WASHINGTON JOURNAL OF LAW, TECHNOLOGY & ARTS
VOLUME 13, ISSUE 3 SPRING 2018

HOW MACHINES LEARN: WHERE DO COMPANIES GET DATA FOR MACHINE LEARNING AND WHAT LICENSES DO THEY NEED?

Rachel Wilka, Rachel Landy, and Scott A. McKinney[*]

Cite as: 13 Wash. J.L. Tech. & Arts 217 (2018)

http://digital.law.washington.edu/dspace-law/handle/1773.1/1815

ABSTRACT

Machine learning services ingest customer data in order to provide refined, customized services. Machine learning algorithms are increasingly prominent in multiple sectors within the software-as-a-service industry including online advertising, health diagnostics, and travel. However, very little has been written on the rights a company utilizing machine learning needs to obtain in order to use customer data to improve its own products or services.

Machine learning encompasses multiple types of data use and analysis, including (a) supervised machine learning algorithms, which take specific data provided in a tagged and classified format to deliver specific predictable output; and (b) unsupervised machine learning algorithms, where untagged data is processed in

---

[*] Scott McKinney and Rachel Landy are senior associates in the Technology Transactions practice at Wilson Sonsini Goodrich & Rosati, P.C. (WSGR). Scott is an adjunct professor at Georgetown Law and a guest lecturer for Cornell University's Cornell Tech grad program. Rachel represents numerous technology companies on matters relating to intellectual property and commercial contracts. Rachel Wilka is Corporate Counsel at Zillow Group, Inc. and lead counsel for Zillow Rentals, hotpads, Inc., and dotlop, Inc., supporting all product counseling, licensing, commercial partnership, and risk management matters. Rachel was previously a technology transactions associate at Wilson Sonsini Goodrich & Rosati. The views expressed in this article are the authors' own and do not represent the views of WSGR, Zillow, or any of the authors' other clients. Thank you to Manja Sachet and Rob Philbrick.

order to look for patterns and correlations without a specified output.

This Article introduces the reader to the types of data use involved in various machine learning models, the level of data retention normally required for each model, and the risks of using personal information or re-identifiable data in connection with machine learning. The paper also discusses the type of license a commercial provider and consumer would need to enter into for various types of machine learning software. Finally, the paper proposes best practices for ensuring adequate rights are obtained through legal agreements so that machines may self-improve and innovate.

TABLE OF CONTENTS

INTRODUCTION

Machine learning—it's been a technology catch-phrase for at least five years, a tagline for any company purporting to "innovate a new future," but what does it actually mean? Machine learning services ingest data in order to provide refined, customized services to users.[1]

Real world utilization of machine learning increases daily, as more and more companies use the technology for market trend analysis, price setting, development of company (or industry) best-practices, medical diagnoses, insurance—virtually any industry that has representable and analyzable output information can be optimized through machine learning.[2]

---

[1] *See What is Machine Learning?*, COURSERA, https://www.coursera.org/learn/machine-learning/lecture/Ujm7v/what-is-machine-learning (last visited 4/19/2018).

[2] *See* Louis Columbus, *10 Ways Machine Learning is Revolutionizing Marketing*, FORBES (Feb. 25, 2018), https://www.forbes.com/sites/louiscolumbus/2018/02/25/10-ways-machine-learning-is-revolutionizing-marketing/#803e5fe5

The algorithms that drive machine learning are increasingly prominent within the software-as-a-service industry, where machine learning can be leveraged for multiple industries, including online advertising, health diagnostics, and travel.[3] Despite the increased use of machine learning across business sectors, the rights a company utilizing machine learning needs to obtain in order to use outside data to improve its own products are often amorphous and misunderstood. As machine learning becomes integral to companies across all industries and those companies become more and more reliant upon datasets for use in their machine learning analysis, the data itself (and the corresponding rights in such data) becomes increasingly important.

This Article examines the legal data rights a company needs to obtain in order to use data for machine learning, and how those rights change depending on the machine learning model and business application. Part I of this Article defines machine learning and analyzes the various use cases for machine learning based on differing data rights. Part II discusses how companies may use data for different purposes. Part III discusses the varying degrees of data retention a company may undertake. In Part IV, we follow that discussion with an overview of data sources a machine learning company could access. Part V discusses the laws and legal risks relating to the use of data (including personally identifiable information ("PII")) in machine learning applications across commercial sectors. Lastly, Part VI provides recommendations and considerations for drafting data licenses.

I. BACKGROUND

A. Definition of Machine Learning

The term "machine learning", which is widely credited to ex-

---

bb64.

[3] *See* Forbes Technology Council, *Looking Ahead: The Industries That Will Change The Most As Machine Learning Grows,* FORBES, https://www.forbes.com/sites/forbestechcouncil/2017/03/08/looking-ahead-the-industries-that-will-change-the-most-as-machine-learning-grows/#4c45248c647b

IBM employee Arthur Samuel,[4] is the ability of computers ("machines") to learn without being guided or re-programmed.[5] Samuel's initial machine learning example was a machine that can be programmed to play checkers better than the person who designed the program. Remarkably, a computer could be trained to do this in eight to ten hours of playing time *over sixty years ago* using machine learning.[6] All that was necessary to train the computer was to provide it with the rules of the game, a general sense of direction regarding how the game worked, and a list of parameters that were thought to have something to do with the game, but whose correct background signs and relative importance were unknown and unspecified to the computer.[7] In relatively short order, the machine learned how to play checkers better than its programmer, without the programmer having to revise the initial computer code or manually train the computer in strategy.[8]

The use cases for modern machine learning are virtually boundless. Machine learning is best used in tasks for which designing code with explicit task-specific instructions is difficult or impossible, such as ranking, optical recognition, complex problem solving, and filtering.[9] Machine learning applications typically involve feeding (relatively) automated programs a large data set of inputs, and solving problems or identifying issues using results-driven decisions based on the data set.

