How To Identify, Prepare And Package Data For Monetization In AI

By Charley Macedo, Nicole Spence, Benjamin Beavan and Barry Brager

As computers have become more powerful and the collection of data has exploded, monetization opportunities for an organization with respect to its treasure troves of data are becoming more and more prevalent. In order to assist organizations on how to prepare and package data for monetization in artificial intelligence(AI)/machine learning(ML) opportunities,[1] this article provides an overview of key considerations outlined at the 2020 LES USA and Canada Annual Meeting in a virtual presentation that was shared with the same title.

Part I provides a primer on how ML works and the role data plays in it.

Part II provides a summary of key contractual considerations for licensors and users of data for ML applications based on IBM's typical clauses for The Weather Channel (a licensor of data) and Watson (a licensee of data).

Part III provides insights on how Getty Images has developed an innovative business in licensing data sets of images and videos for use with ML applications.

Part IV provides an explanation of two real world examples of business systems built using ML technology from Perception Partners.

Part I: Primer on AI/ML and Training Sets

Everybody has data. In order to be useful as a training set, the data collection will need to include a large and consistent set of data. To illustrate how training sets are developed, consider the real world type of experience shared by everyone of the collection of health-related data from medical providers.

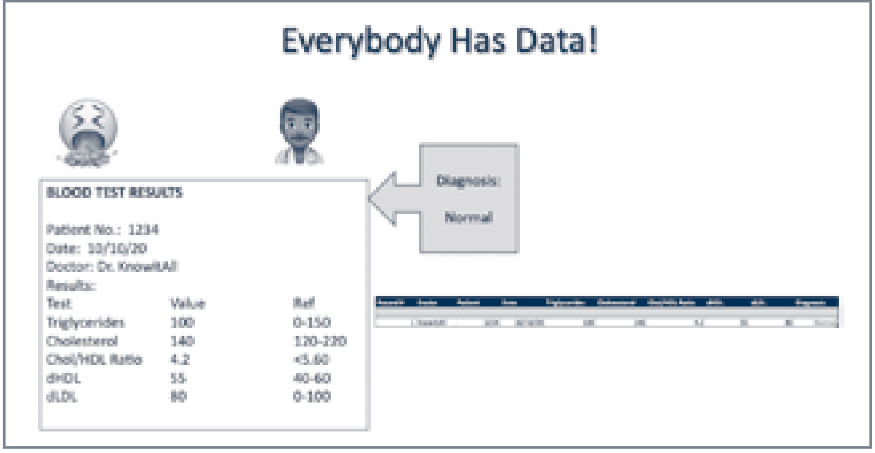
A. Building the Training Set

The process begins when an individual patient feels sick and visits his/her doctor. The process will begin with the doctor, after obtaining an appropriate consent, running perhaps one or more tests, like a blood test on the patient, and the doctor making a diagnosis.

Figure 1 illustrates an example of such data collection. In this example, the doctor "Dr. Knowitall" took a sample on October 10, 2020 of "Blood Test Results" for patient identified as "Patient No. 1234," with various tests and data items. Each test is labeled by the type of test taken and includes a value of the specific test result. While not always necessary, the reference provided with each respective blood test can be used to see if the specific value is normal or abnormal. The doctor will also provide a diagnosis based on the test results, which may be tagged to the sample.

This information typically gets transferred to a database, which can be represented as a row in a table. Figure 1 illustrates how all the information contained in the Blood Test Result taken by the doctor for the patient on the specified date is included in a single row of the table. This is called a "sample," and each row in this database reflects a single sample that will ultimately be used to build the training set. The different "features" of the training set are the different blood tests that were taken, in this case triglycerides, cholesterol, cholesterol/HDL ratio, dHDL, and dLDL. All of these different features have the label of what the test is in the header column and the value or the data items of the actual results of that test in the respective role for each sample. Each sample will have its own set of values and data items, while the labels will remain the same for all the different values that are of the same feature. The tag of what the diagnosis is reflects the categorization of the data sets that ultimately will be used by an ML algorithm to classify an incoming query

Figure 1. Example Of Blood Test Results Being Transferred To A Sample In A Database



BLOOD TEST RESULTS

Patient No.: 1234
Date: 10/10/20
Doctor: Dr. KnowitAll
Results:

Test	Value	Ref
Triglycerides	100	0-150
Cholesterol	140	120-220
Chol/HDL Ratio	4.2	<5.60
dHDL	55	40-60
dLDL	80	0-100

[1] ML is a type of AI. This article will focus in particular on training sets for ML.

(as discussed below). In this case, the tag characterizes this data set as "normal."

Of course, the same patient may visit the doctor many times over the course of treatment. Each time the patient visits the doctor, a new sample of test data may be taken, and a separate diagnosis tagged to the sample. As illustrated in Figure 2, each test may be entered as a separate sample (or row) in the database. Figure 2 illustrates separate samples of the same features being taken by the same doctor for the same patient with different sample dates, respective sample values and respective tags of diagnosis.

However, the same doctor may have in his or her database other patients. Figure 3 illustrates the database including samples taken not only from the same patient 1234, but additional samples taken from other patients. Here, there are additional samples that were taken from patient 5431 on different dates. Those different dates for those different samples are reflected with different diagnoses: sometimes normal, sometimes abnormal.

