Google News Initiative

7 LESSONS - 42 MIN

Hands-on Machine Learning

This course teaches journalists how to use Machine Learning for their reporting. You will learn how to train a Machine Learning model to identify and classify images in vast datasets.

Tools Used:
Google Cloud AutoML Vision

Created by:

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

JournalismAi

TEXTY.ORG.UA

Google News Initiative

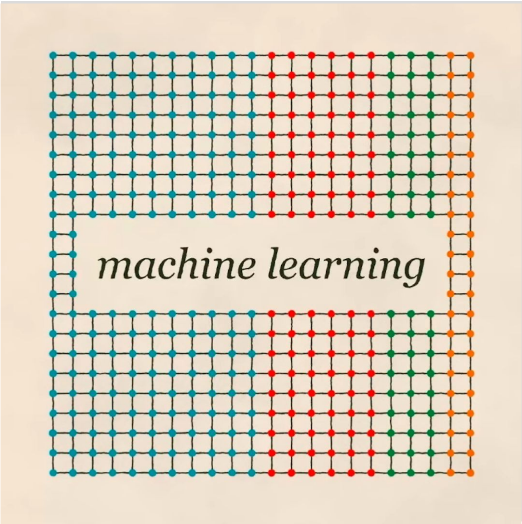
LESSON 1

What is Machine Learning?

Machine Learning for journalists. What you will learn in this course.

Lesson Overview

Journalism and Machine Learning



How can journalists use machine learning (ML) to enhance their journalistic work? This is the question that we will explore in this course.

The course will help you understand what are the situations when machine learning is the right tool to support your reporting and will teach you how to train a machine learning model.

This is a follow-up to the course Introduction to Machine Learning. If you haven't taken it yet, we encourage you to do so before proceeding with this course.

For more lessons, visit:

newsinitiative.withgoogle.com/training

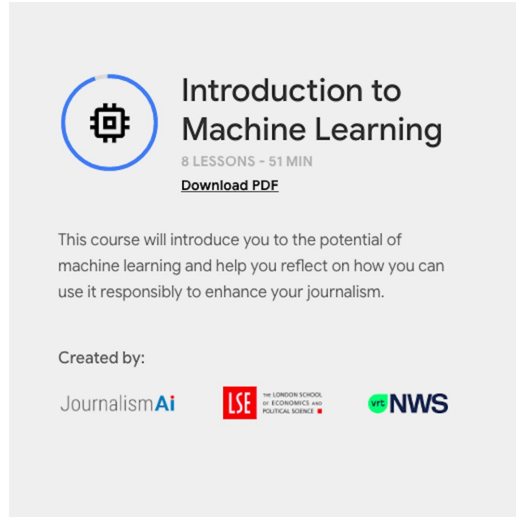
What you can expect from the course

SINGLE STEP

Introduction to Machine Learning explores the potential offered by machine learning to news organisations and explains how journalists can use it responsibly to enhance their reporting.

This course wants to go one step further and show, via a real-life example that we will introduce in the next lesson, what results journalists can achieve by using machine learning. If you want to learn how machine learning works in practice and how you can use it to report your stories, this course is for you.

Will you be an expert machine learning designer and data scientist at the end of this course? No, sorry. But you will learn the steps that underline how most machine learning processes work and you will be able to run experiments on your own.



Defining Machine Learning

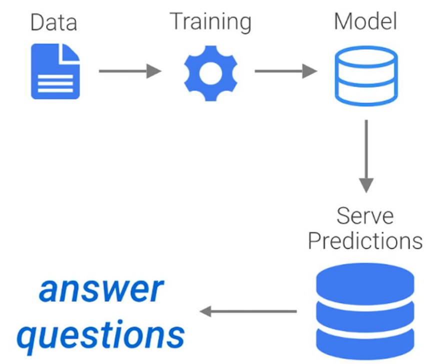
SINGLE STEP

Before we continue, let's make sure we know what we are talking about. What is machine learning?

Machine learning is part of a collection of technologies that are grouped under the umbrella term "artificial intelligence" (AI). As with most of the terminology in the field of artificial intelligence, there is no unique definition of machine learning.

In simple terms, machine learning is a technology that uses data to answer questions. More formally, it refers to the use of algorithms that learn patterns from data and are able to perform tasks without being explicitly programmed to do so.