To be clear, machine learning (in the classic sense) is *not* artificial intelligence. Although machine learning does involve learning by experience, a machine learning algorithm does not act intelligently,[10] and is not flexible in changing environments.[11] However, we see the concepts become increasingly conflated, as

---

[4] *See* A.L. Samuel, *Some Studies in Machine Learning Using the Game of Checkers*, 3 IBM JOURNAL OF RESEARCH AND DEVELOPMENT 210 (1959).

[5] *Id*.

[6] *Id*

[7] *Id*.

[8] *Id*.

[9] ETHEM ALPAYDIN, INTRODUCTION TO MACHINE LEARNING 6–8 (3rd ed. 2014).

[10] *See* discussion *infra* Part I.B.

[11] DAVID POOLE ET AL., COMPUTATIONAL INTELLIGENCE: A LOGICAL APPROACH 1 (1998).

algorithms are commonly programmed with artificial intelligence, and as machine learning algorithms come to make up a greater part of the artificial-intelligence ecosystem.[12] Machine learning should not be conflated with data mining, either.[13] Unlike data mining, which usually focuses on uncovering previously unknown properties of a dataset, machine learning typically focuses on better-predicting outcomes or revising an algorithm based on already-known properties of that dataset.

Below we discuss the common types of machine learning and the different levels of data use associated with different machine learning models.

B. Types of Machine Learning

Machine learning can be split into three major categories: (1) supervised, (2) reinforcement, and (3) unsupervised.[14] We discuss each in turn below.

1. Supervised

With supervised machine learning, one knows the desired output of the algorithm based on a dataset, usually referred to as "training data," that is used to optimize a performance criterion.[15] Supervised machine learning algorithms are typically "taught" using a training dataset. If the algorithm provides unexpected or incorrect results

---

[12] *See, e.g.*, Fred Jacquet, *Exploring the Artificial Intelligence Ecosystem: AI, Machine Learning, and Deep Learning*, DZONE/ AI ZONE (Jul. 4, 2017), https://dzone.com/articles/exploring-the-artificial-intelligence-ecosystem-fr.

[13] *But see* ALPAYDIN, *supra* note 9, at 2 (describing the application of machine learning methods to a database as "data mining."). Opinions regarding, and semantical definitions of the term "machine learning" vary.

[14] *See generally* OLIVIER CHAPELLE, ET AL., SEMI-SUPERVISED LEARNING (2006). *available at* http://www.acad.bg/ebook/ml/MITPress-%20SemiSupervised%20Learning.pdf; *see also* Vishal Maini, *Machine Learning for Humans, Part 5: Reinforcement Learning*, MEDIUM.COM (Aug. 19, 2017), https://medium.com/machine-learning-for-humans/reinforcement-learning-6eacf258b265.

[15] *Id.*; *see also Data Sets and Machine Learning*, DEEP LEARNING FOR JAVA https://deeplearning4j.org/data-sets-ml (last visited Mar. 31, 2018); ALPAYDIN, *supra* note 9, at 3.

after analyzing the base data using the training dataset, the programmer can make algorithmic tweaks (or changes to the training data) to right the course. In supervised machine learning, all of the data within a training data set is "labeled" (or assigned a value), which allows the machine to easily compare analysis data against the training set baseline.[16] The algorithm generates information based on its analysis of the training data, and uses that information to produce inferred or revised functions. These revised functions can be used by the end user to discern new trends regarding a dataset, or to refine the algorithmic analysis itself.[17] Analyzing enormous data sets at a speed only computers can achieve, the algorithm can identify trends, flag otherwise unidentified issues, and give the algorithm operator other desired results that can be tweaked using variations in the algorithm or training data.

2. Unsupervised

In unsupervised machine learning, there is no training data, and the outcomes are unpredictable.[18] Unsupervised machine learning algorithms can solve problems using input datasets alone, with no reference or training data, by recognizing patterns in the data and grouping together reoccurring or common data characteristics.[19] Unlike supervised algorithms, which rely on labeled data, unsupervised machine learning uses functions to uncover previously unknown properties of a dataset using unlabeled data. For example, say you had a dataset comprised of apples, oranges, and bananas, and want to analyze and identify trends in the fruit. The problems are: the data set is huge, the fruit are all jumbled together, and none of the data is labeled as an "apple," an "orange," or a "banana." In a supervised machine learning scenario, if the algorithm was not "taught" to identify an apple, it would not know to look for, nor group together, apples. In contrast, an unsupervised machine learning algorithm is able, over time, to recognize that data across the datasets have similar characteristics, such as being shiny, red,

---

[16] *Id*.
[17] *Id*.; *see also* DEEP LEARNING FOR JAVA., *supra* note 15.
[18] ALPAYDIN, *supra* note 9, at 11.
[19] *Id*.

and generally apple-shaped. Unsupervised algorithms can identify these similarities and group together the apples with the apples, the oranges with the oranges, and the bananas with the bananas. Unsupervised machine learning can seem to border on artificial intelligence,[20] and companies often use it to analyze large datasets of customer transactions, generate common trends or characteristics based on the past transactions, group those customers into clusters, and use that cluster of information to refine the company's business model.[21]

There is a sub-class of supervised machine learning called "semi-supervised" machine learning, in which an algorithm-operator uses a small amount of labeled training data to inform a much larger unlabeled dataset.[22] Semi-supervised machine learning is usually thought of as halfway between unsupervised and supervised learning.[23] Both supervised and semi-supervised machine learning tend to lend themselves to relatively predictable outcomes, and are often used by companies to optimize user experiences based on predicted or predetermined outcomes.

