White Paper

An Introduction to Machine Learning Training Data

appen™

Table of Contents

This white paper is intended for executives who have either already invested or are planning to invest in machine learning within their organizations.

Introduction

Computers need to be told what to do. That's been the traditional wisdom. They only know 0s and 1s, yes or no, on or off. They can't do anything without being instructed or programmed.

But artificial intelligence (AI) has ushered in a new paradigm, one in which computers learn without being explicitly programmed. Powering this AI is machine learning, which is the practice of feeding computers a huge amount of training data that the computers use to find patterns. These patterns help computers identify the correct response to various situations.

Today, we are seeing an explosion in AI applications, and a corresponding demand for machine learning solutions. Why? Because we are asking computers to solve increasingly complex problems that mimic human functions like speech and sight. These problems are more complex than the mathematical models that traditional programming can handle.

AI requires machine learning, and machine learning requires data—a lot of the right kind of data. Without it, no project can get off the ground. In fact, in a recent study from Oxford Economics and ServiceNow, 51% of CIOs cite data quality as a substantial barrier to their company's adoption of machine learning.[1]

Appen can help. We've been in the data business for over 20 years and have developed deep experience working with leading global technology companies, governments, and other organizations, across a variety of data types. We collect and annotate speech, sound, image, video, and text data, and use it to fuel our clients' many machine learning projects. We also review and annotate data from live products to improve products and their user experience.

We created this white paper for business executives embarking on—or looking to improve—their own machine learning project, to share a few guiding principles about your data quantity, quality, and sources.

[1] The Global CIO Point of View, October 2017

The Importance of Quantity: Why Machines Need a Huge Volume of Data to Learn

The first thing to know about machine learning data is that you need a lot of it. Remember, machine learning helps computers solve problems that are too complex for an algorithm alone. What makes these problems complex? Often, it's the amount of inherent variation—there are hundreds, thousands, or millions of variables. And the resulting system must be able to cope with them all.

Think of machine learning data like survey data: the larger and more complete your sample size, the more reliable your conclusions will be. If the data sample isn't big enough, it won't capture all the variations or take them into account, and your machine may reach inaccurate conclusions, learn patterns that don't actually exist, or not recognize patterns that do.

Take a speech recognition system, for example. Spoken languages and human voices are extremely complex, with infinite variations among speakers of different genders, ages, and dialects. You could work with a mathematical model to train a machine on textbook English, but the resulting system would likely struggle to understand anything that strays from the textbook: loose grammar, people with foreign accents or speech disorders, and those who use slang, jargon, and filler words or sounds like "ah" and "um." If you were employing that system for email or text, it would also trip up on the emojis and abbreviations (such as LOL) that appear in typical chat sessions. You would have spent a lot of time and money on something that would utterly fail in the market.

The more your machine learning data accounts for all the variation the AI system will encounter in the real world, the better your product will be. Some experts recommend at least 10,000 hours of audio speech data to get a recognizer to begin working at modest levels of accuracy.

The more your machine learning data accounts for all the variation the AI system will encounter in the real world, the better your product will be.

This same principle applies to new and established products alike. You need a lot of data to get to market with the best AI solution you can make, as well as to improve and update it. Search engines on retail websites, for example, need constant training to keep up with changing inventory: adding new products, removing discontinued products, adding and removing seasonal items, etc. To ensure customers see relevant results, it's critical to regularly tune the onsite search algorithm.

The Importance of Quality: Why the Right Data Is Essential for Success

Machine learning requires a huge volume of data. It also requires the right kind, because ultimately the system will do what it learns from the data. You can have the most appropriate algorithm, but if you train your machine on bad data, then it will learn the wrong lessons, come to the wrong conclusions, and not work as you (or your customers) expect. On the flip side, a basic algorithm won't hold you back if you have good data (and enough of it, of course). Your success, then, is almost entirely reliant on your data.

What defines "bad" data? Many things. The data may be irrelevant to your problem, inaccurately annotated, misleading, or incomplete.

You can have the most appropriate algorithm, but if you train your machine on bad data, then it will learn the wrong lessons, come to the wrong conclusions, and not work as you (or your customers) expect.

Consider search engine evaluation. To improve a search engine's performance, Appen works with human judges to rate how good each result is for a particular query. For example, if you were searching to find the depth of the Grand Canyon and got two results, one from the U.S. National Park Service and one from a local tour guide, both of which had the answer, which would you trust more? Most people would trust the National Park Service more, because the site itself looks more trustworthy, or because the Park Service isn't trying to sell anything. This comparison illustrates why preference and relevance is important, and the value of that additional level of human discernment. A computer can find the data—the depth of the Grand Canyon—but doesn't know which source is better unless its told.

Similarly, a query for "cotton cardigan sweater" may get some results for cotton sweaters, some wool sweaters, some pull-over styles, and some cardigans. People evaluating those results must ask themselves, "What was this person looking for, and did they get it?" The judges then rate each result, giving higher marks to the cotton cardigans than the wool ones, and pull-overs of both materials.