Moreover, a defining feature of machine learning systems is that they improve their performance with experience and data. In other words: they learn.



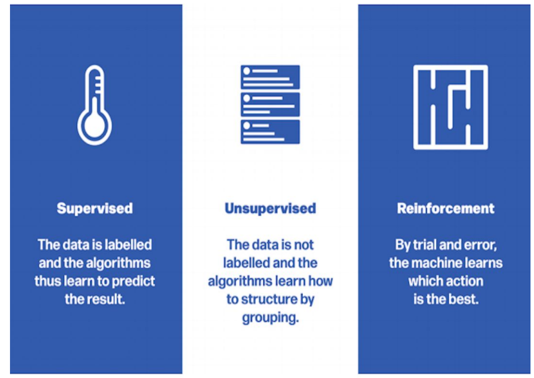
There are various ways to learn

SINGLE STEP

You should keep in mind that there is not one only way for a machine to learn. Different approaches to machine learning are commonly distinguished by the kinds of problems they try to solve, as well as the type and amount of feedback provided.

Broadly, we can divide machine learning into three subareas: 1) Supervised learning; 2) Unsupervised learning; 3) Reinforcement learning. Review the [Introduction to Machine Learning] to find out more about what differentiates these three categories.

For the purpose of this course, we will focus on supervised learning. This means that we will use labelled examples to train an algorithm to automatically assign the correct label to each new example we will ask it to analyse.



Exploring the potential of Machine Learning

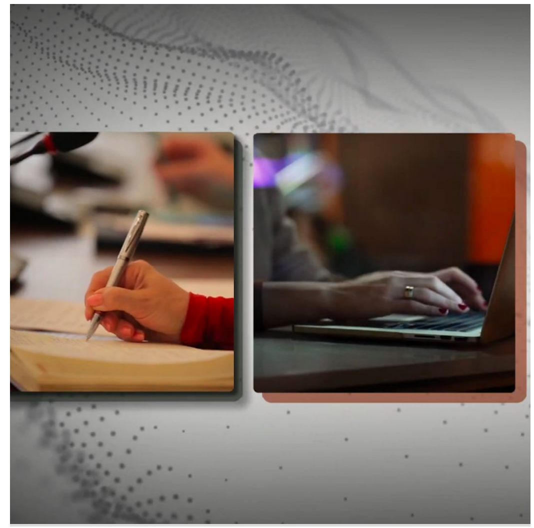
SINGLE STEP

Now that we have reviewed the basics, we are ready to wrap up this introduction and move on.

In the next two lessons, we will introduce a case study that will ground our exercise in a concrete journalistic example and the algorithm that we will use to understand the dynamics behind most machine learning processes.

The following lessons will then focus on the practical step-by-step guide: How to source and prepare the data, how to train your machine learning model and how to test and evaluate its performance.

The last lesson will summarise the key learnings, help you understand how to apply them in your day-to-day reporting and recommend other resources you can use to dive even deeper into the world of machine learning.



LESSON 2

Investigating stories with Machine Learning

How you can use Machine Learning in your reporting

Lesson Overview

Machine Learning for investigations: a case study

In 2010, the price of amber on the global market started to surge. Due to the high demand, in the following years parts of north-western Ukraine, rich in amber, attracted foreign and local interest and became the scene of an illegal "amber rush", a new "Wild West".

Hundreds of hectares of forests and agricultural land were turned into a lifeless moon landscape, with the most intense mining activity taking place between 2014 and 2016 but continuing over the following years.

1 Leprosy of the Land, an investigation by Texty

2 Finding examples of illegal amber mining

3 Is ML the right tool for this problem?

4 Focus on the process

For more lessons, visit:

newsinitiative.withgoogle.com/training

Leprosy of the Land, an investigation by Texty

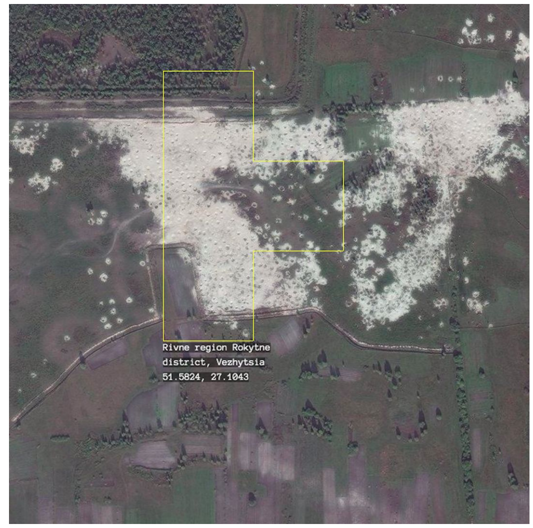
SINGLE STEP

In 2018, Ukrainian data journalism agency Texty published Leprosy of the Land, an investigation in which they used machine learning techniques to detect cases of illegal amber mining across Ukraine.

First, an algorithm divided sections of satellite images into visually uniform subsections. So if an image was half green forest and half dirt field, it would split the image into those two subsections.

Another algorithm found which subsections most resembled the existing examples of amber mining, which have a distinctive pockmark-like pattern of holes in the ground.

Finally, the journalists examined the examples the algorithm found, to make sure that what it thought looked like amber mining wasn't actually something else, like deforestation.



Finding examples of illegal amber mining

SINGLE STEP

In this course, we will focus on the methods used by Texty to train an algorithm to recognise visual examples of illegal amber mining in a huge amount of satellite images, previously divided in subsections by another algorithm.

As mentioned in the first lesson, this means we will experiment with supervised learning. You will learn how the algorithm can learn from labelled examples to recognise the same pattern in images it has never seen before.

You will also learn how you can replicate the process for your own stories: from finding the examples you need, to training a machine learning model to recognise what you are looking for, and then to testing and evaluating the model to make sure it provides reliable results.

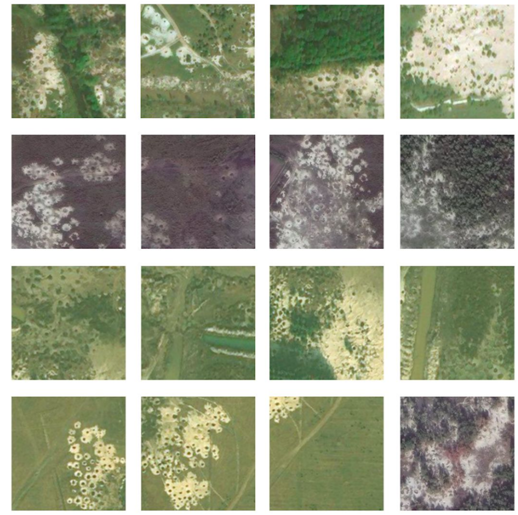
Is ML the right tool for this problem?

SINGLE STEP

But why was machine learning the right tool to find the information that Texty was looking for?

Classical programming requires you to specify step-by-step instructions for the computer to follow. While this approach works for solving a wide variety of problems, it isn't up to the task of recognising examples of illegal amber mining in a huge amount of satellite images. There are just so many visual elements that the computer would need to consider that it's impossible to come up with a step-by-step set of rules that could teach the software to distinguish between real examples of illegal amber mining and things that might just look similar to it.

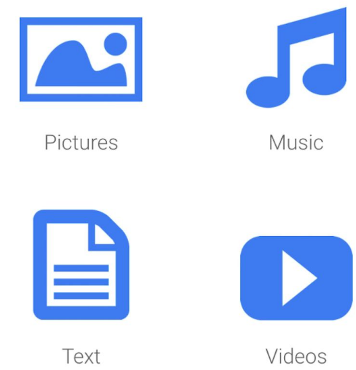
Fortunately, machine learning systems are well-positioned to solve this problem.



Focus on the process

SINGLE STEP

Keep in mind that what you will learn in this course – how to spot illegal amber mining – is only one example. Following the same process, machine learning can be used to perform a number of different journalistic tasks and can even be applied to analyse different types of content, not only images. We will review some other use cases at the end of the course. As we go through the exercise, remember to focus on the process rather than on the specific case study.



Now, before we start the actual exercise, we need to dedicate a few minutes to meeting and setting up the tool we will learn to use in the next lessons: Google Cloud AutoML Vision.