3. Reinforcement

Reinforcement learning is based on an algorithm that has a concept of how an environment should behave, and learns an optimal behavior for such an environment by analyzing repetition and repeated failures over time.[24] Unlike supervised machine learning, reinforcement learning algorithms are not presented with input/output pairs for correction—instead, the algorithm is performance-driven.[25] One well-known example of reinforcement

---

[20] *See* Bernard Marr, *Supervised V Unsupervised Machine Learning – What's The Difference?*, FORBES (Mar. 16, 2017, 3:13 AM), https://www.forbes.com/sites/bernardmarr/2017/03/16/supervised-v-unsupervised-machine-learning-whats-the-difference/#4ecd3f80485d.

[21] ALPAYDIN, *supra* note 9, at 12.

[22] CHAPELLE, ET AL., *supra* note 14, at 2–3.

[23] *Id*.

[24] *See* Leslie Pack Kaelbling, Michael L. Littman & Andrew W. Moore, *Reinforcement Learning: A Survey*, JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH 4, 237 (1996).

[25] *Id.*

learning is the self-driving car industry.[26] Many self-driving algorithms are not artificially intelligent in the traditional sense, but instead use repetition (i.e. driving thousands of test miles and tracking driving errors and successes) to optimize the algorithm and underlying technology in a way that human programmers could never do on their own.[27] Another way to think about reinforcement learning is "trial-and-error", but on a massive scale accomplishable only by computers.[28] Over time, the software learns what to do, and what not to do, until its functionality is optimized for the task at hand.

II. LEVELS OF DATA USE ASSOCIATED WITH DIFFERENT MACHINE LEARNING MODELS

The use case for machine learning implementation dictates the data rights that must be obtained, as well as the applicable data retention and use policies. For example, consider these three different use cases:

- OpenTable recommends restaurants, but can only do so based on the information it collects (e.g. where the user has dined before, not the actual dish he or she actually eats—information OpenTable does not have).[29]

- To predict which show a user will want to binge next, Netflix wants to know that user's viewing history, and some relevant demographic information, such as age, gender, and location.[30]

- Accolade's Maya Intelligence Option inputs information

---

[26] *See* Will Knight, *Reinforcement Learning*, MIT TECHNOLOGY REVIEW (March/April 2017), https://www.technologyreview.com/s/603501/10-breakthrough-technologies-2017-reinforcement-learning/.

[27] *Id*.

[28] Maini, *supra* note 14.

[29] *OpenTable Privacy Policy*, OPENTABLE, https://www.opentable.com/legal/privacy-policy (last updated May 15, 2017).

[30] *Netflix Privacy Statement*, NETFLIX, https://help.netflix.com/legal/privacy (last updated Nov. 30, 2016).

> about an individual's health insurance, medical history, medications, test results, and other personal health information in order to provide personalized healthcare support.[31]

Like all companies that depend on machine learning, these companies obtain, use, and retain data in different ways, depending on their business model and their machine learning models.

A. Supervised

Supervised machine learning presents clearer use cases. The outcome is predictable, and in fact, programmed. Netflix and OpenTable, for example, ingest user preference data to produce individualized recommendations to that user. These algorithms do not necessarily rely on extraneous data inputs—they are trained to provide recommendations if certain inputs are present. But by continuously ingesting new data, the engine can be refined and perfected on an ongoing basis. For example, over time, Netflix may be able to distinguish between medical-drama fanatics who want to binge Grey's Anatomy and those who prefer ER. For this reason, the results of supervised machine learning can be highly valuable to companies in any industry, but especially those industries that are consumer-facing.

However, for both Netflix and OpenTable, the use of the data (recommendations) is not these companies' core business; rather, it is an added feature that has helped propel both companies to the top of their respective industries. Without compelling recommendations, Netflix would still be a video streaming service. However, it relies on data to enhance the user's experience, thus adding value to the service.[32] Netflix does this by ingesting and inferring from a user's preferences. For example, it knows if you watched one episode of Gilmore Girls, or if you watched every

---

[31] ACCOLADE, https://www.accolade.com/solutions/ (last visited March 30, 2018).

[32] Chris Raphael, *How Machine Learning Fuels Your Netflix Addition*, RTINSIGHTS (Jan. 5, 2016), https://www.rtinsights.com/netflix-recommendations-machine-learning-algorithms/.

season five times, and it can use that information to determine whether you were a superfan or lost interest quickly.

The same is true, to a lesser extent, with OpenTable. OpenTable bases its recommendations largely on collections of user ratings.[33] However, OpenTable's capabilities are limited. Its model does not know whether its users actually ate at a restaurant booked through OpenTable. It only knows how that user feels about the restaurant if he or she rates it on the app. Furthermore, the app does not know, for example, whether dietary preferences affected that rating.

One benefit of supervised machine learning algorithms is that, in the early stages, potential data sets can be separated into those that are necessary and those that are merely helpful. A company may find that data sets with particular characteristics are subject to more extensive regulations than the data required to successfully implement a machine learning solution. As a result, the company will either utilize the data differently, or avoid implementation of the data altogether. For example, Netflix, in its early days, may have found that age was highly useful. However, unless the appropriate controls are in place, gathering other sensitive information, such as children's' names, can result in significant legal risk.[34] Nevertheless, using machine learning, a start-up company may find that it can estimate age based on user habits, thereby making it unnecessary to undertake the legal risk of gathering that information directly.[35]

B. Unsupervised

Using unsupervised machine learning is a process best thought of as "high risk, high reward." Without a clearly defined desired

---

[33] Pablo Delgado & Sudeep Das, *Using Data Science to Transform OpenTable Into Your Local Dining Expert*, presentation at SparkSummit 2015, *available at* https://www.slideshare.net/SparkSummit/using-data-science-to-transform-opentable-into-delgado-das.

