *-Annotations and the images in the dataset are licensed differently, since the dataset is typically used with both the images and annotations, we consider them together. If only annotations are used commercially then the obligations are DD - B+E+D, RPEAI - B, CAI - B

4.2 Case study results

We describe our findings below by first discussing our results regarding the characteristics of the studied datasets in terms of the licenses that they use, and then we discuss the results of our license compliance analysis.

Characteristics of studied dataset licenses.

[Result 1] 4 out of the 6 studied datasets do not use any standard dataset license (e.g., Creative Commons 4.0). Among the studied datasets, only FFHQ and VGGFace2 use standard dataset licenses. MS COCO dataset only licenses their annotations under a standard license, they refer to the Flickr terms of use for the license of the images contained in their dataset. Both ImageNet and Cityscapes use a custom license. CIFAR-10 dataset does not explicitly mention any license. The authors of CIFAR-10 only request for a citation from the users of their dataset.

[Result 2] The two datasets that use a standard dataset license use licenses from Creative Commons (CC) license family. FFHQ and VGGFace2 creators license their datasets using CC-NC-SA-4.0 from the CC family of licenses. The MS COCO dataset creators license their dataset's annotations under CC 4.0.

[Result 3] Except Cityscapes dataset (which was created from scratch), all the other studied datasets are created by collecting data from one or more data sources. In fact, CIFAR-10 and ImageNet datasets have more than one data source from which they collect their data points. Such a result affirms our need for the approach that we outline in Section 3 which traces a dataset's lineage and identify its data sources, so the the license compliance can be verified.

License compliance analysis results.

[Result 4] All of the studied publicly available datasets might result in potential license compliance violation if they are used to build commercial AI software. From Table 3 we observe that, none of the studied licenses explicitly allow practitioners the right to commercialize an AI model trained on the data or even the output of the trained AI model. Such a result also effectively prevents practitioners from even using pre-trained models trained on these datasets. Publicly available datasets and AI models that are pre-trained on them are widely being used commercially [40]. In the case of the MS COCO dataset, only the annotations are licensed under CC 4.0, since users of the dataset typically require the images along with the annotation, using MS COCO dataset under any commercial scenario might result in potential license violation. However,

但是也没有explicitly ban

if only the annotations from MS COCO dataset is used to build commercial AI software, they are allowed to do so. However, in such a case practitioners are obligated to provide a link to its license, retain the copyright notice and not use MS COCO dataset's annotations to endorse their product. Failure to do so might once again result in a license compliance violation.

[Result 5] 4 out of the 6 studied publicly available datasets might result in potential license compliance violation if the dataset is modified. More specifically, for CIFAR-10, ImageNet, MS COCO and Cityscapes. Sometimes when a dataset is being used commercially, it is desirable for a dataset to either be modified or enhanced with additional data points or different labels [44]. For instance, as Cui et al. [11] show, augmenting the data (i.e., rerepresenting) can be a useful way of improving the performance of an AI model. However, doing so on the above publicly available datasets might result in potential license compliance violation. Even the other datasets, which permit modifications, outline the obligations for practitioners to indicate the exact changes that were made to the dataset. Failure to do so might result in potential license compliance violation.

[Result 6] 4 out of the 6 studied publicly available datasets might result in potential license compliance violation if the dataset is distributed along with the commercial AI software. As we notice from Table 3, distribution of CIFAR-10, ImageNet and CityScapes is not allowed. Despite this restriction, many platforms distribute these datasets [40]. For instance, DeepAI [12] distributes ImageNet and CIFAR-10 even though it is not allowed to do so. Even in the case of the remaining two datasets, where distribution of the dataset is allowed, the obligations outlined by the license state that they must be distributed with the same license of the dataset.

[Result 7] In certain cases, it might be necessary to safeguard the model against reverse engineering efforts or adversarial attacks [42, 46]. We note that for the Cityscapes dataset, model reverse engineering is not allowed. If a model trained on Cityscapes data is made available, it must safeguard against attempts that aim to recover the dataset from the trained AI model. Failure to do so might result in a license compliance violation.