In addition to each doctor having multiple patients, a training set may include samples taken by multiple doctors for each of their various patients. Figure 4 illustrates the database also including additional data or samples that were taken by a different doctor, "Dr. Imagood," for a different patient, identified by patient number 9783 on different dates with different diagnoses and test result values or data items being filled in.

Collectively, when all of these samples are put together, a training set is formed. There could be thousands, hundreds of thousands, or even millions of samples included that can make up what becomes the training set.

B.How a Training Set is Used in ML

Section A began with the premise that everybody has data, and the data can be put together to form what can be considered a training set. With that foundation, this Section B discusses how a training set fits into a ML system.

A ML system is thought of as a black box. It has an input and an output. The input is usually in the form of a query data set. The query data set would include data items that match features in the training set. The output would match the tags in the training set. When the input is put into the black box, the ML algorithm is applied to the training set to come up with the solution in the form of the output (see Figure 5, page 233).

In Section A, a training set based on alphanumeric data items was described. However, training sets can include other kinds of data. For example, training sets can include images, videos, and/or sounds.[2]

C. Real World Applications of AI and ML Systems

There are many relevant examples of real world applications that are used today. For example:

- In healthcare and medicine, ML can be and is used to diagnose and stratify patients and treat disease based on those diagnoses.
- In financial services, ML can be and is used to identify and predict optimal investments. It can also be used to detect fraud.
- Marketing systems use ML to identify and present personalized advertisements.
- Recruitment and human resource systems can identify best potential candidates and/or problem employees.

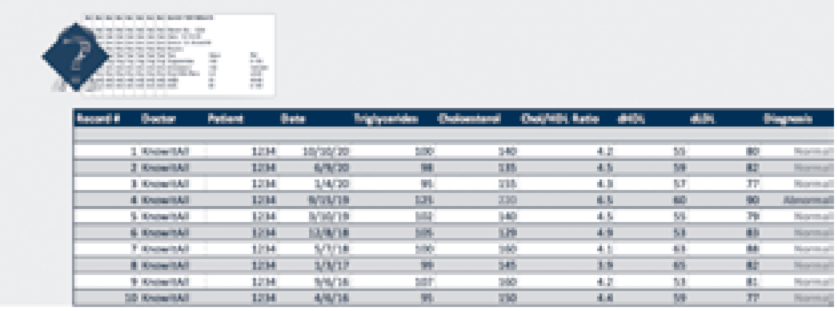
■ Charles R. Macedo,
Partner,
Amster, Rothstein &
Ebenstein LLP,
New York, NY U.S.A.
E-mail: cmacedo@arelaw.com

■ Nicole Spence,
Senior Attorney, IP Law,
IBM Corporation,
Armonk, NY U.S.A.
E-mail: Nicole.Spence@ibm.com

■ Benjamin Beavan,
Global Director of
Strategic Development,
Getty Images, Inc.,
London, England
*E-mail: Ben.Beavan@
gettyimages.com*

■ Barry Brager, CLP, RTTP,
IP Strategy Leader,
Perception Partners,
Atlanta, GA U.S.A.
*E-mail: bbrager@
perceptionpartners.com*

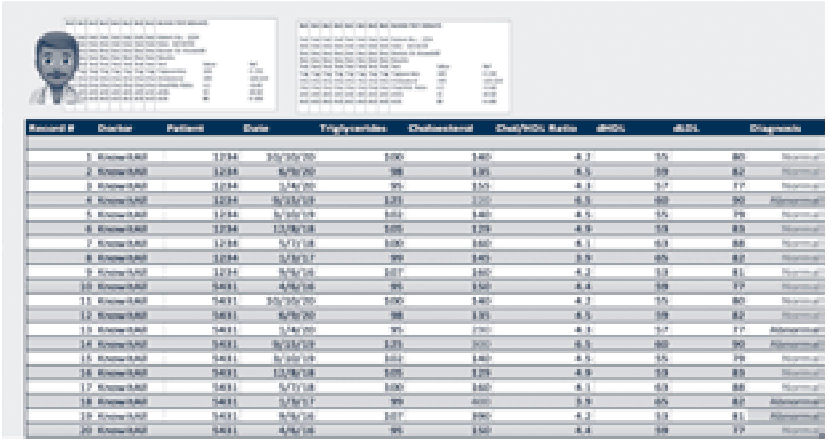
Figure 2. Illustration Building A Dataset Based On Multiple Samples From The Same Patient



Record #	Doctor	Patient	Date	Triglycerides	Cholesterol	Chol/HDL Ratio	dHDL	dLDL	Diagnosis
1	KnowItAll	1234	10/10/20	100	140	4.2	55	80	Normal
2	KnowItAll	1234	6/9/20	98	135	4.5	59	82	Normal
3	KnowItAll	1234	1/4/20	95	115	4.3	57	77	Normal
4	KnowItAll	1234	9/15/19	125	220	6.5	60	90	Abnormal
5	KnowItAll	1234	3/10/19	102	140	4.5	55	79	Normal
6	KnowItAll	1234	12/8/18	105	129	4.9	53	83	Normal
7	KnowItAll	1234	5/7/18	100	160	4.1	63	88	Normal
8	KnowItAll	1234	1/3/17	99	145	3.9	65	82	Normal
9	KnowItAll	1234	9/5/16	107	160	4.2	53	81	Normal
10	KnowItAll	1234	4/6/16	95	150	4.4	59	77	Normal

