White Paper

appen®
data with a human touch™

How to Develop a Training Data Strategy for Machine Learning

Table of Contents

Introduction 3

Budgeting for Your Machine Learning Program 5

Sourcing the Right Data for Your Use Case 7

Choosing a Data Labeling Platform 12

Ensuring Data Quality 13

Ensuring Data Security 17

Why Outsource Data Collection and Annotation 19

Conclusion 21

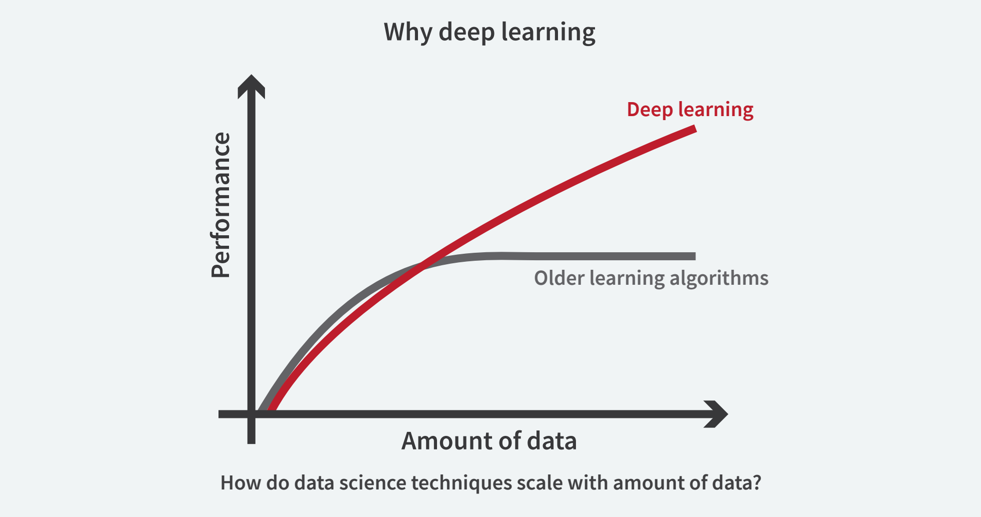
appen®
data with a human touch™

The purpose of this white paper is to help organizations create a strategy for developing training data to build and improve artificial intelligence systems, from determining budget, to managing quality and throughput, to dealing with data security.

Introduction

Artificial intelligence (AI) systems learn by example, so it stands to reason that the more examples they have, the better they'll learn. The higher quality those examples are, the better still. Insufficient or poor training data can result in an unreliable system that reaches the wrong conclusions, makes poor decisions, can't handle real-world variation, and introduces or perpetuates bias, among other problems. It's also expensive. IBM estimated that poor data quality in the United States costs the country's economy roughly $3.1 trillion per year.[1]

Data—the quality and quantity of it—can make or break your machine learning project. That's why it's crucial to have a clear data strategy in place before you begin your AI initiative. Without a well-defined strategy for collecting and structuring the data you need to train, test, and tune your AI systems, you run the risk of delayed projects, not being able to scale appropriately, and ultimately, competitors outpacing you.



Source: "The Unreasonable Effectiveness of Data," Alon Halevy, Peter Norvig, and Fernando Pereira, Google, 2009

[1] Harvard Business Review, "Bad Data Costs the U.S. $3 Trillion Per Year," September, 2016

Machine learning is based on algorithms that crunch through all that training data, and with traditional machine learning, there has been some debate about whether the algorithm or the data is more important.

With newer deep learning models based on artificial neural networks—developed to take advantage of deeper models, increased compute power, and access to more and varied data—the data has clearly taken the central role. In traditional methods, as shown in the chart here, the algorithm will hit a performance plateau. With deep learning, more data improves performance at scale.

In addition to quantity, machine learning also requires high-quality data. Your system will only be as good as the data it trains on. If you use irrelevant, inaccurate, misleading, incomplete, or biased data, the resulting AI will reflect one or more of these problems.

> **“Only 18% (of businesses) currently have a clear strategy in place for sourcing the data that enables AI, now is the time to define the way forward for your organization.”**

If you don’t think data quality is a challenge for your project, consider that a 2017 Harvard Business Review study [2] revealed that 47% of newly created data records have at least one critical error and only 3% of companies’ data meets basic quality standards.

