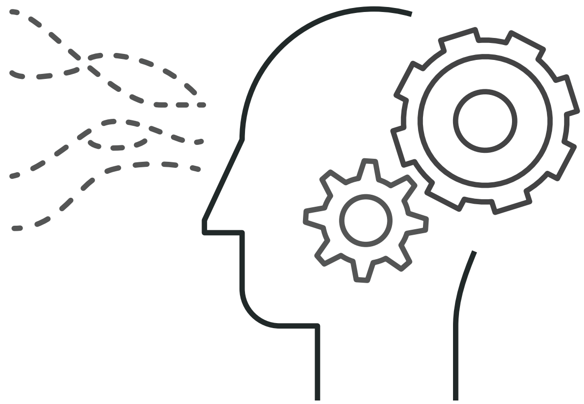
How to put active learning to work for your enterprise

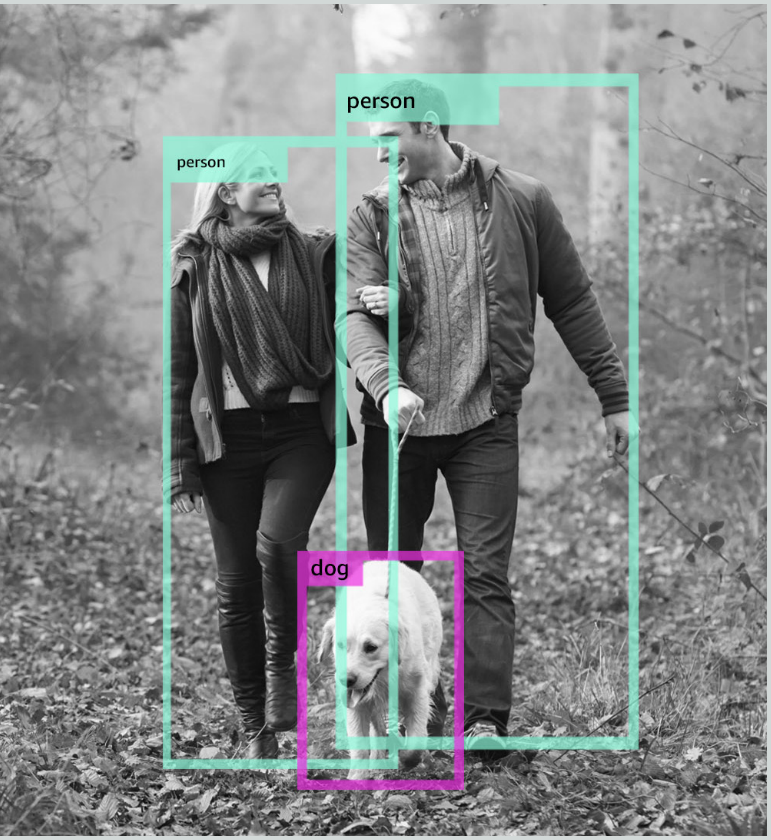
Choosing the right machine learning strategy

By now, most people are familiar with the promise of machine learning across nearly every business.



It can increase efficiency, lead to smarter decisions, and even alter entire industries.