If the evaluators don't interpret the original intention correctly, then they train the search engine to return bad results instead of good results. In this example, qualifying your evaluators is key to ensuring you're creating high-quality data.

For speech and pattern recognition, "bad" data might be incomplete or inaccurate. For example, if the machine thinks the sound of someone saying the word "cat" corresponds to the text of the word "rat," that's going to create a frustrating user experience for someone trying to order cat food from a home assistant.

It's important to note that the data for this example is not just the speech samples. The machine also needs to know what, exactly, was said, which comes from human annotation, or people transcribing the speech. So that the machine correctly learns the correspondences between written and spoken words, it's also useful to tell it about the speaker who made the sounds: gender, age, dialect, etc.

With speech, as well as other data, machines need to learn to differentiate important information from non-important. For instance, the sound of someone breathing in sharply is similar to a "ch" or "sh" sound. In this case, a machine ends up being able to recognize what to pay attention to and what to ignore by the context of the sound, not necessarily the sound itself.

It's a substantial task to gather and prepare a sufficient amount of suitable machine learning data. Fortunately, it's much easier if you know what you're looking for—and where to get it.

Sources: Where Machine Learning Data Comes From

Where will your data come from? Broadly speaking, there are four main sources: real-world usage data, survey data, public data sets, and simulated data.

Real-World Usage Data

When your AI products are already in-market, real-world data from actual users is a great resource. With a search engine or search feature, for example, you can look at queries, total results, which results people click on, and what they look at and purchase. Social media sites can gather data about what users post, like, share, and comment on. Speech recognition solutions from smartphones, in-car systems, or home assistants can collect spoken queries and the machines'

responses. There's also broadcast data from music services and sites like YouTube that may track what people look at.

The benefits of using real data are that you know it accurately reflects how people use your system, and you don't have to pay to create it.

However, there are legal questions associated with collecting it, as well as privacy concerns. Some companies have had trouble collecting this data and faced lawsuits when they overstepped. To do it right, you must allow users to opt-in to sharing it. Also, because you don't have control over what people say or do, you may have to collect and process way more data than you need to get the training data you're looking for. Additionally, with real data, you'll always be making an educated guess about context: what users truly intended, why they click on some things and not others. You can't go back and ask users what they meant, but you can ask a vendor like Appen to train a crowd of people to go through and label it, turning it into useful machine learning data.

Also, data gained in the real world requires labelling, annotating or transcribing to be really useful, and that's an extra step.

Survey Data

The second source for machine learning data is surveys. You go directly to your users, or prospective users, and ask what they like or don't like, and what you can improve about the product.

This approach gives you data from actual users, and gets around privacy concerns and legal issues as, by taking the survey, people are opting to participate. Surveys provide context and the opportunity to follow up on anything that's unclear. You also have some control over what people say and do in that you can direct them to the specific topics you want to address.

On the other hand, survey data is somewhat unreliable, because what people say they do and want on your survey might be quite different than what they actually do and want. Additionally, survey data is often skewed toward dissatisfied users, as people who get what they want are less motivated to provide feedback.

Public Data Sets

There are a number of different types of public data sets available from search engines, social media, Amazon Web Services, Wikipedia, universities, data science communities, and other data repositories. There's also an enormous amount of public data from academic efforts in speech

and language processing from the last 40 years, licensable from various organizations. Appen has an extensive catalog of off-the-shelf, licensable linguistic resources for text-to-speech, speech recognition and other systems that rely on speech. For most commercial purposes, affordability is the real advantage of these data sets. This kind of data is often used for creating or refining commodity technologies, like basic language recognition or machine translation.

Engineered or Collected Data

The fourth main way to collect quality data is to make it yourself. This is often the only way to proceed with a new solution, when there aren't any users or usage data yet. You can simulate the user experience by hiring native speakers and professional raters, gathering and annotating the data your project specifically needs. You can mimic the conditions where people will use your product: driving in a car on a city street, calling into a call center, etc.

On one hand, you can get exactly what you need faster this way because you're in control. You always know the context. You can follow-up with your raters and speakers if there's a question. And, since you're not using real data, there are no legal or privacy concerns. Most important, your model will produce a better end result.

On the other hand, this type of data collection will require a larger investment. To do it well, be sure you work with an experienced data collection vendor. There's a lot of management involved to ensure you're getting the right kind of data—and not just the sample itself, but also the metadata. This annotation, like the same done for real-world data, also requires qualifying crowds of people to ensure they can label and categorize the data, allowing machines to know what to do with it.

If you spend the time and money to build a custom database solution, it would be a waste to end up with messy data from an inexperienced vendor.

Simulated data is also not something most companies should attempt themselves. For training a voice-activated, in-car infotainment unit, for example, an auto manufacturer could drive around in every car it makes with dozens of speakers to record the data—and do this for every language in which the unit will operate. Or, the manufacturer could hire a data collection vendor to create a set of modular data that can be mixed with different types of road noise and cabin impulse response

measurements (how sound carries through different spaces) to simulate many real-life conditions. That approach is much more cost effective. Again, the process is easier and faster if you know how to create efficiencies—which experienced vendors, like Appen, do.