This white paper covers how to plan your training data strategy, including budgeting, options for data sourcing, ensuring quality and security, and how outsourcing the collection and labeling of training data can help scale your AI initiatives. With 71% of companies expecting their AI investments to increase, but only 18% currently having a clear strategy in place for sourcing the data that enables AI [3], now is the time to define the way forward for your organization.

[2] Harvard Business Review, “Only 3% of Companies’ Data Meets Basic Quality Standards,” September 2017

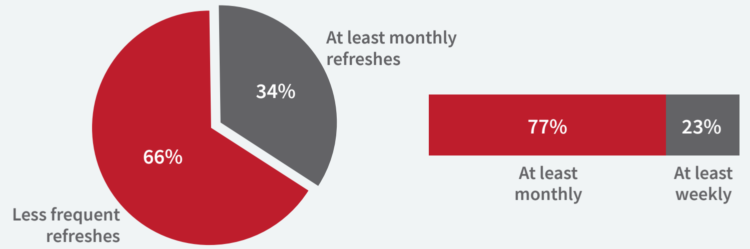
[3] McKinsey & Company, “AI adoption advances, but foundational barriers remain,” November 2018

Budgeting
For Your Machine Learning Program

When launching a new machine learning project, the first thing to define is the objective you’re trying to achieve. That will tell you what type of data you need and how many “training items,” or data points that have been categorized, you’ll need to train your system. Once these have been determined, you can then evaluate your options for sourcing data, and calculate a budget.

As an example, training items for a computer vision or pattern recognition project would be image data that has been labeled by human annotators to identify the contents (trees, stop signs, people, cars, etc.). The number of items you’ll need can range from several thousands to millions of images, depending on the complexity of your computer vision or pattern recognition scenario. If your project is to sort images of products into several broad categories (e.g., shirts, pants, shoes, socks, dresses), you may only need several thousand images to train it. For a more complex project, such as sorting images into thousands of specific categories (e.g. men’s running shoes, women’s fashion heels, baby shoes, etc.), the system might require several million correctly labeled images.

AI Applications require regular data refresh



- **One third of AI applications require frequent, monthly data updates**
- **One quarter of those require weekly updates**

Source: McKinsey & Company, “Notes from the AI Frontier: Insights from Hundreds of Use Cases, April 2018

It’s important to be clear-eyed about the time and money investment required to get your AI initiative off the ground, maintain it over time, and evolve the features and functionality — along with your business — so the solution stays relevant and useful to your customers.

Additionally, plan for your pilot program to be iterative, with at least two rounds of revisions. No machine learning project gets everything right in the first release. That's not a failure; it's just the nature of these projects. Machine learning models must be trained, then tested, then tuned.

We recommend that clients launch with a small subset and analyze what comes back, then adjust the instructions or other parts of the experimental design to improve the next release. One adjustment is enough for most projects, but some need more. Again, this depends on the complexity of the project.

> **“Starting a machine learning program is a long-term investment. Getting a great return requires a long-term strategy.”**

Depending on what kind of solution you're building, your model may need to be continuously retrained or refreshed. Models which change infrequently, such as voice recognition, will need less frequent data refreshes compared to solutions like fraud detection models, where the inputs are constantly changing. Then, to keep your solution up-to-date, plan to train it regularly on a smaller subset of data with a similar distribution to the initial training data. Your solution might require quarterly, monthly, or even weekly updates. As your business changes, so should your data.

Starting a machine learning program is a long-term investment. Getting a great return requires a long-term strategy.

Sourcing the Right Data

For Your Use Case

The type of data you'll need depends on the kind of solution you're building. For example, a speech recognition solution that understands spoken human commands must be trained on high-quality speech data that has been translated to text. A search solution needs text data annotated by human judges to tell it which results are the most relevant.