LESSON ##

Google Cloud AutoML Vision

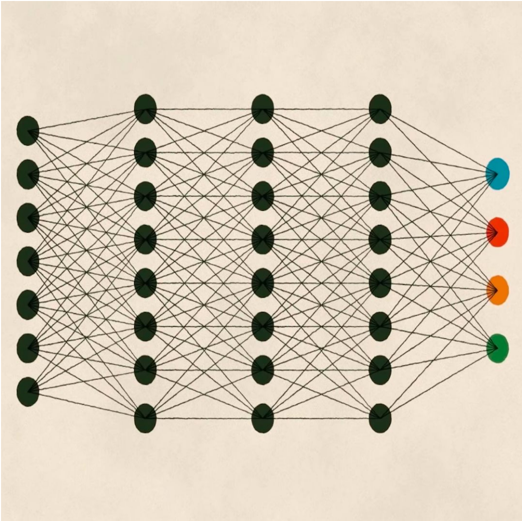
Learn how to set up AutoML Vision to prepare for the course exercise

Lesson Overview

Classifying images with Machine Learning

As mentioned in the previous lesson, Texty used two different algorithms in the production of Leprosy of the Land.

After the first algorithm allowed them to divide sections of satellite images of Ukrainian forests into visually uniform subsections, they needed a second algorithm that could identify which sections of satellite images most resembled the existing image examples of amber mining. What they needed was a so-called "custom classifier".



1	Using labelled examples to learn
2	Choosing an algorithm
3	Setting up your Google Cloud account
4	Moving forward

For more lessons, visit:

newsinitiative.withgoogle.com/training

Using labelled examples to learn

SINGLE STEP

A custom classifier is a type of machine learning model that you can deploy when your use case requires you to apply pre-defined labels to classify a dataset of images you want to investigate.

In our case, those pre-defined labels are simple: "YES: this image includes visual elements consistent with patterns that usually show amber mining activity" and "NO: this image doesn't include visual elements that suggest amber mining activity".

Google Cloud AutoML Vision enables us to do just that. We will learn how to use it to perform supervised learning, that is to say that we will train a machine learning model to apply the appropriate YES and NO labels to a dataset of images we will feed it.

Google Cloud
AutoML Vision

Choosing an algorithm

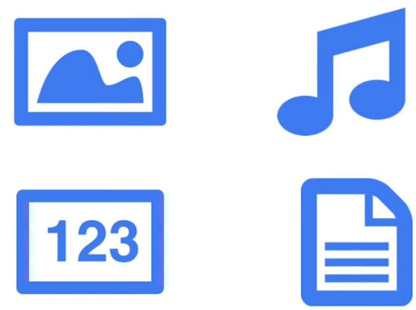
SINGLE STEP

As Jeremy Merrill of the Quartz AI Studio said in his Crash Course in Classifying Text with Machine Learning, "for your purposes as a journalist – it doesn't matter much which algorithm you pick, as long as you pick an algorithm that does the right kind of thing."

AutoML Vision is not the only tool we could use to achieve our desired goal. Actually, it is not the algorithm that Texty used during their investigation. The reason why we are using AutoML Vision in this course is its accessibility: you don't need to have any coding skills in order to learn how it works and to train a high-performing model on your data.

If you do have coding skills already and you want to dig deeper, have a look at fast.ai's Practical Deep Learning for Coders.

Choosing a Model

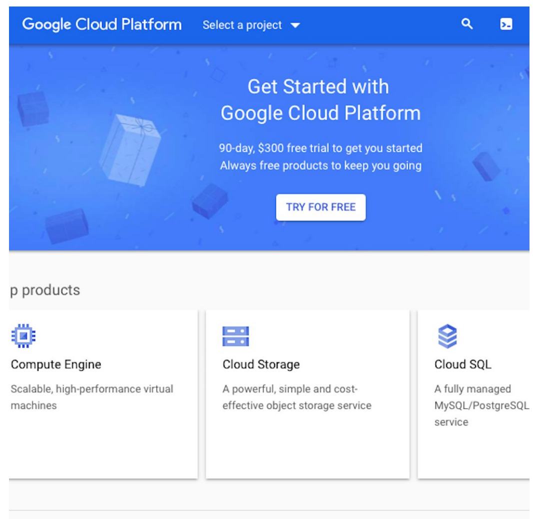


Setting up your Google Cloud account

To use AutoML Vision, you'll have to [sign up for a Google Cloud account](). Upon signing up, you will be granted $300 of credit to start your experiments. Each exercise in training a machine learning model, like the one we'll do in this course, costs about $20. Follow this step-by-step guide:

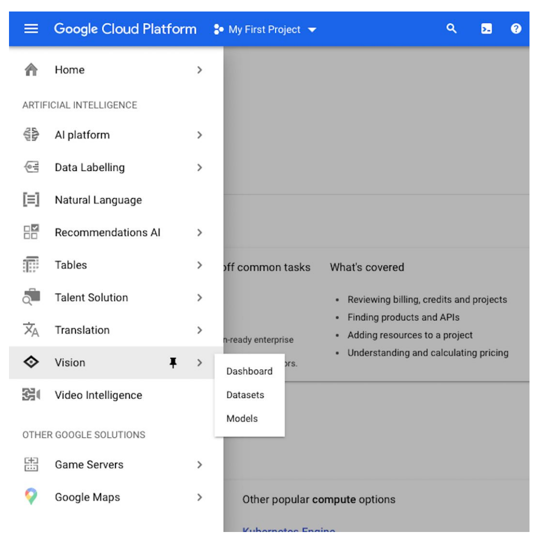
STEP 1 OF 5

Click on “Try for Free” under “Get Started with Google Cloud Platform” and follow the instructions to create your account.



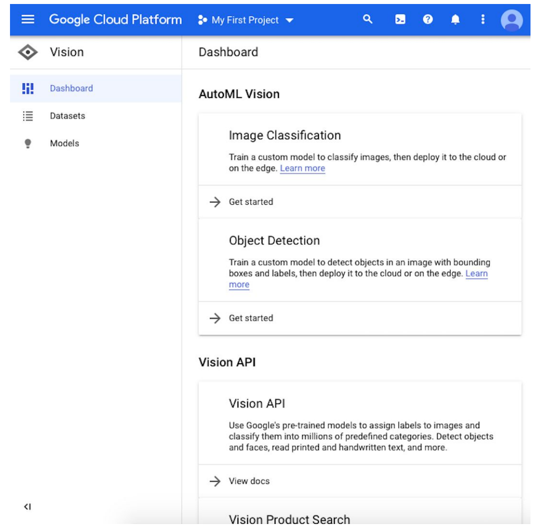
STEP 2 OF 5

When you have created the account, open the navigation menu on the left side of the page and scroll to the very bottom to find “Vision” in the “Artificial Intelligence” section. Click on “Dashboard”.



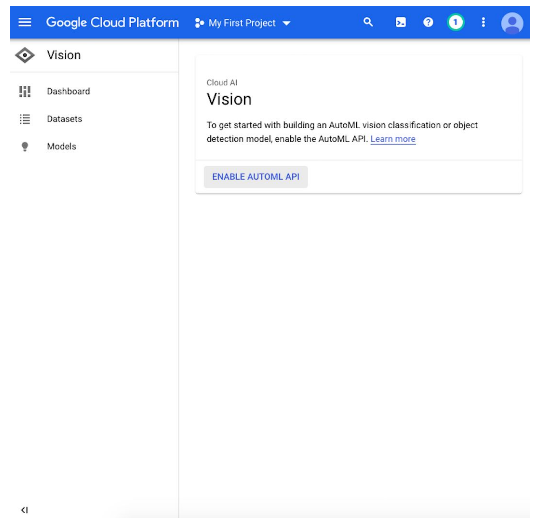
STEP 3 OF 5

You have now accessed your workspace, which shows the Google Cloud “Vision” tools, including the one we will use: “Image Classification”. Click on “Datasets” in the navigation menu on the left side.



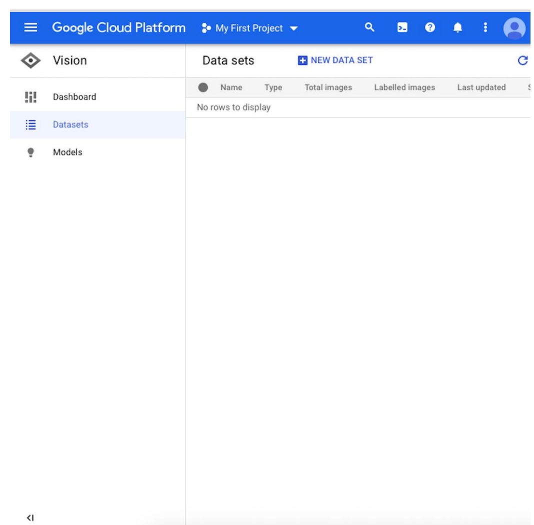
STEP 4 OF 5

Next, click on “Enable AutoML API”. The process might take a few seconds. Then, click on “Get Started”.