[Result 8] In addition to rights, all six datasets outline obligations that must be followed. Failure to do so might result in potential license compliance violation. Table 3 outline the obligations that one must satisfy when using a given dataset for the outlined commercial usage scenarios. Failure to do so might result in license compliance violation. For instance, ImageNet's license explicitly requires practitioners to indemnify the ImageNet team against any claims arising from use of the dataset. FFHQ, VGGFace2 and MS COCO datasets' annotations require that, if they are distributed or modified, to be presented under the same license. In addition, among the studied datasets we observe that the obligations are standardized (i.e., similar obligations appear across multiple licenses). However, since we studied only 6 datasets, it is challenging to know if our results generalize.

Summary: Publicly available datasets might not be suitable to build commercial AI software. All of the six studied datasets contain potential risks associated with at least one commercial usage context.

5 FEEDBACK FROM INDUSTRY PRACTITIONERS

We presented our findings to the practitioners within our AI research teams and gathered feedback. In particular, we present our approach and the analysis results on one of the sample datasets. Then we ask these practitioners to perform license compliance analysis on new datasets of their interests using our approach.

Prior to our proposed approach, whenever the practitioners wished to use a publicly available dataset, they presented that dataset license to the lawyer and enquired if they could use the given dataset commercially. They only seek to interpret the license associated with the dataset without any consideration to the dataset's provenance or lineage. Furthermore, they were simply seeking if they could use the dataset in their AI software or not without distinguishing the type of use (e.g., using the dataset to train the AI model, or distribute the dataset). After using our proposed approach they realized that for several datasets (e.g., CIFAR-10) they could find the potential license compliance violations only after tracing a dataset's provenance and lineage. In addition, they expressed the usefulness of our approach and conveyed that our systematic approach and our case study results can help them demonstrate due diligence. They further expressed a keen interest in pushing our approach for wider adoption in their teams. Similarly, all the practitioners expressed the usefulness of the schema that we present as apart of our approach. In particular, they expressed that our enhanced MDL schema provided a great way to standardize our approach and document the rights and obligations associated with using a publicly available dataset. Furthermore, they facilitate better understanding of what commercial use cases can be undertaken.

While majority of the feedback were positive, most of the practitioners expressed concerns over the amount of manual effort involved in our approach. They further stated that the involved manual effort might be prohibitive in some cases to the adoption of our proposed approach. They wished for some automated tools that would help them with the application of our approach to assess license compliance violations.

6 RECOMMENDATIONS FOR AI ENGINEERS

Based on our case study results and feedback from the industry practitioners, here we provide three recommendations for the AI engineers who use publicly available datasets to create AI software.

Recommendation #1: Employ caution while using publicly available datasets to build commercial AI software. Just because a dataset is publicly available, it doesn't imply that it could be used to create commercial AI software. Therefore, it is pivotal for the AI engineers to use our approach to verify if the publicly available dataset that they are interested in using can indeed be used in the commercial context that they intend to use.

Recommendation #2: While assessing the license compliance of datasets, it is essential to use systematic approaches and clear documentations to demonstrate due diligence. We recommend the AI engineers to use our proposed approach and the schema to document various aspects (e.g., provenance, lineage and the rights/obligations) of a dataset and its license(s).

Recommendation #3: Share the knowledge regarding the risks associated with using a given publicly available dataset commercially. We recommend that AI Engineers to openly share and maintain the use cases where a given dataset can or cannot be used for the following three reasons: (1) to advocate the importance of this area of work, (2) to minimize the duplicated effort while analyzing the license compliance for the same datasets, and (3) to further validate the analysis via crowd sourcing.

7 THREATS TO VALIDITY

In this section, we describe the threats that pertain to our study.

7.1 External validity

In our paper we only explore if a given publicly available dataset can be commercially used without violating the license associated with the given dataset. However, we do not explore other factors like privacy and ethical issues that may impact the commercial usage of a given dataset. So even if our process finds that a publicly available dataset maybe commercially used, the user of the the

given dataset must verify if other factors might impact its usage. Finally, we only seek to answer if a given publicly available dataset can be used commercially which may not cover other cases such as using the given dataset for internal research or academic purposes. We also do not seek to ascertain if a given dataset can be used under fair use or fair dealing or other similar laws as it typically needs to be verified by a court of law. In addition, all of our studied datasets are image datasets. Different datasets like video or text based datasets might pose different challenges. We invite the future research to examine if our proposed approach can be used to assess the potential license compliance risks associated with using those datasets commercially.