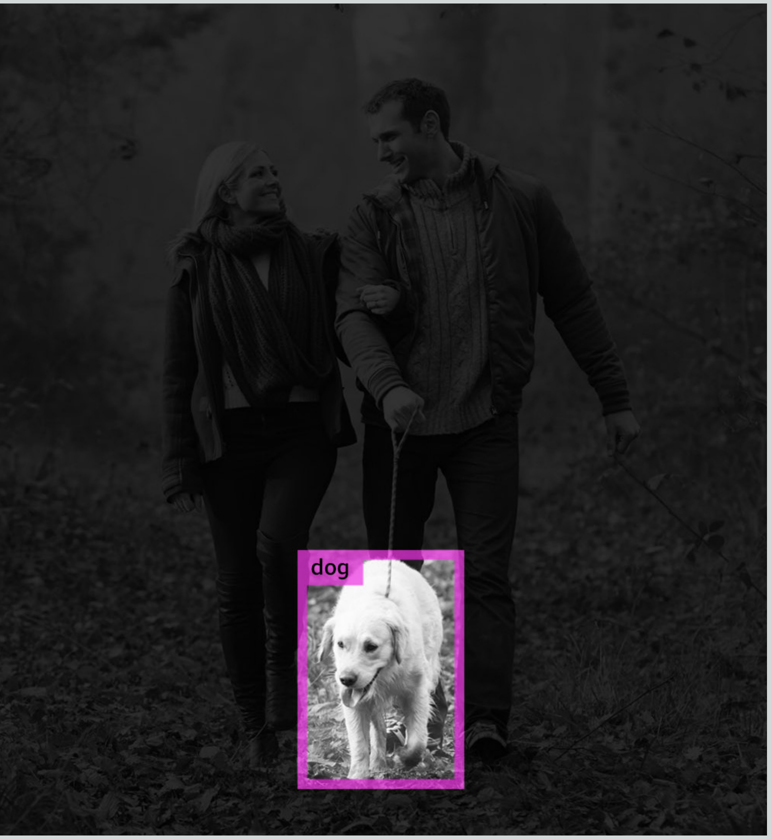
Yet, while many people understand the benefits of machine learning, far fewer understand which approach works best for their specific needs. After all, machine learning isn't a one-size-fits-all arrangement. What might work for a hedge fund might not work for a social media startup or a company training autonomous vehicles. In this eBook, we'll start by defining a few high-level machine learning categories, and continue by explaining **active learning**, a powerful solution for myriad real-world business problems.



Supervised vs. unsupervised machine learning

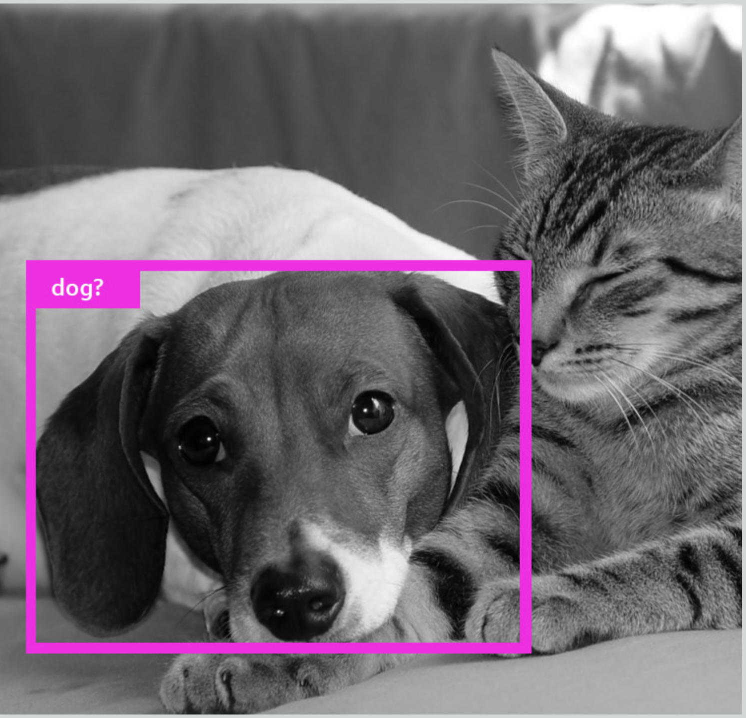
GENERALLY SPEAKING, YOU CAN SPLIT MACHINE LEARNING INTO TWO MAIN CATEGORIES: SUPERVISED AND UNSUPERVISED MACHINE LEARNING.