STEP 5 OF 5

At this point, you should see a mostly blank screen as you have not yet updated any dataset. This is what we will do in the next lesson.



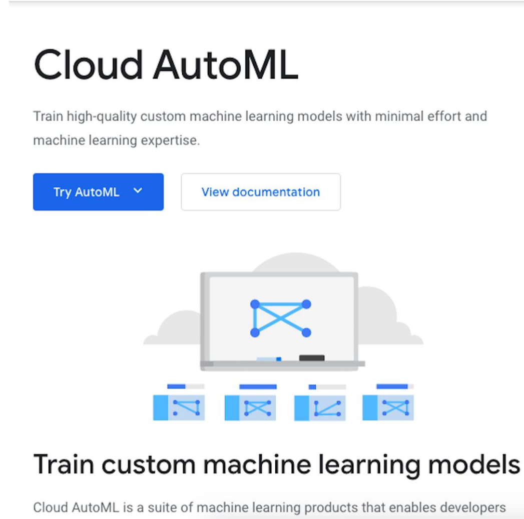
Moving forward

SINGLE STEP

Now you are ready to use AutoML Vision. In the rest of the course, we will learn how to use it to achieve our desired outcome: to train a machine learning model to recognise illegal amber mining.

Texty and JournalismAI are partnering in the production of this course. Thanks to this partnership, we will be able to use a sample of the actual satellite images used by Texty while investigating [Leprosy of the Land](#).

Before we move forward, make sure to check the other [AI and Machine Learning products](#) offered by Google Cloud, including Natural Language, Translation, Speech-to-Text and Text-to-Speech, and much more.



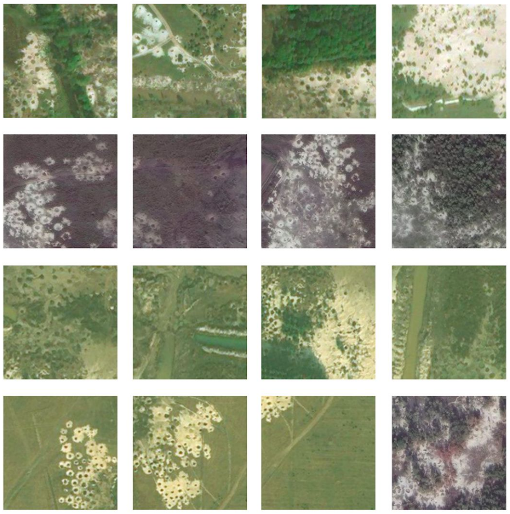
LESSON 4

Data preparation

Assess your use case, source and prepare your data

Lesson Overview

What is training data?



If you have properly set up your Google Cloud account, you are now ready for the exercise. In this lesson, you will learn what questions you should ask while gathering the training data and how to prepare it to be used by AutoML Vision.

With training data, what we mean is examples of what we want our ML model to be able to recognise and categorise. In our case, this means providing a set of satellite images and telling the algorithm which ones are examples of amber mining and which are not.

1 Start with your use case

2 Assess your use case

3 Source your data

4 Prepare your data

For more lessons, visit:

newsinitiative.withgoogle.com/training

Start with your use case

SINGLE STEP

While putting together the dataset, always start from the problem you are asking ML to help you solve. Consider the following questions:

I. What is the outcome you're trying to achieve?
II. What kinds of categories would you need to recognise to achieve this outcome?
III. Is it possible for humans to recognise those categories? Although AutoML Vision can handle many more images and categories than humans can, if a human cannot recognise a specific category, then AutoML Vision will have a hard time as well.
IV. What kinds of examples would best reflect the type and range of data your system will classify?