7.2 Internal validity

In our study, we only consider the license of the data sources and do not consider the licenses associated with the **data points** (e.g., individual images) extracted from the data sources. Individual data points could have licenses as well. However, each dataset/data source has thousands of data points, making it impossible to extract the licenses of all the individual data points without an automated method. Even if we do manage to extract all their licenses, it is impossible for a lawyer to interpret these licenses and assess their compatibility. Although our proposed approach can demonstrate due diligence, this remains a threat. In addition, there can be multiple copies of the same dataset available across different sources with different licenses available. To disambiguate and identify the appropriate license(s) associated with a given dataset, our approach specifically extracts a given dataset's provenance and lineage information.

7.3 Construct validity

In our study, it is possible that the provenance and lineage that we extract for each of the studied datasets is not accurate. For instance, when extracting provenance, we determine the origin of the dataset by checking if the origin has a publication associated with it. However, it is impossible to know for certain if it is indeed a reliable indicator of when (and where) the dataset originated. Similarly, for some of the datasets, such as ImageNet, the exact data sources are unknown. In our paper, we make the best effort to infer data sources from the associated papers and documentations so that we can infer their licenses.

8 CONCLUSION

Publicly available datasets are being widely used to build commercial AI software. One can do so if and only if the license associated with the publicly available dataset provides the right to do so. However, it is not easy to verify the rights and obligations provided in the license associated with the publicly available datasets. Because, at times the license is either unclear or potentially invalid.

Therefore, in our study, we propose several schema and an approach that allows one to assess and anticipate the potential license compliance violations that may occur when using a publicly available dataset to build AI software. We demonstrate the usefulness of our approach through a case study that we conducted on six commonly publicly available datasets. Through our case study we find the publicly available datasets might not be suitable to build commercial AI software.

DISCLAIMER

Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of Huawei.

ACKNOWLEDGMENTS

We thank Mario Aguilar-Simon from Teledyne Scientific for his valuable comments regarding the interpretation of MS COCO dataset's license.