[34] *See, e.g.*, Children's Online Privacy Protection Act of 1998, 15 U.S.C. §§ 6501–6506 (1998).

[35] This is contrary to companies operating in the healthcare space, which almost always need some level of personal health information—another highly regulated category of data. For those companies, the risk is inherent in the business and should be priced into the model for customers.

output, the company may not know what it needs, or even what it is likely to get, from the algorithm. On the other hand, a company might get results that it did not anticipate or even think were possible. Unsupervised machine learning is popular in the health-tech industry because making a diagnosis requires analyzing many variables that human doctors cannot necessarily test for individually.[36] Machine learning gives doctors the assistance they need to take in a large amount of data and then spit out all known potential diagnoses. The Maya Intelligence Option, for example, could benefit from taking in numerous health data points in order to generate a potential treatment plan, the scope of which would not be pre-defined.

Unsupervised machine learning, by its nature, requires that the operator have more flexibility in its use of data sets. As a result, the data use rights obtained from data providers (discussed in Part V) for use in unsupervised machine learning analysis should be broader than data use rights for supervised machine learning. For example, speech recognition software operators obtain broad rights to use data collected through the software (i.e. users' speech). The Apple Terms of Service state: "By using Siri or Dictation, you agree and consent to Apple's and its subsidiaries' and agents' *transmission, collection, maintenance, processing, and use* of this information, including your voice input and User Data, *to provide and improve* Siri, Dictation, *and dictation functionality* in other Apple products and services."[37] While Apple's main purpose in collecting this data is likely to tune its engine to recognize speech more efficiently, such a broad license also allows the operator to use the speech for a number of ancillary purposes, such as understanding dialects, intonations, and speech impediments. Thus, the operator is not sure what the results will be or how those results may be used in the future. Indeed, an operator may find that certain data sets once considered vital turn out to be useless. Prior to implementation, the machine learning algorithm cannot necessarily predict which data is valuable and

---

[36] *See, e.g.*, Chip M. Lynch, Victor H. van Berkel, Hermann B. Frieboes & Bin Liu, *Application of Unsupervised Analysis Techniques to Lung Cancer Patient Data*, PLOS ONE (Sept. 2017), available at http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0184370.

[37] *Apple Ios Software User Agreement*, APPLE INC., at 3 (emphasis added) available at https://www.apple.com/legal/sla/docs/ios6.pdf (last revised 2012).

which is not. This uncertainty necessitates a broader, less restrictive scope of operator rights than in other scenarios. In some cases, this may mean that the operator must assume the additional risks of using, collecting, or storing data that is subject to regulation.

Overall, companies' use cases and data supply needs should help inform whether their algorithms are unsupervised, reinforced, or supervised. Accordingly, the rights to be obtained to that data, discussed in Part V, should reflect those business decisions. Moreover, in addition to the data use rights that must be obtained, we must also consider the data storage and retention issues associated with machine learning.

III. RETENTION

In addition to determining whether an algorithm should be supervised or unsupervised, any machine learning company must determine the scope of its data retention policy. Data retention policies track how data is stored, shared, and deleted to ensure consistency of data treatment and compliance with contractual obligations, applicable law, and best practices. As discussed in Part II, the particulars of a data retention policy for a machine learning company rely on the use case for the algorithm and the data-treatment requirements imposed by the data source.

For example, a supervised machine learning environment may only need to retain training data if it is not using new data to improve its capabilities. Or, it may only need to retain the data for a limited period of time in order to establish overall patterns or features to include in training data. In our Netflix example, it may be helpful for Netflix to know that over a two-year period, a user watched all of Dawson's Creek, Gilmore Girls, and 7th Heaven, but not Buffy the Vampire Slayer.[38] Knowing, in context, that the user prefers real-life teen dramas to science-fiction teen dramas can help improve the algorithm.

By contrast, an OpenTable user's eating habits may not follow predictable patterns. The fact that a user ate at a Chinese restaurant five days in a row is helpful for understanding the user's culinary tastes that week. But that same user could then decide she's had

[38] This assumes that all of the programs mentioned are available on Netflix.

enough Chinese food for a year, and move on to sushi. Thus, for OpenTable, pattern analysis is less important than it is for Netflix; it can simply build on each data input individually without a longer-term analysis. Where Netflix may be able to determine that a user had a child based on a change in viewing habits (and could adjust accordingly), OpenTable's use case doesn't require a long data retention period to provide a benefit.

Ultimately, assuming the operator has obtained the requisite rights from users (discussed in Part V), the operator ought to retain the data for as long as is commercially reasonable (although the relevant industry market approach may dictate that data be destroyed after a certain amount of time). To mitigate the potential harm of data destruction requirements, an operator should always retain the training data it used to fix bugs and help tune the algorithm. Other than the training data, a company could find that it need not retain a lot of individual data inputs so long as the algorithm has previously ingested, responded, and reacted to the data.

Some data providers try to contractually require data destruction after the term of an engagement.[39] Operators of unsupervised algorithms should always push back; the nature of those algorithms is such that there could always be a golden needle in a data-haystack, so an operator should try to retain the right to continue to mine the data for as long as possible. If a customer is insisting on destruction, an operator may promise anonymization and aggregation of the data so the customer could not be identified. Ultimately, the operator must determine at what point the algorithm (and the operator's business) will be able to live without the data, i.e., when it has obtained sufficient replacement data to be self-sustaining. In other words, what retention term is reasonable for the company? The operator may be able to compromise by agreeing to only use a customer's data in perpetuity where that data is anonymized and aggregated with other customers' data sets. A company that destroys data will also need to develop an appropriate support policy if the original reference set is eventually deleted.