Supervised learning refers to scenarios where you have correct input-output pairs as training data from which a machine learns. Training a machine learning computer vision model with a set of labeled images is an example of supervised learning. You might show a machine a picture of a dog with a label that says "dog," or a piece of social media labeled with a positive sentiment. By showing the machine enough of these labeled examples, it can take the "input" of an image of a dog and provide you the "output" of the label "dog."



Unsupervised learning is much different. Here, there are no correct input-output pairs. There's really no "right" answer. Instead, unsupervised learning refers to scenarios in which a machine finds patterns and relationships from analyzing a large amount of data. One example of this method would be segmenting your entire database of customers into subgroups based on purchasing patterns, geography, demographics, and related attributes.

While unsupervised learning can be a powerful method for solving certain problems, in this eBook, we'll be covering supervised learning, specifically a subset of supervised learning called "active learning." We will explain what it is, how it works, and areas in which it can be particularly effective.



What is active learning and how does it work?

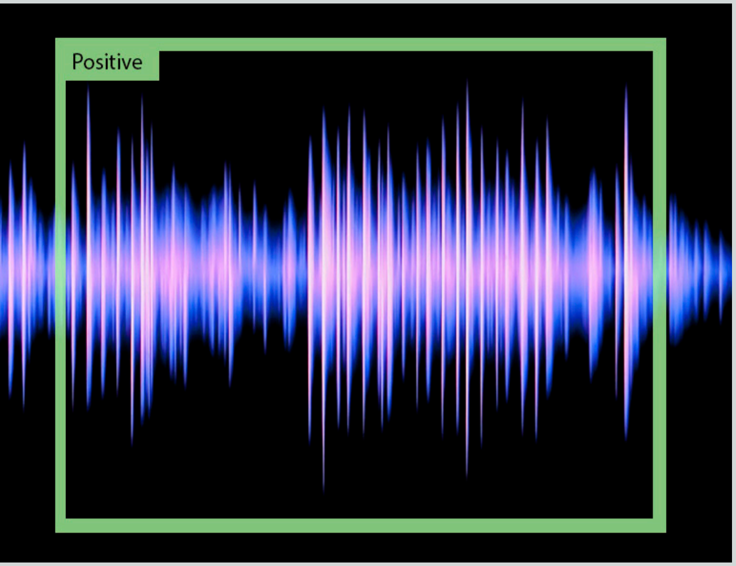
It's important to remember that active learning is a branch of supervised learning. In other words, with active learning, there are correct, labeled input-output pairs in the training data. The goal of your model should be providing correct outputs from unlabeled inputs. For example, if a model has been trained with labeled images, the goal of a model might be to output the correct labels from unlabeled images (e.g. understanding that an unlabeled picture of a dog is a dog).

Active learning, sometimes called **human-in-the-loop**, is different than some kinds of supervised learning in that it leverages human intelligence as part of an iterative training process. In active learning, humans validate, correct, or provide labels made by a model. From there, those human judgments are fed back into the model to improve it.

Using our simple vision model as a guide, here are two examples of how active learning works in the real world:

- A model doesn't have sufficient confidence to tell if an unlabeled image is a dog. A human makes the determination, and this data point is fed back into the model so it has additional training data to learn from.
- A model believes an image is a dog. A human either validates that output or corrects it, and this data point is fed back into the model so it has additional training data to learn from.

Done properly, active learning continually improves a machine learning model's accuracy over time. It tells a model when it's right, when it's wrong, and provides answers when it's unsure.



How do supervised machine learning models actually learn?

Supervised machine learning models learn from labeled datasets. These datasets are populated with correct input-output pairs. Such a dataset is generally referred to as **training data**.

In supervised learning, it's important that your training data is labeled. While we've used the image of a dog with the label of "dog" here, training data could consist of other information, such as a short spoken sentence with a transcription. In this case, the input is the sound file, and the output is the text of what was spoken. Another example would be a short piece of text labeled with a sentiment (the input could be a tweet, and the output in that case would be "positive" or "negative").