9 REFERENCES

- [1] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. 2019. A Marketplace for Data: An Algorithmic Solution. In *Proceedings of the 20th ACM Conference on Economics and Computation, Phoenix, AZ, USA, June 24 - 28*.
- [2] Mark Allen and Dalton Cervo. 2015. *Multi-domain master data management: Advanced MDM and data governance in practice*. Morgan Kaufmann.
- [3] AWS. 2021. GluonCV YOLOv3 Object Detector. https://aws.amazon.com/marketplace/pp/prodview-5jlv43tsn3ny?sr=0-1&ref_=beagle&applicationId=AWSMPContessa [Last visited on 08-30-2021].
- [4] Misha Benjamin, Paul Gagnon, Negar Rostamzadeh, Chris Pal, Yoshua Bengio, and Alex Shee. 2019. **Towards Standardization of Data Licenses: The Montreal Data License**. *arXiv preprint arXiv:1903.12262* (2019).
- [5] Travis Breaux and Annie Antón. 2008. Analyzing Regulatory Rules for Privacy and Security Requirements. *IEEE Transactions on Software Engineering* (2008).
- [6] Travis D Breaux, Annie I Antón, Kent Boucher, and Merlin Dorfman. 2008. Legal Requirements, Compliance and Practice: An Industry Case Study in Accessibility. In *Proceedings of the 16th IEEE International Requirements Engineering Conference*.
- [7] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*.
- [8] European Commission. 2019. Ethics Guidelines for Trustworthy AI. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> [Last visited on 08-30-2021].
- [9] European Commission. 2021. Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206> [Last visited on 08-30-2021].
- [10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2015. The Cityscapes Dataset. In *Proceedings of the 2015 CVPR Workshop on the Future of Datasets in Vision, Boston, MA, USA, June 11*.
- [11] Xiaodong Cui, Vaibhava Goel, and Brian Kingsbury. 2015. Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23 (2015), 1469–1477.
- [12] DeepAI. 2021. Datasets for machine learning, artificial intelligence, and statistics. <https://deepai.org/datasets> [Last visited on 09-01-2021].
- [13] Josef Drexler, Reto Hilty, Luc Desautettes-Barbero, Franziska Greiner, Daria Kim, Heiko Richter, Gintare Surblyte, and Klaus Wiedemann. 2016. Data Ownership and Access to Data-Position Statement of the Max Planck Institute for Innovation and Competition of 16 August 2016 on the Current European Debate. *Max Planck Institute for Innovation & Competition Research Paper* (2016).
- [14] Yuanrui Fan, Xin Xia, David Lo, Ahmed E Hassan, and Shanping Li. 2021. What Makes a Popular Academic AI Repository? *Empirical Software Engineering* (2021).
- [15] Centers for Disease Control and Prevention. 2021. Health Insurance Portability and Accountability Act of 1996 (HIPAA). <https://www.cdc.gov/php/publications/topic/hipaa.html> [Last visited on 08-30-2021].
- [16] Daniel German and Massimiliano Di Penta. 2012. A Method for Open Source License Compliance of Java Applications. *IEEE Software* 29 (2012), 58–63.
- [17] Daniel M German and Ahmed E Hassan. 2009. License Integration Patterns: Addressing License Mismatches in Component-based Development. In *Proceedings of the 31st International Conference on Software Engineering, ICSE, Vancouver, BC, Canada, May 16 - 24*.
- [18] Dave Gershgorin. 2017. The Data that Transformed AI Research—and Possibly the World. <https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/> [Last visited on 08-30-2021].
- [19] Dave Gershgorin. 2021. GitHub’s automatic coding tool rests on untested legal ground. <https://www.theverge.com/2021/7/7/22561180/github-copilot-legal-copyright-fair-use-public-code> [Last visited on 09-03-2021].
- [20] GitHub. 2021. GitHub Copilot: Your AI pair programmer. <https://copilot.github.com/> [Last visited on 09-03-2021].
- [21] Michael W Godfrey. 2015. Understanding software artifact provenance. *Science of Computer Programming* 97 (2015), 86–90.
- [22] David G Gordon and Travis D Breaux. 2014. The role of legal expertise in interpretation of legal requirements and definitions. In *Proceedings of the 22nd International Requirements Engineering Conference (RE)*. IEEE, 273–282.
- [23] TensorFlow Hub. 2021. deeplabv3-mobilenetv3-cityscapes: Lightweight deep learning model for semantic image segmentation. <https://tfhub.dev/sayakpaul/lite-model/deeplabv3-mobilenetv3-cityscapes/1/default/2> [Last visited on 08-30-2021].
- [24] Silvia Ingolfo, Alberto Siena, John Mylopoulos, Angelo Susi, and Anna Perini. 2013. Arguing regulatory compliance of software requirements. *Data & Knowledge Engineering* 87 (2013), 279–296.
- [25] Luke Jones. 2021. Microsoft’s GitHub Copilot Met with Backlash from Open Source Copyright Advocates. <https://winbuzzer.com/2021/07/06/microsofts-github-copilot-met-with-backlash-from-open-source-copyright->