---

[39] *See, e.g.*, *Data License Agreement*, PRACTICAL LAW COMPANY INTELLECTUAL PROPERTY & TECHNOLOGY, available at https://us.practicallaw.thomsonreuters.com/w-004-3938.

IV. SOURCES OF DATA

Companies looking to obtain data to create or train machine learning algorithms tend to look to four sources: (a) data sets sold through data brokers; (b) batch uploaded data from software installed on-premises for customers; (c) ongoing customer data collection from network-connected software as a service offering (both for customer-facing improvements and other company purposes); and (d) open public data sets.[40]

A. Data Sets Sold Through Data Brokers

Data brokers are companies that have gradually built databases of consumer data. These databases were originally built for "marketing, fraud detection, and credit scoring purposes."[41] Companies can go to data brokers to purchase data sets, usually with personally identifiable information removed. Data brokers may offer a database (or set of databases) that tracks behaviors the operator wants to build a machine-learning algorithm around. Data broker databases can include demographic data, court and public records data, social media and technology data, consumer interests data, financial data, health data, and purchase behavior data.[42] However, some observers doubt whether data broker databases are sufficiently anonymized to avoid business or regulatory risk.[43] Another downside of purchased data is that the purchaser runs the

---

[40] *See, e.g.*, SEATTLE OPEN DATA PORTAL, https://data.seattle.gov/ (last visited May 10, 2018).

[41] Bernard Marr, *Where Can You Buy Big Data? Here Are The Biggest Consumer Data Brokers*, FORBES (Sept. 7, 2017), https://www.forbes.com/sites/bernardmarr/2017/09/07/where-can-you-buy-big-data-here-are-the-biggest-consumer-data-brokers/#48d997096c27.

[42] *See* Leo Mirani & Max Nisen, *The Nine Companies That Know More About You Than Google or Facebook*, QUARTZ (May 27, 2014), https://qz.com/213900/the-nine-companies-that-know-more-about-you-than-google-or-facebook/.

[43] *See* Alex Hern, *Anonymous Browsing Data can be Easily Exposed, Researchers Reveal*, THE GUARDIAN (Aug. 1, 2017), https://www.theguardian.com/technology/2017/aug/01/data-browsing-habits-brokers.

risk of the data not being tailored to its exact needs, thereby making it less useful in providing the desired predictive output.[44] The largest American data brokers include Axciom, Corelogic, and Datalogix.[45]

B. Ongoing Customer Data Collection From Network-Connecting Software as a Service Offering

The most common method of collecting training data is to collect data directly from users of an operator's service. Data collected from consumers can be acquired in different ways: (a) web activity, provided when a consumer interacts with the company's website; (b) consumer surveys and other feedback mechanisms; (c) mobile user data, provided through consumer interaction with a company app; and (d) social media.[46] In order to obtain necessary rights to consumer data, the operator should include a license in its governing user agreement (e.g., the consumer terms and conditions of use) and accurately disclose the data collection and use in its privacy policy. We discuss obtaining rights to service user data in more detail in Part V.

C. Batch Uploaded Data From Software Installed On-Premises for Customers

For customers not connected to the operator's network automatically (i.e., customers that do not use a hosted or software-as-a-service product), operators can choose to negotiate the right to receive a bulk package of use data through a manual upload or other transfer mechanism. This type of data collection most often occurs where the operator's product is installed on-premise, which may be due to: (a) industry privacy sensitivity, for example, in the medical and financial sectors; (b) consumer desire for customized

---

[44] *See, e.g.*, INFOBASE, https://www.acxiom.com/what-we-do/infobase (providing a large user database with numerous information points gathered, over time, in response to different requests).

[45] Mirani, *supra* note 42.

[46] *See* DEALNEWS, *How Online Retailers Collect and use Consumer Data*, CULT OF MAC (May 26, 2016) https://www.cultofmac.com/430158/how-online-retailers-collect-and-use-consumer-data-deal-news/.

solutions;[47] or (c) the nature of the product lends itself better to on-site installation.[48] On-premise software can involve a negotiated paper agreement (instead of a shrink-wrap or click-through agreement), so companies need to be careful that the necessary data rights are not negotiated out of the agreement.

D. Open Source Public Data Sets

Finally, academic institutions, individual researchers, and 'open-source advocates'[49] have created pre-populated data sets for common machine-learning algorithm problems. For example, the University of California at Irvine currently maintains 413 data sets that are open to the public for use in machine learning algorithms.[50] Generally, the rights to these data sets are less restrictive than one would find in a negotiated bilateral agreement, as open source licenses tend to be permissive by nature. However, operators should still evaluate the applicable data license terms to be aware of any requirements to contribute developed technology back to the open source community, and other requirements of the license (e.g., to provide attribution). Descriptions of most common open source licenses are maintained by the Open Source Initiative.[51]

V. LAWS/LEGAL RISKS AROUND USE OF DATA/PII IN MACHINE LEARNING

The legal risks of using data generally depend on the following

---

[47] *See* Thomas Peham, *On-Premise vs. Cloud Software: A Comprehensive Comparison*, USERSNAP, https://usersnap.com/blog/comparison-of-cloud-vs-on-premise-enterprise-software/ (last visited Mar. 31, 2018).

[48] *See* HOST ANALYTICS, https://hostanalytics.com/blog/on-premises-versus-cloud-based-epm-software-which-is-right-for-your-business/.

[49] Open source advocates are generally thought of as zealous individuals, who believe that as much of the internet and developing software as possible should be made open to the public. *See, e.g.*, CBSNEWS, *Oracle names Open-Source Evengelist*, CNET (Sept. 7, 2005), https://www.cnet.com/news/oracle-names-open-source-evangelist/.

[50] *See* UCI MACHINE LEARNING REPOSITORY, http://archive.ics.uci.edu/ml/index.php (last visited Mar. 31, 2018).