Data types

The most common types of data used in machine learning are image, video, speech, audio, and text. This data can be structured to develop or enhance many different kinds of solutions, including:

- **Automatic speech recognition**
- **Autonomous vehicles**
- **Customer experience management (CEM) / customer relationship management (CRM)**
- **Computer vision**
- **Data analytics**
- **eCommerce**
- **Fraud detection**
- **In-car infotainment**
- **In-car navigation**
- **Machine translation**
- **Medical imaging**
- **Risk management models**
- **Proofing tools**
- **Search relevance**
- **Semantic search**
- **Social media**
- **Social media analytics**
- **Text-to-speech**
- **Virtual assistants and chatbots**

Before they're used for machine learning, training data items must be annotated, or labeled, to identify what they are. Annotation tells a model what to do with each piece of data. For example, if a data item for a virtual home assistant is the recording of someone saying "Order more double-A batteries," the annotation might tell the system to start an order with a particular online retailer when it hears "order," and to search for "AA batteries" when it hears "double-A batteries."

Image and Video Data

Computer vision solutions, such as those used in autonomous vehicles and on social media sites, can recognize images and video as well as humans do. Developing an effective one requires large volumes of high-quality, annotated image and video data specific to your product. Self-driving car systems need to be trained on images and video from the areas where they'll be on the road, annotated with information about what each object is (stop signs, traffic lights, pedestrians, emergency vehicles) and how to react appropriately (stop, go, pull over, etc.). A social media platform can be trained with annotated data to recognize content in images and videos to improve user experiences by delivering better personalization.

Speech Data

To work well, automatic speech recognition systems, such as those for in-home assistants and in-car infotainment units, should be able to understand and respond to human speech in a variety of environments and contexts. They can learn to recognize language spoken with different accents or speech impediments, at varying volumes, in dialects, and from people of different ages and genders. Machines can also learn how to read emotions through tone—helpful for identifying customer frustration levels.

To do all of this, computers need to train on human-annotated speech data specific to the context where they'll be deployed: in the cabin of a car, on the phone, in a home, or elsewhere. It's also important to collect natural language utterances (things like "uh," "um" and "er" in English), which help train and test applications to recognize the many nuances of human speech.

Audio Data

The world is filled with sounds, so any AI system that will be "listening" in a real-world environment must be able to distinguish between speech and non-speech sounds. In-home assistants should be able to identify sounds to ignore, like appliance fans, vacuum cleaners, and air conditioning units. Autonomous cars need to recognize sirens and car horns, and operate in a variety of traffic and road conditions. Surveillance systems can be trained to discern breaking glass, gunshots, dogs barking, and sounds of distress. On a lighter note, machines are also at work interpreting sounds in nature—like bird songs, crickets, and frogs—to aid in environmental research.

Every one of these solutions must be fed thousands of pieces of annotated audio data points relevant to the environment where they'll operate.

Text Data

Text data provides the foundation for machine learning projects where systems must learn to respond to specific questions. An example is a chatbot on an insurance website that answers customer inquiries about insurance products.

Additional systems that rely on text data include language translation engines, the prompts and grammar specifications for voice-interactive devices and automated phone systems, search engines, proofreading tools, content moderation tools, social media monitoring and analytics, and others.

Data sources

There are several options for sourcing your training data: real-world usage data, survey data, public datasets, engineered data, "dog food" data, and synthetic data.

> **"Because you can't control what users say and do, you may have to collect and process a lot of extra data to extract what you need."**

Real-World Data

Real-world usage data is a great option if you already have a product in market that is generating the kind of data you'll need. The benefits are that you know it's an accurate reflection of how people already use your system, and you don't have to spend the time or money to create it.

The challenges include legal and privacy concerns, requiring that you give users the choice to share their data or not. Also, because you can't control what users say and do, you may have to collect and process a lot of extra data to extract what you need. Another concern is scale. A company might not have a large enough customer base or enough data to cover all the real-world variation the system will encounter, either at launch, when accommodating new customers, or expanding to new markets with their own cultural nuances.

Additionally, with real data, you'll always be making an educated guess about what people intend. For this data to be really useful, we recommend that you work with a vendor like Appen with a skilled crowd to annotate and/or transcribe it. It's an extra step before you begin, but it will ensure that you train your solution on high-quality data.

Survey Data

Another source for machine learning data is surveys. Ask your users what you want to know: what they like or don't, what you can improve. This approach gives you data from actual users without the legal and privacy concerns, because people must opt-in to provide feedback. It also allows you to direct participants to specific topics, giving you greater control over what kinds of data they produce.

The drawback here is that surveys are somewhat unreliable, because what people say they do and what they actually do can be different things. Additionally, survey data is often skewed toward dissatisfied users, as they have a greater motivation to provide feedback.