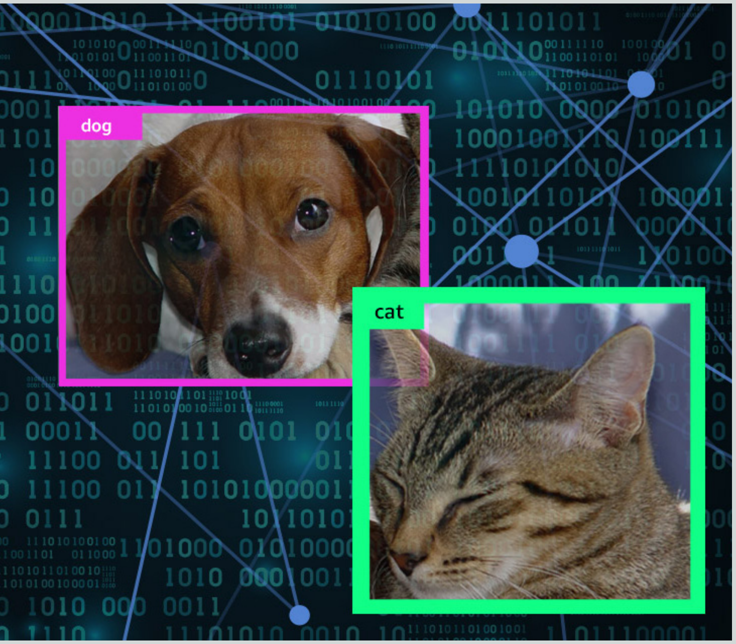
Labeled examples teach supervised machine learning models about the relationships between inputs and outputs. In essence, they teach them what it is that you are actually looking to automate.

Creating a robust training data set is a key part of any supervised learning project. The same, of course, is true for active learning. Oftentimes, training data will need to be labeled by human annotators. Think of this as providing the correct outputs for unlabeled inputs. For example, you might show a human annotator an image and ask that annotator what it is. That labeled input-output pair is a “correct” answer that a supervised model can use to learn what that object actually looks like.

How to create training data for your specific needs

A key step in creating the right training data for your active learning model is understanding your goals: that is, what it is that you actually want your model to accomplish. To do so, it's crucial that you understand both the inputs being provided to your model, and your desired outputs from that model.

This may seem rudimentary, but it's incredibly important. Let's go back to our image classifier example. Here, you largely know your inputs: those are simply pictures. However, your outputs are up to you. If it's important that your model identifies pedestrians versus dogs, that input-output pair would need to be in your training data. In other words, you would need to provide your model with both labeled images of pedestrians and dogs. However, if your business goals do not involve understanding different objects–say cars and cats–you likely do not need those labels in your training data.



This holds true no matter the business case. A model that identifies customer opinion for a restaurant can be trained to classify simple “positive” or “negative” emotions or to predict the reasons behind those emotional responses. You simply need to provide the schema for those reasons (e.g. “the food was cold”).

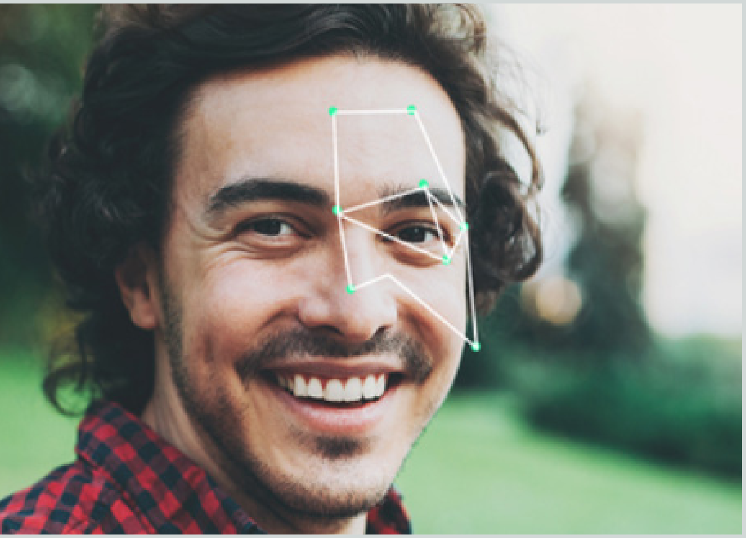
Remember, your model will eventually provide outputs from unlabeled data; it is not able to create new labels. An active learning model that is not shown what a cat looks like will never really know what a cat looks like. You need to provide the input-output pairs–or the training data–to actually teach it to recognize a cat image.