[51] *See* OPEN SOURCE INITIATIVE, https://opensource.org/ (last visited Mar. 31, 2018).

factors: (a) the relative sensitivity of the data; (b) the types of predictions to be produced; (c) the agreement governing the acquisition and use of the data; and (d) the impact on a broader industry or market.

A. Use of Sensitive Data

The legal risk associated with a machine learning algorithm is determined, at least in part, by the sensitivity of the source data. In other words, if regulated data is an input, then the output is also likely to be regulated (or considered sensitive data of the same category). Sensitive data is more often regulated, and penalties for non-compliance with regulatory schemes for sensitive (e.g., personally identifiable) data often carries harsher penalties.[52] In addition, data providers (like business-to-business operators or data brokers) may be more hesitant to agree to provide sensitive data that is subject to extensive regulations, due to their fear of being held accountable for misuse by a third party of data they originally collected.

The primary categories of what we often consider sensitive data are not surprising: (a) health data; (b) financial data; (c) educational data; (d) location data; (e) visual data (photos of a consumer); and (f) data regarding children. Importantly, if an operator seeks to use sensitive data to make predictions within the given industry, the operator will fall under the purview of industry regulators.[53] For example, if educational data is used to predict educational outcomes for students, or financial data is used to determine credit-worthiness, the resulting predictions would likely be subject to similar regulatory schema.

In addition, operators may be required to handle data in a

---

[52] *See, e.g.*, *Legal Resources*, FEDERAL TRADE COMMISSION, https://www.ftc.gov/tips-advice/business-center/legal-resources?type=case&field_consumer_protection_topics_tid=250 (last visited May 10, 2018).

[53] For example, HIPAA will apply to data clearinghouses, processors, and clearinghouses, as well as business associates which will include most health-software providers *See Are You a Covered Entity?*, CENTERS FOR MEDICARE AND MEDICAID SERVICES, https://www.cms.gov/Regulations-and-Guidance/Administrative-Simplification/HIPAA-ACA/AreYouaCoveredEntity.html (last visited May 10, 2018).

specific way, or even store data for longer periods of time, based on the sensitivity of the industry. For example, in the health context, the Health Insurance Portability and Accountability Act requires that certain health-related data (but not all) be retained for at least six years.[54] Particular categories of health providers are subject to additional retention requirements. For example, Medicare managed care providers must retain records for at least ten years.[55] While the operator itself may not be a managed care provider, it may be a subcontractor to one who is required to be bound by the same retention policies. In those cases, it is common for the "covered entity" (i.e., the entity bound by the law) to contractually "pass through" certain data retention requirements under HIPAA to all of its subcontractors.

B. The Output Use Case

Certain machine learning outputs may create undue legal risk, even if the data is collected in compliance with any applicable laws. For example, an operator's use of data to predict a consumer's credit-worthiness will result in a company being classified as a "Credit Reporting Agency."[56] Credit reporting agencies are subject to burdensome regulations.[57] As another example, the use of data in a device to predict health outcomes can lead to a product or service being classified as a medical device, which is subject to regulation by the Food and Drug Administration, including things like fitness

---

[54] *See* Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191.

[55] 42 C.F.R. § 422.504(d)(2)(iii) (2011).

[56] *See Credit Reporting*, FEDERAL TRADE COMMISSION, https://www.ftc.gov/news-events/media-resources/consumer-finance/credit-reporting (last visited Apr 1, 2018); *see also What is a credit reporting company?*, CONSUMER FINANCE PROTECTION BUREAU (May 25, 2017), https://www.consumerfinance.gov/ask-cfpb/what-is-a-credit-reporting-company-en-1251/.

[57] Even those who merely furnish information are subject to reporting and notice requirements. *See Consumer Reports: What Information Furnishers Need to Know*, FEDERAL TRADE COMMISSION, https://www.ftc.gov/tips-advice/business-center/guidance/consumer-reports-what-information-furnishers-need-know (last updated Mar. 2018).

trackers and massage chairs.[58] As discussed in Part V.A., detection of legal wrongdoing in these cases often does not require analyzing the actual data use, and can be determined solely from the resulting product.

C. Breach of Contract/License

One of the larger areas of legal risk for operators using data in machine learning algorithms is the risk of non-compliance with the agreements under which data rights are obtained. If a company relies on a small number of customers for the majority of its revenue, just one dispute can have an enormous impact on the company, especially if the details of the alleged misuse are made public. Such an allegation, even if unfounded, could harm the company's ability to attract future customers. For example, the unauthorized use of a customer's data could be considered a breach of confidentiality (if the data is identified as being subject to confidentiality terms), intellectual property infringement (to the extent any intellectual property rights are embodied in the data), or misappropriation of trade secrets (depending on how the data is misused), which could result in breach of contract claims, claims in tort, or statutory damages for copyright infringement.

Additionally, it is critical that operators relying on a few large enterprise customers use that data correctly (i.e., consistent with the data use rights in the customer license agreement). The loss of one large customer could destroy the viability of the algorithm.

It is important to keep in mind, however, that private actions (e.g., between two private parties) to enforce violations of data use terms are limited by the customer's ability to detect the operator's wrongdoing. It is often difficult or impossible for a customer to know, or to prove, that a company uses individual data in machine learning algorithmic analyses. To address this information imbalance, new methods of detecting illegal collection and use of data have evolved over the last few years. For example, to uncover

---

[58] Given the rise of internet of things, new ways to deal with these devices/requirements are being explored. *See FDA Selects Participants for New Digital Health Software Precertification Pilot Program*, FOOD AND DRUG ADMINISTRATION (September 26, 2017), https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm577480.htm.