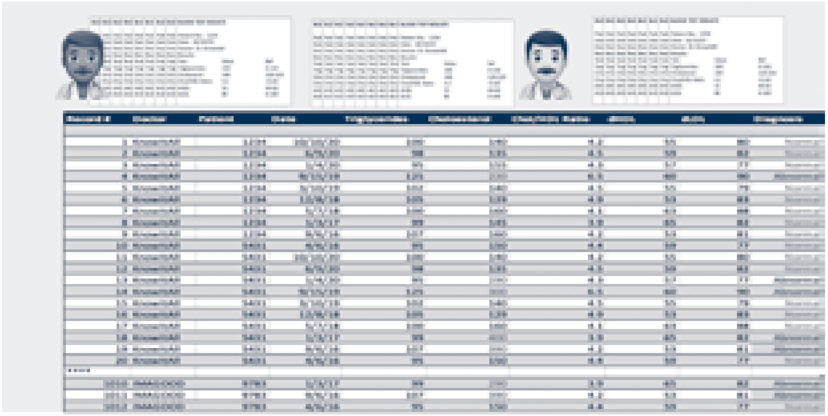
[2] For a more detailed description of how ML works, see Charles R. Macedo, "INSIGHT: What Every Lawyer Should Know About Machine Learning Tech," *Bloomberg Law*, July 20, 2020 (available at *https://news.bloombergtax.com/daily-tax-report/insight-what-every-lawyer-should-know-about-machine-learning-tech-1*).

Figure 3. Illustration Building A Dataset From Samples For Multiple Patients Of The Same Doctor



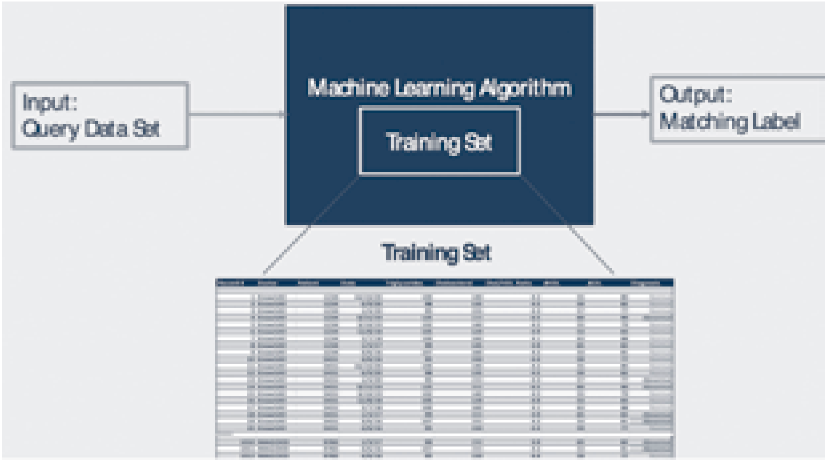
Record #	Doctor	Patient	Date	Triglycerides	Cholesterol	Chol/HDL Ratio	dHDL	dLDL	Diagnosis
1	KnowItAll	1234	10/10/20	100	140	4.2	55	80	Normal
2	KnowItAll	1234	6/9/20	98	135	4.5	59	82	Normal
3	KnowItAll	1234	3/4/20	95	155	4.3	57	77	Normal
4	KnowItAll	1234	9/13/19	125	230	4.5	60	90	Abnormal
5	KnowItAll	1234	5/10/19	102	140	4.5	55	79	Normal
6	KnowItAll	1234	12/8/18	105	129	4.9	53	85	Normal
7	KnowItAll	1234	5/7/18	100	160	4.1	63	88	Normal
8	KnowItAll	1234	1/3/17	99	145	3.9	65	82	Normal
9	KnowItAll	1234	9/9/16	107	160	4.2	53	81	Normal
10	KnowItAll	5431	4/9/16	95	150	4.4	59	77	Normal
11	KnowItAll	5431	10/10/20	100	140	4.2	55	80	Normal
12	KnowItAll	5431	6/9/20	98	135	4.5	59	82	Normal
13	KnowItAll	5431	3/4/20	95	290	4.3	57	77	Abnormal
14	KnowItAll	5431	9/13/19	125	300	4.5	60	90	Abnormal
15	KnowItAll	5431	5/10/19	102	140	4.5	55	79	Normal
16	KnowItAll	5431	12/8/18	105	129	4.9	53	85	Normal
17	KnowItAll	5431	5/7/18	100	160	4.1	63	88	Normal
18	KnowItAll	5431	1/3/17	99	400	3.9	65	82	Abnormal
19	KnowItAll	5431	9/9/16	107	390	4.2	53	81	Abnormal
20	KnowItAll	5431	4/9/16	95	150	4.4	59	77	Normal

Figure 4. Illustration Building A Dataset From Samples For Multiple Patients Taken By Multiple Doctors