How active learning iteratively improves your model

One of the most attractive aspects of active learning is the continuous, iterative process of model improvement.

Active learning leverages human intelligence to train your model after it has learned from your original training dataset (i.e. your input-output pairs). Essentially, once a model is trained and given new unlabeled data, the model should be able to produce an output prediction (for example, “this image is a car”). However, the model may not always label data correctly. It could be wrong, it could be correct but with low confidence, or it could be unable to provide a prediction. This is where active learning really shines.

This is why active learning is also called human-in-the-loop machine learning. Here, the loop is the iterative process of improving models with clarifications and better data. The human is in the loop, of course, because the human is the one providing those validations or corrections.



Use Figure Eight, AWS, and active learning to meet your business goals

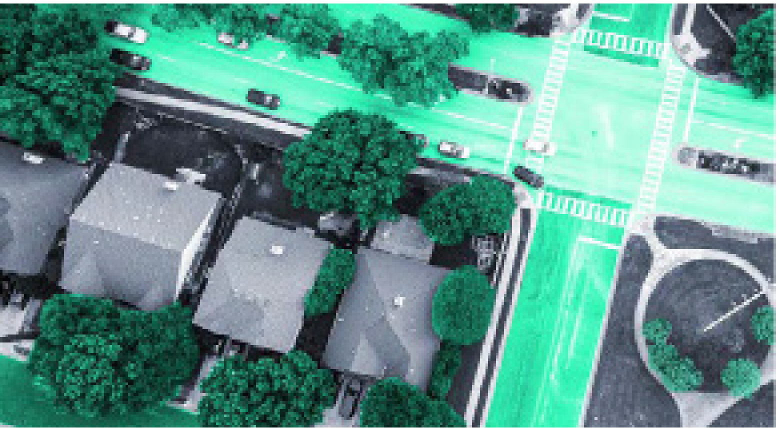
Effectively, what you're doing is finding areas where your model errs and teaching it exactly what it needs to know. If your model makes poor predictions about a certain output, you show it more input-output pairs, giving it more examples so it can better understand the relationships it needs to in order to make confident predictions.

Figure Eight can facilitate both the building of training datasets and the human-in-the-loop process of active learning. Figure Eight's platform is flexible, allowing you to annotate, score, or judge any kind of data–text, images, audio, or video. Built-in quality controls keep data accuracy high and their platform reliably generates massive amounts of human-labeled training data. You can also feed your algorithms predictions into Figure Eight to iteratively improve your model with active learning.

Figure Eight runs on Amazon Web Services (AWS) and integrates with AWS services including Amazon S3 and Amazon SageMaker. The solution takes advantage of the interoperability, scalability, and flexibility of AWS so that your models perform well, regardless of the size of the workload.

Figure Eight features:

- A platform that allows you to clean, collect, and label large training datasets at enterprise-scale.
- Access to on-demand human-in-the-loop annotators.
- Robust quality controls that ensures your data is of the highest quality.
- Consulting and technology services to help your machine learning project achieve your desired results.



How to get started with Figure Eight

Learn more about how Figure Eight and AWS can help you get started with human-in-the-loop machine learning. Contact us to set up a free 30-minute consultation with our experts.

LEARN MORE