Public Datasets

While public datasets are available for all the data types listed above, affordability is the only real advantage to them. Often, they are not specific enough to meet a particular project's unique requirements, and aren't large enough to effectively train an algorithm.

They can be useful, however, for creating commodity technologies, like basic language recognition and machine translation. Appen has an extensive catalog of off-the-shelf, licensable linguistic resources for these kinds of systems that rely on speech.

Engineered or Collected Data

By "engineered" data, we mean making or collecting it yourself. This often the only way to train a new solution that doesn't have real-world users yet. You can simulate the user experience by hiring native speakers, human judges, and other professionals who will create, gather, and annotate the specific training data you need. This type of data collection will require a larger investment, however.

With this approach, you can get exactly what you require without the legal or privacy concerns of real-world data. You can also set the context, and follow up with the team if there's a question. To do it well, we recommend working with an experienced vendor, as there is a lot of management involved to ensure a high-quality dataset.

> **“Appen has an extensive catalog of off-the-shelf, licensable linguistic resources for these kinds of systems that rely on speech.”**

Engineered data is also a good option if you have a limited amount of real-world data but not enough for training a model. "Data augmentation" is the term for this hybrid approach. Work with a vendor to create the data volume, or to fill in the specific gaps you are missing until you have a complete training dataset. Typical strategies include adding different background noises to speech data, rotating angles or adjusting light conditions for image or video data, adding word variation to text data, etc. This is a quick and affordable way to expand a dataset, but does have the same drawbacks of any synthetic data — namely, that it can introduce bias and errors into your training set. Again, we recommend working with an experienced vendor to control for these possible negative impacts.

“Dog Food” Data

Some teams will collect and label data by opening up a solution for a pre-launch period and invite internal employees to use it to generate “dog food” data. The benefits include helping company employees become familiar with the AI solution and understand a potential end-user’s pain points. However, your internal team’s time is costly compared to what you’ll pay professional data vendors. And since your team will likely have expectations about product functionalities and what the system should conclude, you run the risk of building bias into the data.

Synthetic Data

Synthetic data is similar to engineered and "dog-food" data in that it’s manufactured specifically to train your system, but differs in that computers create it instead of people. You can create synthetic data by feeding a real-world dataset to a generative model. The model will learn from that and create more data like it. Synthetic data is useful to quickly and affordably conduct low-risk experiments or to expand a real-world dataset. Disadvantages include the challenge of creating high-quality synthetic data, the increased possibilities of introducing bias, and the likelihood of not accounting for all the various scenarios of a real-world deployment.

Choosing a Data Labeling Platform

Data labeling (also known as annotation) is critical to any AI initiative — having correctly annotated data is what actually trains your machine learning models to make the right decisions. Different types of labels result in different outcomes for the algorithm. The more intricate or nuanced your training data is, the better the outcome. Most organizations need large volumes of high-quality training data, fast and at scale. To achieve this, they must build a data pipeline that delivers sufficient volume at the speed needed to refresh their models. That's why choosing the right data annotation technology is a key piece of your training data strategy,

When evaluating data labeling platforms, it's important to keep the following considerations in mind:

- The tools must handle the appropriate data types for your initiative (i.e., speech tools that recognize text and characters in multiple languages, image annotation tools that provide semantic segmentation and multiple ways to draw bounding boxes).
- The platform should be designed to allow for flexible labeling workflow designs to accommodate various use cases, as well as experimentation, so your team can create datasets to meet the specific needs of your project.
- Workforce management is a critical component of any data labeling project. The technology should be able to both manage an individual annotator's quality and throughput for data labeling tasks, as well as overall project quality and efficiency metrics.
- User experience for data annotators using the platform is also important; data labeling tools should be free of usability issues that may negatively impact the ability to create high-quality training data.
- The platform should provide machine learning-assisted data labeling, to augment data annotators' overall performance, and speed up time to value.

> **“ Appen, through its recent acquisition of Figure Eight, provides the industry's leading solution for data collection. ”**

Choosing the right technology stack for your machine learning initiative will help you collect and structure the necessary data, at the speed, scale, quality, and security levels required by your project.

Appen, through its recent acquisition of Figure Eight, provides the industry's leading solution for data collection, annotation, and crowd workforce management.