Bing's practice of copying data and functionality, Google inserted false hits in their search engine functionality and monitored Bing to see if the false stories or incorrect results also appeared in Bing's results in the same order. Additionally, parties more frequently negotiate contractual auditing rights to allow searching for wrongful use of data directly in the service provider's files.[59]

D. Impact on the Larger Market/Industry

Finally, because widely-adapted machine learning algorithms are a relatively recent technological development, novel regulations and industry controls are being created in an attempt to police new concerns as they arise. Outside of the United States, the Australian government is looking into whether machine learning should be considered anti-competitive in particular use cases because it can create the ability to more easily base pricing off of a competitor and allow parties without any actual direct communication to participate in a tacit price fixing scheme.[60]

VI. WHAT NEEDS TO BE CONSIDERED WHEN DRAFTING AN AGREEMENT FOR A MACHINE LEARNING SERVICE

Different operators will rely on different license terms to obtain data depending on the proposed data use. First, an operator must determine whether it is interested in the rights to the results output, or just improvements to the algorithm. Second, the operator must determine if it is attempting to buy data or simply collect data through a service it is already offering. Third, the operator must visualize the desired machine learning output. The actual output will often dictate the terms of the license required to offer the machine learning service.

---

[59] *See* Marc Silverman, *The Right to Audit Clause*, WITHUM, SMITH & BROWN, https://www.withum.com/kc/right-audit-clause/ (last visited Apr. 1, 2018); *see also* Danny Sullivan, *Google: Bing Is Cheating, Copying Our Search Results*, SEARCH ENGINE LAND (Feb. 1, 2011), https://searchengineland.com /google-bing-is-cheating-copying-our-search-results-62914.

[60] *See* Tas Bindi, *Big Data and Machine Learning Algorithms Could Increase Risk of Collusion*, ZDNET (Nov. 16, 2017), http://www.zdnet.com/article/big-data-and-machine-learning-algorithms-could-increase-risk-of-collusion-accc/.

A. Predictions Versus Algorithm Improvements

Not all machine learning operators have the same level of interest in using the results of an algorithm in future work. Some operators are intimately interested in the accuracy of the result, but not the result itself. For example, a marketing platform that predicts whether an individual will click on an image with particular attributes will not care about whether the consumer goes on to buy the linked product. Instead, it cares only about which attributes the image contains and whether the attributes had the predicted effect (i.e., caused the consumer to click the link). The relevant data are image attributes and the user's "clicks," rather than the customer's content. In contrast, a medical imagery predictive algorithm would want to know if its software successfully or unsuccessfully predicted the presence of a medical condition, *and* all of the specific outcomes that were or were not correctly predicted. As a result, that operator would need a license to obtain more specific data about each diagnosis.

B. Source of Data

As discussed in Part IV, some consumer-facing companies offer data-gathering services and data can also be obtained through wholesale acquisitions of databases. Data gathered through negotiated agreements with customers can vary depending on: (a) whether the company is business-to-business ("B2B") or business to consumer (a business providing a service to an individual consumer) ("B2C"); (b) industry norms and data sensitivity; and (c) customization of the product and algorithm.[61] Operators should be cognizant of the different rights negotiated with each customer, and maintain minimum acceptable terms to avoid violation of customer agreements. By contrast, purchased data generally has fewer limitations which may only restrict the purchaser from specific high-risk activities, like predicting credit-worthiness or re-identifying

---

[61] *See* Daniel Glazer et al., *License Scope and Restrictions and Original versus Derived Data,* available at https://us.practicallaw.thomsonreuters.com/4-532-4243.

individuals.[62]

C. Output

Finally, both public perception and potential legal consequences of machine learning data use are dependent on the final output of the algorithm. Consider the medical industry. Given the public interest in improving and refining medical care, consumers may be more likely to allow companies to use their data to develop software that will diagnose a specific ailment based on individual attributes. The customers themselves have a stake in the result and thus less resistant to sharing their data. However, information about personal health is highly sensitive. Consumers may be willing to allow the use of their data, but only if it is anonymized. An operator should be aware that in some cases, it is far more likely to get the data sets it needs if it promises to protect the consumer's identity.[63]

D. Recommendations for Drafting

When drafting an agreement to acquire data for use in a machine learning algorithm, there are several aspects of the license one should consider. This Section discusses a number of considerations for data licenses, including: (1) license duration; (2) ownership of created output; (3) requirement for data to be provided in a de-identified/non-sensitive format; (4) combining data with other data sets; and (5) promises that data is gathered in accordance with applicable law.

---

[62] As an example, Acxiom states that data sets from their site: "contain information on individuals and households in the U.S. and are developed from many sources, including public records, publicly available information, and data from other information providers. Acxiom's marketing products are used by qualified companies, non-profit organizations and political organizations in their marketing, fundraising, customer service and constituent service and outreach programs to provide customers and prospects with better service, improved offerings and special promotions." *Highlights for US Products Privacy Policy*, ACXIOM.COM, https://www.acxiom.com/about-us/privacy/highlights-for-us-products-privacy-policy/ (last visited Apr. 20, 2018).

[63] These promises could, of course, expose the operator to significant legal risk if they are broken.

1. License Duration

A data license should not be time-limited. This is particularly important if the algorithm makes continuing reference to source data. If the license itself cannot be perpetual, then the operator should retain perpetual rights to any improvements or derivative works of the data so that the effectiveness of the algorithm is not diminished.

If an operator must agree to a time-limited license that requires the return of data, then it should be aware how difficult it can be to identify exactly which machine learning result is attributable to a specific data set or individual piece of data. The model should improve and evolve with each new data set added. Therefore, the ideal data license will be perpetual, notwithstanding termination of the underlying agreement.