Record #	Doctor	Patient	Date	Triglycerides	Cholesterol	Chol/HDL Ratio	dHDL	dLDL	Diagnosis
1	KnowItAll	1234	10/10/20	100	140	4.2	55	80	Normal
2	KnowItAll	1234	6/9/20	98	135	4.5	59	82	Normal
3	KnowItAll	1234	3/4/20	95	155	4.3	57	77	Normal
4	KnowItAll	1234	9/13/19	125	230	4.5	60	90	Abnormal
5	KnowItAll	1234	5/10/19	102	140	4.5	55	79	Normal
6	KnowItAll	1234	12/8/18	105	129	4.9	53	85	Normal
7	KnowItAll	1234	5/7/18	100	160	4.1	63	88	Normal
8	KnowItAll	1234	1/3/17	99	145	3.9	65	82	Normal
9	KnowItAll	1234	9/9/16	107	160	4.2	53	81	Normal
10	KnowItAll	5431	4/9/16	95	150	4.4	59	77	Normal
11	KnowItAll	5431	10/10/20	100	140	4.2	55	80	Normal
12	KnowItAll	5431	6/9/20	98	135	4.5	59	82	Normal
13	KnowItAll	5431	3/4/20	95	290	4.3	57	77	Abnormal
14	KnowItAll	5431	9/13/19	125	300	4.5	60	90	Abnormal
15	KnowItAll	5431	5/10/19	102	140	4.5	55	79	Normal
16	KnowItAll	5431	12/8/18	105	129	4.9	53	85	Normal
17	KnowItAll	5431	5/7/18	100	160	4.1	63	88	Normal
18	KnowItAll	5431	1/3/17	99	400	3.9	65	82	Abnormal
19	KnowItAll	5431	9/9/16	107	390	4.2	53	81	Abnormal
20	KnowItAll	5431	4/9/16	95	150	4.4	59	77	Normal
....									
1010	IMAGOOD	9783	1/3/17	99	400	3.9	65	82	Abnormal
1011	IMAGOOD	9783	9/9/16	107	390	4.2	53	81	Abnormal
1012	IMAGOOD	9783	4/9/16	95	150	4.4	59	77	Normal

Figure 5. How AI/ML Works



- Product development uses a different kind of AI to identify and design key product features and designs. This type of AI may be generative rather than merely identifying a result based upon the query.[3]
- Security systems also use a wide variety of AI and ML systems, such as facial recognition, identity recognition, and security risks detection, among others.

The possibilities and applications of AI and ML systems today and in the future go beyond our imagination and cover just about any type of system that involves large collections of data.

D. Protecting Training Sets

Of course, efforts to protect, defend and license training sets can prove to be difficult to the extent that all that may be included in such training sets is raw data, which the U.S. Supreme Court has made clear is not generally protectable. The basic tools of intellectual property protection that are available, *i.e.*, patent, trademark, copyright, trade secret and contract law, are not always a perfect fit for training sets. A full discussion of how to

[3] A famous example of a generative AI system among patent practitioners is "Device for the Autonomous Bootstrapping of Unified Sentience" ("DABUS"), a patented (see U.S. Patent No. 10,423,875) Artificial Intelligence system created by Stephen Thaler. DABUS, a system of interconnected neural network modules that have been trained in the field of a general endeavor that postulate and test new designs for products and other inventions, such as "warning light" and a "food container." See "ARE Patent Law Alert: DABUS AI: The Test Case For When A Computer Can Be An Inventor On Patents For Innovation It Purports To Create," November 23, 2020 (available at *https://www.arelaw.com/publications/view/are-patent-law-alert-daubus-ai-the-test-case-for-when-a-computer/*).

defend and protect training sets is beyond this article, but available in other writings.[4]

Part II: Representative Contract Considerations by IBM as a Licensor and Licensee

At IBM, we have two main use cases, a licensee (consumer of data) and a licensor (provider of data). In particular, IBM Watson AI is an example of a licensee, and The Weather Channel is an example of a licensor. Due to the different use cases, each of these business units encounter unique licensing challenges and possess different standard contractual considerations.

A. IBM Watson (as Licensee)

As a licensee, IBM Watson AI generally procures data through licensing to train AI models.

IBM Watson AI has a vast portfolio of products and offerings that vary from chat bots to operational flow systems providing real-time feedback for IT assistance. Each of these products and offerings use AI models, which need training sets or data.

There are a variety of data types used to train our AI models. For example, data types may include text (*e.g.,* articles or publications), images, videos, graphical data (*e.g.*, data points on a map) and factual data (*e.g.*, temperatures in New Delhi, India during a particular time period); similarly, the data sources may vary.

From a legal standpoint, there are several key contractual obligations when negotiating for the use of data to protect IBM or our clients, namely (1) rights possessed by the data provider; (2) privacy and confidentiality considerations; (3) accuracy, timeliness and completeness; and (4) limitations/restrictions on data use.

1. Rights Possessed by the Data Provider. One important distinction is whether the data provider is a data generator or data procurer. A data generator (or data owner) is a person or entity that generates the data. As the original owner of the data, the data generator would possess the appropriate rights to license or sell the data. A data procurer, however, is a person or entity that may obtain the data from another person or entity. As such, it is important to ensure that the data procurer has sufficient rights to license the data to a third party, which may require due diligence into whether the data procurer has the proper authority from the original owner to license the data to a third party. If the data procurer does not have the appropriate authority to license or sell the data to a third party, then IBM Watson AI cannot obtain data from that source.

In sum, regardless of whether the licensor is a data generator or a data procurer, IBM typically requires identification of the data source and confirmation of the rights to the data possessed by each source. This requirement serves to protect IBM and its clients from any infringement claims related to that particular data.

2. Privacy and Confidentiality Considerations. We also consider whether the data includes personal information or confidential information. In general, the data may be anonymized, de-identified, or pseudonyms may be used in lieu of personal information. Even though any of these methods for handling personal information may be implemented by the data provider or IBM, it is preferred that the data provider implements these methods before the data is provided to IBM, thereby removing any privacy or confidentiality concerns.