Ensuring Data Quality

Once you know where your data is coming from, the next step is planning for quality control. Using poor-quality data to train your machine learning system is like preparing for a physics test by studying geometry. You'll learn something, but your efforts probably won't help you answer your test questions correctly.

How to Manage Data Annotation Quality

Data annotation is the process of tagging or labeling unstructured data, providing crucial metadata that helps a machine to learn about the contents. Depending on the task, data annotation can be a relatively simple activity — but it's also repetitive, time-consuming, and difficult to do right consistently. It requires a human touch, because computers are not always able to deal with ambiguity, and many systems will need to be adjusted for local markets to handle cultural nuance. The stakes are high, because if you train a model on inaccurate data, the model will do the wrong thing. For example, if you train a computer vision system for autonomous vehicles with images of sidewalks mislabeled as streets, the results could be disastrous.

When we talk about quality here, we're talking about both the accuracy and consistency of those labels. Accuracy is how close a label is to the truth. Consistency is the degree to which multiple annotations on various training items agree with one another.

Standard Quality-Assurance Methods

Most vendors use three standard methods for ensuring accuracy and consistency: gold sets, consensus, and auditing.

- Gold sets, or benchmarks, measure accuracy by comparing annotations (or annotators) to a "gold set" or vetted example. This helps to measure how well a set of annotations from a group or individual matches the benchmark.
- Consensus, or overlap, measures consistency and agreement amongst a group, and does so by dividing the sum of agreeing data annotations by the total number of annotations. This is the most common method of quality control for projects with relatively objective rating scales. The goal is to arrive at a consensus decision for each item. Any disagreement amongst the overlapped judgments is typically arbitrated by an auditor.
- Auditing measures both accuracy and consistency by having an expert review the labels, either by spot-checking or reviewing them all. This method is important for projects where arriving at a consensus judgment may not be feasible—tasks such as transcription, where auditors review and rework the content until it reaches the highest levels or accuracy.

Appen's Additional Quality Controls

Beyond these, Appen has developed further methods to ensure our annotators consistently surpass the bar for quality and productivity.

Custom quality solutions: We develop custom solutions for each project, ensuring we address the data type being collected, and how our clients will use the data. For example, in a subjective task, we look at consistency and logic instead of only using a consensus measurement.

Quality measurement metrics: We use multiple methods of quality measurement—even if our clients only require a single metric. To illustrate, we may use consensus, accuracy, rating distribution, and regular auditing as quality measurements in an image annotation project.

Weekly data deep-dives: Our project management team investigates data weekly and sets stretch productivity and quality goals to exceed client expectations. If a client requests quality of 92%, we'll set a stretch goal of 95% and will not accept work under 92%.

> “We develop custom solutions for each project, ensuring we address the data type being collected.”

Real-time learning methods: We use real-time learning methods like rapid evaluation feedback (REF) sets and spot checks so annotators can receive feedback as they work and apply what they learn to their next task. This approach facilitates rapid improvement of both performance and the data compared to waiting for test results and seeing improvement a week later. In practice, if an Appen annotator makes an incorrect evaluation and receives feedback, they must acknowledge the feedback before starting on the next task.

Management testing and auditing: Appen's project managers, supervisors, and leads all must complete the project exam, annotation work, and audits as well. This gives the management team a 360-degree view of the project and a full understanding of the annotators' experience. When project managers know the task and its challenges, they can create better resources and design better solutions.

How to Avoid Bias

Bias is often invisible until it hits the market in your AI product—at which point it can be difficult to fix. One reason we put so much emphasis on data quality is to help our clients avoid bias, which generally comes from blind spots or unconscious preferences in the project team or training data, from the outset of a project. Bias in AI can manifest as uneven voice or facial recognition performance for different genders, accents, or ethnicities. As AI becomes more prevalent in our culture, now is the time to address built-in bias.

> **“One reason we put so much emphasis on data quality is to help our clients avoid bias.”**

Mitigating Project-Level Bias

To avoid bias at the project level, actively build diversity into the teams defining your goals, roadmaps, metrics, and algorithms. Hiring a diverse team of data talent is easier said than done, but the stakes are high. If the internal “us” of your team doesn’t represent the external “us” of your potential customers, then the end product risks only working for, or appealing to, a subset of people, and missing a mass-market opportunity. Or worse, bias in products can lead to discrimination in hiring, housing, lending, and other areas of modern life.