Additionally, an operator must be aware that a large enterprise customer could insist that a data license be revocable in the event of an operator's breach of the underlying agreement. If the license were revoked, the operator would likely be required to return all data. As discussed, that can be an incredibly cumbersome task to undertake. As a result, it is critical for the operator to ensure compliance with its data license agreements to avoid a license revocation that compromises the algorithm. Concerns about time limitations in a license are less of an issue with data licensed from data brokers, as data brokers often grant perpetual licenses.

2. Ownership of Created Output

Ownership of the output of a machine learning algorithm is another important consideration. Enterprise customers, particularly those with negotiating leverage, will often attempt to claim that any technology, intellectual property, or other output developed by referencing their original data belongs to them. That approach is reasonable in a consulting arrangement with a defined project scope, but not necessarily in the machine learning context, where the operator continuously uses its customers' data to offer an improved product to every current and future customer.

Therefore, it is critical that the operator maintains ownership of its algorithm, as well as the improvements to the algorithm

generated based on its customers' data in order to protect the operator's key intellectual property. As a fallback position, the operator could attempt to transfer ownership of any custom developed features for the specific client or consumer-data reliant improvement if: (a) that improvement or model alone is unusable by the customer in any context other than the operator's algorithm; and (b) the operator is granted a perpetual, unlimited, royalty-free, sublicensable license to the developed model or improvements for use in its products and services.

3. Requirement for Data to be Provided in a De-Identified/Non-Sensitive Format

Machine learning operators often do not want to assume the risk of hosting a platform which produces predictions that could inadvertently reveal an individual's personally identifiable information ("PII"). If the operator gathers data from customers, it must ensure that customers strip their data of any PII or otherwise take on the risk of removing PII. Some enterprise customers, on the other hand, may refuse to provide any PII and will agree to represent that no PII is included in their data sets. Data brokers may also agree to similar terms, or undertake removal themselves. In any event, the customer's privacy policy (if it is required to have one) should ensure that the customer has the right to provide the data to the operator. The operator can then ask the customer to represent and confirm that all data is provided in compliance with the privacy policy.

4. No Prohibition on Combining Data With Other Data Sets

Machine learning algorithms, by their nature, improve with exposure to more and more data, regardless of the source. If data is collected in bulk from an external source, any prohibition on commingling that data with data from other sources undermines the usefulness of that data set. This issue often arises when purchasing data from data brokers, who may have negotiated no commingling provisions with their providers that are passed on to purchasers of the data. An operator could address this issue in its agreement with a data broker by agreeing that there will be no commingling that

results in the identification of individuals or that connects PII to an anonymized/de-identified data set.

Obtaining the rights to combine data sets can be especially important since demonstrating compliance with a contractual requirement to keep data sets separate can be nearly impossible. Certain aspects of data may be present in multiple data sets, and machine learning output may be reliant on multiple data sets, so showing that particular data came from one source and not another is not feasible.

5. Representation That Data was Gathered in Accordance With Applicable Law

Finally, when obtaining data from an external data source, a machine learning operator will have little control over how the data was originally gathered, and very little insight as to whether the collection complied with applicable law. As such, the operator must rely on the representations and warranties of its data providers as to the legality of the data, and should ensure that the applicable representations and warranties are in the underlying data agreement. The operator should insist on these representations and warranties and refuse to deal with any provider that will not agree to them.

CONCLUSION

While the concept of machine learning is not new, the ubiquity of machine learning applications has seen a significant upswing over the past five to ten years. In the legal sector, drafting appropriate license language and associated data use rights for machine learning applications requires lawyers to understand what exactly machine learning is and how it differs from traditional software licensing or service provider scenarios. The most important point to take into consideration when drafting a machine learning license is that all data use is not created equal. How data is gathered, processed, and stored will depend on the type of machine learning model and the goals of the organization using the data. Therefore, to appropriately draft a license, attorneys should examine the data cycle with their client to understand how data will be gathered, processed, stored, and retained. The specifics of the data type, use, processing and

storage will affect a multitude of legal and contractual issues relevant to the data use license itself, including, but not limited to, breadth of license, data use timeframe, and handling of derivatives. Attorneys should also take into consideration sensitivity of data use, collection and retention within a given industry, as well as factors such as consumer perception and the machine learning algorithms' output to help them better advise clients on the "real-world" risks of using different types of data in their business.

PRACTICE POINTERS

- License duration (*term of the agreement versus perpetual*): Understand how long the company needs to refer back to the data (including whether data will be needed for fixing later-discovered flawed outcomes) and whether the data can be separated from the algorithm without affecting functionality.
- Ownership of created output (*customer-owned or company-owned*): Understand whether output is customer specific or increases the value of the algorithm as a whole, and whether the algorithm using training data continues to process improvements from both old and company-created data inputs.
- Data Identifiability (*anonymous versus individual characteristics*): Understand which data is likely to be used as a predictor, and whether anonymization of data would affect the ability to create valuable output. Additionally, consider the federal and state statutes applicable to the type of data processed by the algorithm (e.g., HIPPA for health-related data).
- Data Set Combination (*allowed or prohibited*): Understand whether data-set combination is likely to re-identify personally identifiable information regarding individual data subjects, and which attributes of a data set need to be correlated with to produce valuable output.
- Responsibility for gathering data in compliance with law (*company versus outside data source*): If data is gathered in bulk from an outside source (including from a data broker, a

white-labeled incorporation of the algorithm, or an open source set), the outside party should bear primary responsibility for gathering the data in compliance with law. For data gathered directly from a customer, the company will likely bear primary responsibility for informing the consumer and obtaining consumer consent. For data gathered from the internet (via webscraping or other similar techniques) without the express consent of the data source, the attorney should analyze whether such data collection (1) violates law, or (2) violates online terms of service agreements, and the attorney and company should together conduct a risk-benefit analysis of such data collection.