- advocates-cxwbn/ [Last visited on 09-03-2021].
- [26] Kaggle. 2021. Popular Datasets Over Time. <https://www.kaggle.com/benhamner/popular-datasets-over-time/code> [Last visited on 09-22-2021].
 - [27] Georgia M Kapitsaki, Frederik Kramer, and Nikolaos D Tselikas. 2017. Automating the License Compatibility Process in Open Source Software with SPDX. *Journal of Systems and Software* 131 (2017), 386–401.
 - [28] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4401–4410.
 - [29] Nadzeya Kiyavitskaya, Nicola Zeni, Travis D Breau, Annie I Antón, James R Cordy, Luisa Mich, and John Mylopoulos. 2008. Automating the Extraction of Rights and Obligations for Regulatory Compliance. In *Proceedings of the 27th International Conference on Conceptual Modeling, Barcelona, Spain, October 20-24*.
 - [30] Alex Krizhevsky. 2021. The CIFAR-10 dataset. <https://www.cs.toronto.edu/~kriz/cifar.html> [Last visited on 09-22-2021].
 - [31] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. *Learning Multiple Layers of Features from Tiny Images*. Technical Report. University of Toronto.
 - [32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
 - [33] Internet Archive Wayback Machine. 2021. Internet Archive Wayback Machine. <http://web.archive.org/> [Last visited on 10-02-2021].
 - [34] Internet Archive Wayback Machine. 2021. Internet Archive Wayback Machine Calendar. <http://web.archive.org/web/20100601000000/http://groups.csail.mit.edu/vision/TinyImages/> [Last visited on 10-02-2021].
 - [35] Modzy. 2021. Image Classification Model by Modzy. <https://www.modzy.com/marketplace/model-task/label-or-classify/image-classification-2/> [Last visited on 08-30-2021].
 - [36] Government of Canada. 2021. A guide to copyright. <https://laws-lois.justice.gc.ca/eng/acts/c-42/page-9.html> [Last visited on 09-01-2021].
 - [37] Government of Canada. 2021. Infringement of Copyright and Moral Rights and Exceptions to Infringement (continued). <https://laws-lois.justice.gc.ca/eng/acts/c-42/page-9.html> [Last visited on 09-01-2021].
 - [38] Office of the Privacy Commissioner of Canada. 2021. Submission of the Office of the Privacy Commissioner of Canada on Bill C-11, the Digital Charter Implementation Act, 2020. [Last visited on 08-30-2021].
 - [39] US Copyright Office. 2021. More Information on Fair Use. <https://www.copyright.gov/fair-use/more-info.html> [Last visited on 09-01-2021].
 - [40] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. 2021. **Mitigating dataset harms requires stewardship: Lessons from 1000 papers.** *arXiv preprint arXiv:2108.02922* (2021).
 - [41] Julia Reda. 2021. GitHub Copilot is not infringing your copyright. <https://juliareda.eu/2021/07/github-copilot-is-not-infringing-your-copyright/> [Last visited on 09-03-2021].
 - [42] Maria Rigaki and Sebastian Garcia. 2020. A survey of privacy attacks in machine learning. *arXiv preprint arXiv:2007.07646* (2020).
 - [43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115 (2015), 211–252.
 - [44] Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data* 6 (2019), 1–48.
 - [45] Svetlana Sicular, Jim Hare, and Kenneth Brant. 2019. *Hype Cycle for Artificial Intelligence, 2019*. Technical Report. Gartner, Inc.
 - [46] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. 2017. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*. 587–601.
 - [47] Antonio Torralba, Rob Fergus, and Bill Freeman. 2021. Tiny Images. <https://groups.csail.mit.edu/vision/TinyImages/> [Last visited on 10-02-2021].
 - [48] Antonio Torralba, Rob Fergus, and William T Freeman. 2008. 80 Million Tiny Images: A Large Dataset for Nonparametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30 (2008), 1958–1970.
 - [49] Jean-Paul Triaille, Jérôme de Meeüs d’Argenteuil, and Amélie de Francquen. 2014. Study on the legal framework of text and data mining (TDM). *European Union Studies KM-03-13-426-EN-N* (2014).
 - [50] The Authors Guild v. Google. 2015. US Court of Appeals, New York, 13-4829, 2015.
 - [51] Sander Van Der Burg, Elco Dolstra, Shane McIntosh, Julius Davies, Daniel M German, and Armijn Hemel. 2014. Tracing Software Build Processes to Uncover License Compliance Inconsistencies. In *Proceedings of the 29th ACM/IEEE International Conference on Automated Software Engineering*.
 - [52] Papers with Code. 2021. Papers with Code: The latest in Machine Learning. <https://paperswithcode.com/> [Last visited on 09-03-2021].

- [53] Yuhao Wu, Yuki Manabe, Tetsuya Kanda, Daniel M German, and Katsuro Inoue. 2015. A Method to Detect License Inconsistencies in Large-scale Open Source Projects. In *Proceedings of the 12th IEEE/ACM Working Conference on Mining Software Repositories*.
- [54] Yuhao Wu, Yuki Manabe, Tetsuya Kanda, Daniel M German, and Katsuro Inoue. 2017. Analysis of License Inconsistency in Large Collections of Open Source Projects. *Empirical Software Engineering* 22 (2017), 1194–1222.
- [55] Nicola Zeni, Nadzeya Kiyavitskaya, Luisa Mich, James R Cordy, and John Mylopoulos. 2015. GaiusT: Supporting the Extraction of Rights and Obligations for Regulatory Compliance. *Requirements Engineering* 20 (2015), 1–22.
- [56] Hongyu Zhang, Bei Shi, and Lu Zhang. 2010. Automatic Checking of License Compliance. In *The 26th IEEE International Conference on Software Maintenance*.