3. Accuracy, Timeliness and Completeness. The accuracy, timeliness, and completeness of the data is also considered. As we all know with training AI models: junk in, junk out. If the data is poor quality, then the AI output will be poor quality. Therefore, the accuracy, timeliness and completeness of the data is crucial.

4. Limitations/Restrictions on Data Use. Ideally, a licensee would prefer to have a minimal amount, if any, of restrictions/limitations on the use of the data. However, depending on the intended use of the data, some limitations/restrictions may be less problematic than others. Several common limitations/restrictions on data use include share-alike, derivative works, and commercial uses.

- **Share-Alike:** Share-Alike is commonly included in Creative Commons licenses (*e.g.*, CC-BY-SA 4.0). In a license with a share-alike restriction, the licensee must share the data under the same license as the original work. Such a restriction may conflict with the licensee's standard license terms. As such, it is good practice to negotiate to remove the share-alike restriction to prevent any such conflicts or opt to not use data with a share-alike restriction (or similar restriction) attached to the data.
- **Derivative Works:** If the data may be modified, used to create a derivative work, or may be included in a transformative use, then it is good practice to negotiate to remove the derivative works restriction or to opt to not use the data.
- **Commercial Uses:** As a commercial entity, IBM will most likely use the data within a commercial product or offering. Therefore, we negotiate to remove any restriction prohibiting the use of the data for commercial purposes from the license; otherwise, we will opt to not use the data.

B. The Weather Channel (as Licensor)

As a licensor, The Weather Channel typically licenses the data as "Weather Data Packages" to its clients or

[4] For a full discussion, see Charles R. Macedo, "Protecting Artificial Intelligence Innovations as Intellectual Property: Opportunities and Pitfalls," *Practical Law* (September 2020) (available at *https://www.arelaw.com/publications/view/charley-macedo-authors-practical-law-article-on-protecting-artif/*).

other third parties for use in the client or third party's individual offerings or products. Weather Data Packages are a set of packages that are tailored to the client's brand, particular data consumption, or intended uses by a particular client.

Unlike IBM Watson AI, the data types used by The Weather Channel are generally facts derived from third-party licensed data, data generated by a government entity, or IBM generated data.

Since The Weather Channel takes its reputation and the quality of its data very seriously, the contractual considerations used in any license of data to a third-party and/or client is very important.

Several important contractual considerations include (1) warranty and indemnification to licensee; (2) prohibition on sale of data; (3) restrictions on possible licensees; (4) exclusivity of data provider; and (5) restrictions on usage.

1. Warranty and Indemnification to Licensee. Generally, data is provided as is, with no representations or warranties. The Weather Channel typically will not provide any indemnification to the licensee.

2. Prohibition on Sale. Data provided by The Weather Channel is licensed, not sold. As a result, The Weather Channel is the owner of any derivative works associated with the use of the data provided.

3. Restrictions on Possible Licensees. The Weather Channel typically does not license to any competitors.

4. Exclusivity of the Data Provider. With the exception of any public entities, The Weather Channel is typically the only data provider to its clients. Even though a client may use data from a government entity, a client is prohibited from using data from a private entity.

5. Restrictions on Usage.

- **Internal Use Only.** We generally do not allow clients to provide the data or redistribute it to another third party. They are expected to use it for their internal use or for non-commercial purposes only.
- **Restrictions on Modifications or Alterations.** If the client is providing the data in a third-party application, we have strict limitations on how the data is modified or altered thereby preventing any modification that may be misleading or confusing. Such restrictions on modifications and alterations are intended to ensure that the data is accurate, complete, and timely for all of our clients.

Part III: How Getty Images Transformed Its Stock Photo and Video Library into Licensable Training Sets

At Getty Images, we think about monetizing our data for usage in AI and machine learning. Getty Images has a lot of both pictures and videos, and their use in AI and machine learning is something that is rapidly evolving. This new field allows us to pivot and explore a new way to become a licensor while supporting our contributors and those who provide us with content, as well as supporting our customers' objectives, needs, and requirements.

For 25 years, Getty Images has been the market leader in premium, exclusive, and creative content, which are provided through our Getty Images and iStock brands. Creative content is content that has been shot with models who have been released for their likeness in use of commercials and advertising. Editorial content includes news, sports, entertainment, and historical archives from things that have happened in years gone by. Despite our name, Getty Images is also one of the leading suppliers of video content. This has been a huge area of growth, especially with the acceleration of the digital world during the pandemic.

One of the systems we utilize on our photos is specific image IDs. These are the numbers in the corner of each photograph that can be used to source that asset and assist in searching for it. Next to the image ID is typically the name of the collection or the photographer that asset would come from. This system provides an easy way for our customers to find the pictures and a way for us to index and archive those images.

Our core licensing business has almost a million direct customers through which we license content for use in their traditional projects; however, the opportunity for licensing data for machine-learning projects is growing exponentially. We are continually exploring licensing models that offer value and protection to our customers as well as compensate contributors for the use of their work in AI/ML applications.

Oftentimes, there is legal uncertainty around acquiring data. There are open questions that relate to intellectual property, privacy, and third-party rights. Indemnifications and model releases are the primary methods for offering these kinds of protections.

Encouraging innovation in artificial intelligence is the key to the development of data licensing markets that will power AI and machine-learning models. Transparency and the maintenance of auditable data records are important as well. This essentially details how the data can be attributed back to the content owner.