Data-Level Bias

Bias can creep into training data in a number of ways. Here are six steps to mitigate bias in your training data:

1. Recruit diverse teams to build your models, create your training data, and evaluate it. Consider gender, race, age, geographic location, language preference, education levels, and whatever other characteristics may apply to the real-world customers using your particular solution.

2. When internal team members are labeling the data, they will always add some bias because they have expectations about what their system should conclude. If you decide to use an internal team, Appen’s data evaluation

services can help by bringing an outside perspective to machine learning training data.

3. Find or create a representative training dataset. Quantity always helps, especially if you're using data from internal systems. Try to find the most comprehensive data, and experiment with different datasets, metrics, and segmentation to ensure you've covered the bases.

4. If you're engineering or annotating data, take care to design the instructions and tasks for your creators and annotators in ways that don't bias them from the outset. It's important that annotators have enough instruction to correctly perform the task, but not know what the data will be used for, which can bias behavior. Appen has a team of dedicated, experienced instructional designers that can help.

5. Check for implicit bias in the data as part of your quality-assurance process. Vendors like Appen can help with data visualization and insight tools.

6. Once your product is live, monitor performance using the data it generates to determine whether it's delivering equitable opportunities and outcomes for all users.

Ensuring Data Security

Data security is more important than ever, whether you're working with the personally identifiable information (PII) of your customers, financial or government records, or user-generated content. Increasingly, government regulations are dictating how companies must handle customer information.

Securing this confidential data protects your—and your customers'—information. Being transparent and ethical about your practices and sticking to your terms of service gives you a competitive advantage. Not doing so puts you at risk of scandal and negative impacts to your brand.

How Appen Manages Data Security

Especially with outsourced data projects, you need confidence that your data is secure. As a vendor with over 20 years of experience collecting, transcribing, and translating sensitive data—and over a decade of experience annotating it for machine learning—Appen offers flexible options to meet our clients' varied security and budgetary needs.

Data security issues can come from several areas. We think of data security in three main terms:

1. **Platform security**
2. **Data security in the annotation process**
3. **Securing the annotator workforce**

Platform Security

Platform security is essentially the practice of securing the tools you use to collect and manage your data. For Appen, platform security includes:

- Access management to control proper access to your computer platform services and data with a complete audit trail.
- Change management to identify and track change with process, tools, and techniques that enable auditing and manage change to achieve a business outcome.
- Business continuity, which gives organizations the ability to respond to severe interruptions and continue operating effectively and securely for employees and customers.
- Compliance/certification to ensure companies can meet their legal obligations on data security and privacy to protect the health, safety, and welfare of their customers.

Data Security in the Annotation Process

The previous sections of this paper have made the case that data quantity and quality will make or break your AI projects. A company's data is, increasingly, one of its most valuable assets. So, putting the right security in place is critical. Here's how Appen secures our clients' data during the annotation process.

- Data classification is the first step, where we categorize data based on the level of sensitivity: restricted, private, or public.
- Data protection is next, ensuring that private and restricted data is encrypted end-to-end, both in transit and at rest using the latest security protocols.
- Enterprise deployment options give customers with additional security, control, compliance, and throughput needs, the ability to govern their own data via public cloud, private cloud, and air-gapped, on-premises deployments.

> **“A company's data is, increasingly, one of its most valuable assets. So, putting the right security in place is critical.”**

Securing the Workforce

Global businesses need secure locations for workers in different locations. Appen offers three secure service options to meet clients' needs: secure facilities, secure remote workers, and onsite services.

We maintain our own secure facilities across five continents with a range of security levels from commercial to government-level certifications, including ISO 27001, ISO 9001 and the Cyber Essentials certification. We have sites in multiple locations, including the UK, Philippines, US, and Australia.

We can also secure remote workers globally while maintaining data security with the latest virtual private networking (VPN) Office and Virtual Application technologies.

Additionally, we can meet you where you are with secure onsite services, bringing outside consultants to work onsite at your firm. We'll manage staff onboarding, including background checks, while ensuring compliance with your requirements for data access.

Why Outsource
Data Collection and Annotation?