The Appen Advantage

Appen understands the complex needs of today's organizations. For over 20 years, we've worked with our clients to develop data collection and annotation programs that meet their specific needs. Our unique approach of working with curated crowds allows us to ramp up quickly and provide a diverse range of participants, spanning hundreds of languages and locales. Appen also has a reputation for delivering the highest-quality linguistic data services in the business. Over the years, we've worked in over 180 languages and dialects.

Two Decades of Experience

With decades of experience working all over the world, Appen ensures you'll have the right amount of the right data to fuel your machine learning. We have a huge range of experience with different kinds of data collection and annotation for clients ranging from government agencies to 8 of the top 10 leading technology corporations. That experience helps us plan well, deliver superior data, and always follow through.

Our experience also means we have the knowledge and flexibility to partner with our clients in designing an approach that fits their needs, timeline and budget.

Curated Crowds

Appen is a big believer in people—and lots of them. We rely on curated crowds to do our data collection and annotation work. Curated crowds are consistent groups of people who have been trained, monitored, and coached to complete tasks efficiently and accurately according to quality guidelines.

We find curated crowds the best fit for our data work for a number of reasons: targeted expertise, demographics, language, broad reach, high quality, cost-effectiveness, ability to scale up or down, and long-term engagement.

If you take the approach of hiring a supervised group of employees, they typically all meet at one location to work, which limits their perspectives and experience to certain locations. This differs from Appen's approach. We assemble our curated crowds from our network of over one million people all

over the world. This way, we can easily meet diversity requirements and maintain unique viewpoints or contributions.

Working with curated crowds also ensures high-quality results. We provide guidelines for our crowd, specific to each project, and in some cases each market, so that the workers understand how to perform specific tasks. Part of our strength in this area is in creating these guidelines and working with our clients to continually improve them to ensure the best possible results. Curated crowds become experts on our tasks, increasing quality through continuous engagement, access to resources, and multiple quality-management metrics. One example of quality management is our sophisticated audit system to spot-check data as it's collected, so that if we're not getting the kind of data we need, we can quickly course-correct by working with individuals to help them understand or updating the entire crowd with new information and instruction materials.

Additionally, we take a very human approach. Unlike pure crowdsourcing services that offer work to anyone willing to do it, we take care in our recruiting. We look for people comfortable with the technology they'll be working with, great attention to detail, and the ability to quickly deduce what's going on and annotate it. Various levels of vetting are used to ensure people are qualified and best matched to a project.

For our crowd workers, we offer flexible work opportunities—as evidenced by **FlexJobs naming Appen** as the number one provider of remote jobs in its annual list of the **100 Top Companies to Watch for Telecommuting and Remote Jobs**.

Strong Global Coverage

Appen has professional people on the ground in over 130 markets, and curated crowds distributed around the world.

If you want to improve your product in Azerbaijan, France, or Vietnam, we don't use a US call center staffed with Azerbaijani, French, and Vietnamese speakers. Appen can give you feedback from people on the ground in those countries, and many others, who know the language, culture, current events, and what local people are interested in.

If you're localizing a product for a new market, it takes more than just collecting new language data. You also need expertise in local culture and circumstances to make your product as relevant as possible for your target audience, and to ensure the project goes smoothly. Many vendors have a good track record in a few countries but will struggle to scale up quickly elsewhere. Appen has project managers based all over the world, including Africa and the Middle East, who organize the logistics of our operations.

Conclusion

Investment in AI in 2016 was in the range of $26 billion to $39 billion.[2] Yet, even with the huge advances made in AI solutions in the last decade, and the growing number of them on the market and in our lives, this basic truth holds: AI is only as good as the machine learning data that trained it. To build a successful solution, you need the right data, and a lot of it.

There are several options for sourcing that data, each with its own benefits and drawbacks. No matter which you decide is right for you, an experienced data collection vendor can help you find the most efficient path to collection and implementation. Among data vendors, there's a clear choice: Appen.

AI is only as good as the machine learning data that trained it.

We have over 20 years in the industry partnering with leading commercial and government technology organizations to improve their existing solutions, develop new offerings, and expand into global markets. With access to an experienced crowd of over one million people worldwide, and a team of experienced project managers, we look forward to scaling up your data-collection efforts.

Ensure that all your AI initiatives reach their full potential by working with Appen to source the high-quality data you need. Contact one of our experts today to discuss your specific requirements at **hello@appen.com**, or visit our website at **appen.com**.

[2] Artificial Intelligence: The Next Digital Frontier? McKinsey, June 2017

About Us

Appen is a global leader in the development of high-quality, human-annotated datasets for machine learning and artificial intelligence. With over 20 years of experience, expertise in more than 180 languages, and access to a crowd of over 1 million worldwide, Appen partners with global companies to enhance their machine learning-based products.

12131 113th Ave NE Suite #100
Kirkland, WA 98034

Toll-free inside the US: + 1 866 673 6996
From outside of the US: + 1 646 224 1146