There are many ways that image or video data could be acquired:

- One method is "scraping," which is downloading and saving images from websites. The risks associated with this include poor quality images and videos; no curation, which means it has not been sorted and there is no associated metadata or labeling. In addition, scraping has a wide variety of potential liabilities, including copyright infringement. Therefore, it is not the best practice.

- Open source is another method, wherein there are creative commons licenses for some data sets. These are often used by educational institutions and those in the first stage of training their models. However, several problems arise here, the first being that they tend to be limited-use cases. The second is that there are often multiple users of the same data set, meaning all models have the same output because they have had the same exact input. Finally, these lack in diversity and volume.
- Another way that a customer might find data is to shoot it themselves using their own cameras or network of photographers. However, this can be expensive and time consuming.
- Accessing a database of images and videos, such as is provided by Getty Images. We can also shoot content to meet project requirements.

To address a few applicable use cases, we have several examples:

- **Autonomous Vehicles.** In the early stages of their learning they may require an understanding of where the road is. This then progresses to learning about things like pedestrians and road signs.

  One tool that is utilized is known as bounding boxes, as seen in Figure 6A. These are green boxes that are placed into an image to help locate where specific objects or elements of a picture are. There are other ways of preparing data as well, such as segmentation or localization of objects. Each of these help the machine understand what they are supposed to be looking for.
- **Facial Detection.** The use of faces, one example being facial detection, where a face appears in an image or video and is located. These could be identified with the bounding box or even with facial landmarks. Facial expressions can also be detected. While some of these can be quite subjective, we can tell you how extreme the emotion may be.
- **Object Recognition.** Objects, such as various types of buckets as shown in Figure 6B, can be identified with AI technology and easily located using a keyword search.
- Popular places and scenes, such as famous landmarks.
- Activities and understanding the context of these activities.
- Misinformation, which equates to our editorial imagery. This includes images that have been changed from their original context or their original form.

We take a multi-modal approach, utilizing the relationship between the image or video itself and the caption or metadata keywords that are associated with it. For example, when searching through our 690,000 stock images of cycling in some form or another, one could narrow it down through the association of how many people are in the image, where it was shot, how it was shot, and many more parameters. Typically, these metadata tags help users navigate the library of images and find content more quickly, but those same metadata tags can be used to train their ML models.

These navigation tags help users identify the content and find it. The AI services are also used internally to help customers find content visually. Our materials include traditional photography (still images), video content, as well as 360-degree images and 360-degree video. There are typically over 20 contextual keywords or tags per image or per video, which are added both by humans and AI technology. These keywords are provided both in-house for images that we own and through our contributor network when those are being uploaded. We typically have over 20,000 new assets added per day, which is growing exponentially,

Figure 6A. Bounding Boxes

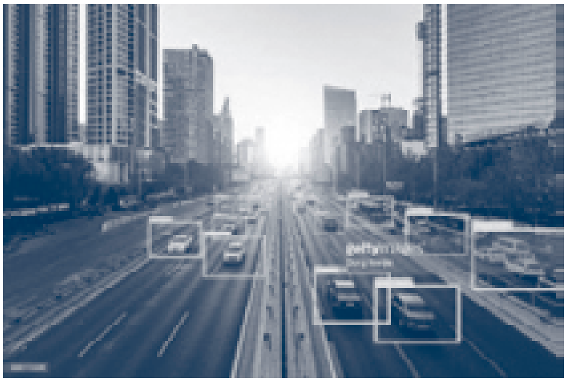
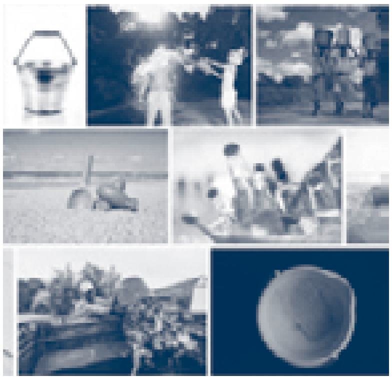


Figure 6B. Object Recognition (Buckets)



and our search engine works in over 15 languages. Thus, training AI in a different language is possible.

The packaging and delivering of our content is also a key aspect. Getty Images has industry-leading API-based connections, which allow our customers who use AI and training set data to be able to receive that content at scale quickly and effectively.

We can create training data sets from existing or from new content. We can also supply different types of annotations or different ways of preparing that data. That could be supplying the caption data, for example, along with the image itself, providing different keywords in different languages, or it could be some kind of visual representation, such as locating a face within an image.

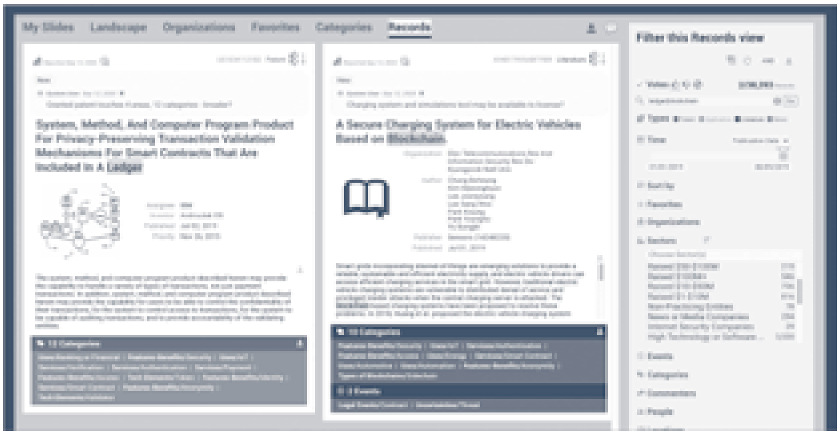
Part IV: Perception Partners's Use of ML in Real World Products