The arguments for outsourcing data collection and annotation are similar to those for outsourcing any aspect of your business. By hiring a qualified vendor, you can scale up faster, test and tune with expert guidance, and keep your own teams focused on building and improving your core business and intellectual property (IP).

Outsourcing with Appen	In-house
✓ Established quality standards	X Variable quality
✓ Dedicated, managed resources remove burden from your team	X Data labeling tasks detract team focus from core IP
✓ Expertise in multiple languages	X Limited language experience
✓ Purpose-built solutions for data security	X Must develop annotation processes to maintain security
✓ Speed and scale help launch solutions faster	X Time-consuming and resource-intensive
✓ Removes internal bias	X High rate of evaluation bias
✓ Customizable billing model	X Full financial burden of project planning & headcount
✓ Low attrition on projects	X Risk for employee attrition
✓ No additional office space	X Must provide additional office space
✓ Fully managed services, from task design & guidelines, to curating a skilled crowd of annotators, to QA, to course correction, to final outcomes	X Full burden of task design, crowd recruitment, QA, and achieving final outcomes

How Appen Delivers Value

Appen can quickly deliver large volumes of high-quality image, video, speech, audio, and text data, by collecting real-world data, annotating your existing data, or creating the data you'll need from scratch. Whether you're entering a new market and need to localize your solution, or building a solution from the ground up, we can quickly recruit large numbers of participants while meeting your diversity, budget, and security requirements.

Appen's Managed Services

Not all vendors deliver the same level of quality at every stage. Here's an overview of how we maintain quality and output:

Recruiting

Attracting, onboarding, and providing opportunities for qualified annotators who meet business requirements

Onboarding

Educating and evaluating annotator competence—before they're deployed on a project

Quality Assurance

Managing the productivity and accuracy of our global crowd to meet and exceed quality targets

Flexible Scaling

Ramping annotator resources—remote, on-site, or in a secured facility—in different geographies, as project requirements change

Program Management

Facilitating end-to-end project design, regular communication, and reporting with client

Service Excellence

Delivering cost-effective, on-time datasets, and the business benefits outlined in the project design

To help create your AI solutions, and keep them up-to-date, we'll work with you to develop an ongoing maintenance plan to test and tune your algorithm and your data so you get the return you want from your investment.

Your data scientists are busy—and their expertise in your business, your IP, is valuable. Make sure they spend their time maximizing your competitive advantage. Leave the sorting, rating and annotating to us and our million-strong, diverse, global crowd ready to create the high-quality training data your machine learning project needs.

No matter what type of project, we'll work with you to develop a customized, end-to-end program. Then we'll enlist our skilled project managers to ensure high-quality results from the outset.

Conclusion

A recent study by IHS Markit[4] revealed that 87% of organizations are adopting at least one form of transformative technology like AI, but only 26% believe that appropriate business models are in place to capture full value from these technologies. A partner like Appen can help plan, develop, launch, and maintain your machine learning project to support your success.

> **“A partner like Appen can help plan, develop, launch, and maintain your machine learning project to support your success.”**

Creating a solid strategy is the first step. That includes setting your budget, identifying your data sources, ensuring quality, and building in security. Developing a clear data strategy can also help provide the steady pipeline of data that most machine learning models need to be regularly refreshed. Appen is trusted by leading companies and governments worldwide as a provider of high-quality training data for machine learning.

Contact us to see how we can put AI to work for your business at **hello@appen.com**, or visit our website at **appen.com**.

[4] HIS Markit, Digital Orbit: Tracking the development, impact, and disruption caused by transformative technologies across key industries, 2019

About Us

Appen collects and labels images, text, speech, audio, and video used to build and continuously improve the world's most innovative artificial intelligence systems.

With expertise in more than 180 languages, a global crowd of over 1 million skilled contractors, and the industry's most advanced AI-assisted data annotation platform, Appen solutions provide the quality, security, and speed required by leaders in technology, automotive, financial services, retail, manufacturing, and governments worldwide.

Founded in 1996, Appen has customers and offices around the world.

Experience working in **130+ countries**

Expertise in **180+ languages**

20+ years working with leading global technology companies

Access to a curated crowd of over **1,000,000** flexible workers worldwide

More than **13 billion** judgments made and **500,000** hours of audio processed

Industry-leading **data annotation platform**

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