Perception Partners is an intellectual property (IP) analytics platform provider that trains machines with text and image big data for competitive intelligence and online brand protection. Perception Partners has evolved its ML approach by productizing human consulting services (over more than 15 years) using expert systems, natural language processing, algorithmic scoring, predictive classification, deep learning and design-centric visualization techniques. These AI-centric processes continuously train on large volumes of data from patents, technical literature, news, litigation and market documents. The process outputs are cleaned, organized and categorized competitive landscape reports delivered on-demand so that busy decision-makers can build consensus with meaningful and visual business stories, without being data scientists themselves.

A. Intellar® Dashboards

The Intellar Dashboards platform curates competitive intelligence for IP, R&D and M&A decision-makers in the cloud, to efficiently map who is (or will be) doing what, when and where in emerging technologies. Intellar derives insights from big data document collections that are mined from dozens of global sources. Intellar displays relevant records by color-coding results for each document type, illustrated monochromatically in Figure 7. For users, the card on the left is coded green in the Intellar display to designate a patent document record and the card on the right is coded violet to designate a technical literature document record.

Figure 7. Intellar® Dashboards - Records View

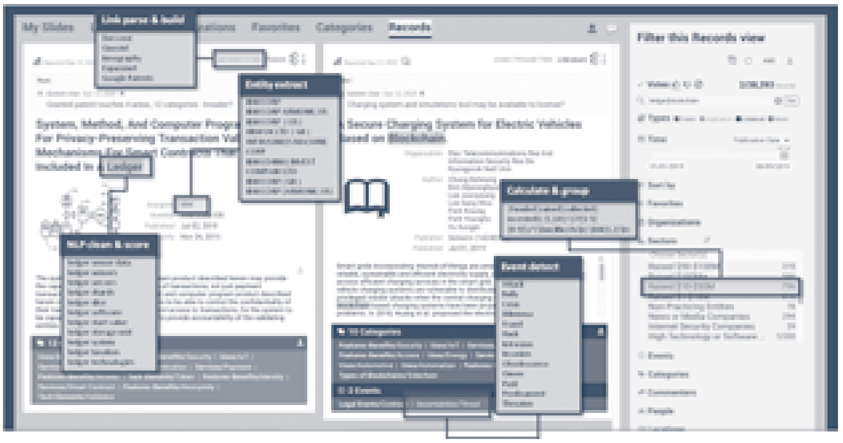

My Slides
Landscape
Organizations
Favorites
Categories
Records
System, Method, And Computer Program Product For Privacy-Preserving Transaction Validation Mechanisms For Smart Contracts That Are Included In A Ledger
A Secure Charging System for Electric Vehicles Based on Blockchain
Filter this Records view
Values
Types
Time
Sort by
Favorites
Organizations
Sectors
Events
Categories
Connections
People

Intellar's visual treatment of heterogeneous data types helps users rapidly identify meaningful competitive signals at a glance, no matter where the signals appear in documents. Intellar utilizes open source and proprietary information that we—and our clients—have mutual rights to use.

The primary value proposition of Intellar Dashboards is to provide shared visualization and interpretation of topics, entities, events and trends – all within understandable landscape views that can be drilled down to individual source records. To create trusted landscapes, Intellar applies machine learning across every bit of data and metadata it can access.

As seen in Figure 8, Intellar's trained machine learning can aid in providing critical landscape insights from individual records. To indicate who is important, Intellar extracts and normalizes entity names from curated records (*e.g.*, corporate parent, subsidiary and acquiring organization names). These names are often misspelled, noisy, obsolete or omitted altogether, depending on the data source (especially non-patent sources), making learned predictions of proper names quite useful. To direct users to the full text of relevant documents (Intellar mines full text, but only shows excerpts), hyperlinks are dynamically generated to connect to sources favored by the user, even if Intellar parsed the record from another source (*e.g.*, an alternative patent database). To provide market and transactional insights, money amounts can be calculated (from text and numeric data) so that records are grouped into quantitative ranges that permit filtering by associated valuations. To enable business trend detection, events can be captured from textual and semantic terms in the underlying records that indicate material, *e.g.*, financial, market, legal, people, research and risk situations. And to better aid topic analysis, key

Figure 8. Intellar's Dashboards With Annotations Showing Types Of Trained Machine Learning


My Slides
Favorites
Categories
Records
Link parse & build
Entity extract
Calculate & group
NLP clean & score
Event detect
Filter this Records view
System, Method, And Computer Program For Privacy-Preserving Transaction Validation Mechanisms For Smart Contracts That Are Included In a Ledger
Secure Charging System for Electric Vehicles Based on Blockchain
10 Categories
2 Events

concept phrases can be captured, cleaned and scored using natural language processing to cluster relevant results for retrieval and analysis.

Intellar categorizes records into taxonomies for technologies and events. Taxonomies are displayed in interactive, hierarchical visualizations that depict topical clusters by color-coded areas, categories and document types. In Intellar's sunburst diagram (Figure 9A): areas are shown in the inner ring; categories are shown in the middle ring; and document types appear in the outermost ring. The taxonomic relationships in the sunburst can also be displayed as a tree (or mindmap) diagram, as illustrated in Figure 9B. A trained record classifier—with results consistently checked by experts—assigns document records into the hierarchy of the taxonomy. By involving experts in checking classifier output throughout training, new synonyms, patterns and terms can be incorporated sooner for more precise results with higher recall on new datasets.

Intellar monetizes competitive intelligence training data by learning the who, what, where, when and how that can be mined from one or more related document records. Intellar—with the help of experts—is constantly comparing new results to prior, while predicting future results in context of deeply learned relationships between entities, topics, events, categories and valuations.

B. Replicatch® Online Brand Protection

In addition to training machine learning to analyze multi-lingual text and numeric patterns, Perception Partners also trains neural networks to classify, match and respond to visual patterns detected in images. Their Replicatch brand protection software platform detects IP infringement by matching IP disclosures to product images. These learned relationships help indicate probable counterfeits and knockoffs online and trigger actions – from risk alerts shared with clients and counsel, to narrative takedown letters sent to offending web hosts. Asserting IP infringement using image-based evidence goes beyond language barriers and may be perhaps the most effective way to remove something problematical from the Internet anywhere, sometimes overnight. Training the Replicatch platform to match any design, image or mark is not trivial and remains a work in progress (with a near-term focus on highly counterfeited items).

Replicatch trained image matching also holds promise for monetizing IP information for purposes other than infringement detection. For example, in media and entertainment, a patented, copyrighted or trademarked design can be detected in video frames (as illustrated in Figure 10). By using IP information to identify products, rights and brands embedded in video content, subsequent advertising and marketing actions can occur such as: offering more information, enabling a user to "follow" a related offering or entity or triggering a purchase.

Figure 9A. Sunburst Diagram

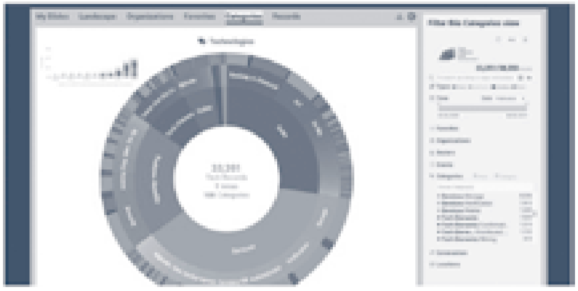
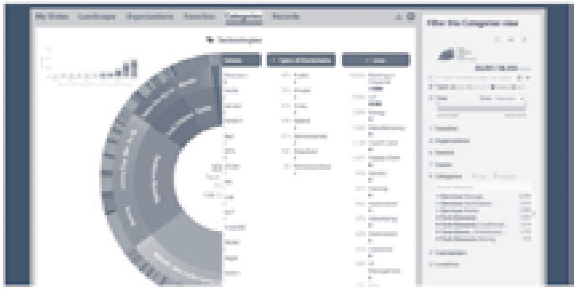
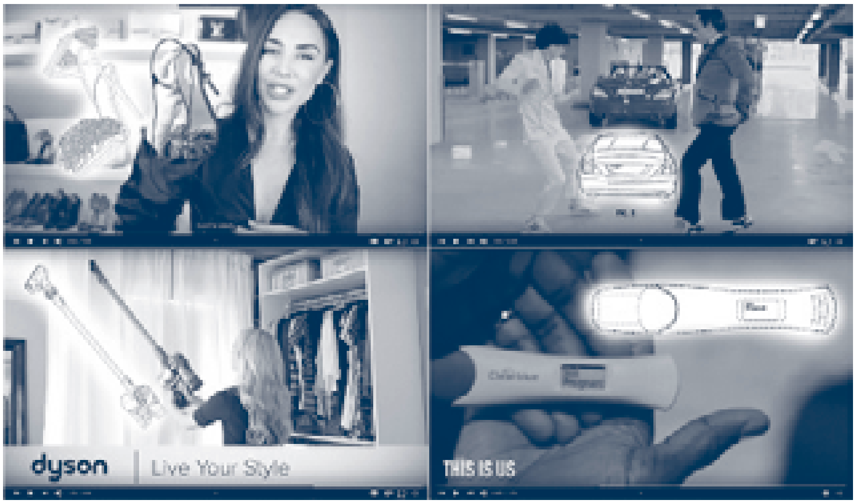


Figure 9B. Equivalent Tree



In the future, trained AI models will be able to learn continuously from designs, copyrights and trademarks as they appear online in near-real time. The ability to link photo and video content to IP assets and their respective owners is a fascinating new paradigm. IP data will be monetized by training AI to help all brand owners extract more value from their protected and recognizable creations during any digital experience.

Figure 10. Replicatch Content Monetization—As Seen On TV



Conclusion

Every day, more and more industries continue to develop and implement AI and ML systems to automate previously manual processes. These processes promise to assist users with better and more efficient decision making, as illustrated herein. Existing data troves can be repurposed as is the case with The Weather Channel sensor data and Getty Images photo and video files. By mixing and matching data from various sources, new and amazing systems can be developed, as IBM demonstrates with Watson, and as Perception Partners demonstrates with Intellar and Replicatch. ■

Available at Social Science Research Network (SSRN): *https://ssrn.com/abstract=3897882*.