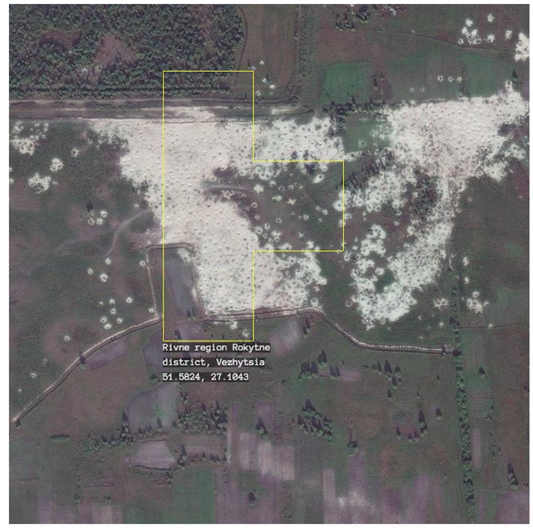
Think about a story you are working on. How do the answers to those questions change your approach to the story and whether you need Machine Learning for it?

Assess your use case

SINGLE STEP

In our case, these could be our answers:

I. We want our model to be able to recognise instances of amber mining in satellite images we will present to it.
II. We only need two categories: "YES: this image includes elements consistent with patterns that usually show amber mining activity" and "NO: this image doesn't include elements that suggest amber mining".
III. Mostly yes: instances of amber mining are quite recognisable in satellite images because of the distinctive pockmark-like pattern of holes in the ground. But we'll see in the testing phase that it might not always be as easy as we think.
IV. Different background, different density of the holes, different colours. The more diverse the examples in our dataset, the better the algorithm will learn.



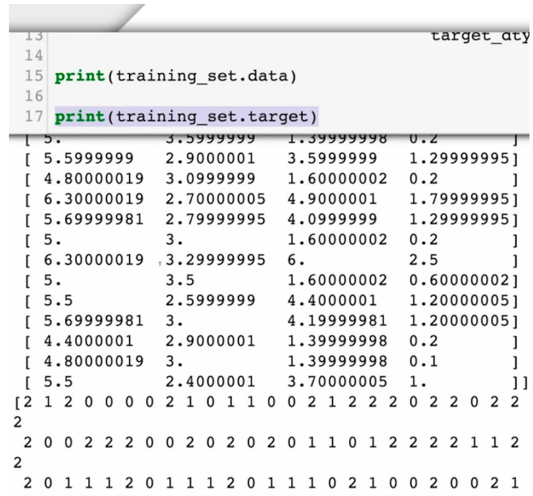
Source your data

SINGLE STEP

Once you've established what data you need, the next step is to find a way to source it. In our case, we already have the dataset provided by Texty. But think of what might be your own use case: How and where can you find the images you need?

You might be able to source them from what your organisation collects or from third-parties. In both cases, make sure to review regulations about data protection in your region and the locations your application will serve.

No training data will ever be perfectly "unbiased", but you can improve your chances of building a "fair" ML model if you carefully consider potential sources of bias in your data and take steps to address them. Review our Introduction to Machine Learning to find out more about it.



```
print(training_set.data)

print(training_set.target)
 [ 5.5999999   2.9000001   3.5999999   1.29999995]
 [ 4.80000019  3.0999999   1.60000002  0.2       ]
 [ 6.30000019  2.70000005  4.9000001   1.79999995]
 [ 5.69999981  2.79999995  4.0999999   1.29999995]
 [ 5.          3.          1.60000002  0.2       ]
 [ 6.30000019  3.29999995  6.          2.5       ]
 [ 5.          3.5         1.60000002  0.60000002]
 [ 5.5         2.5999999   4.4000001   1.20000005]
 [ 5.69999981  3.          4.19999981  1.20000005]
 [ 4.4000001   2.9000001   1.39999998  0.2       ]
 [ 4.80000019  3.          1.39999998  0.1       ]
 [ 5.5         2.4000001   3.70000005  1.        ]]
[2 1 2 0 0 0 0 2 1 0 1 1 0 0 2 1 2 2 2 0 2 2 0 2 2
2
 2 0 0 2 2 2 0 0 2 0 2 0 2 0 1 1 0 1 2 2 2 2 1 1 2
2
 2 0 1 1 1 2 0 1 1 1 2 0 1 1 1 0 2 1 0 0 2 0 0 2 1
```

Prepare your data

SINGLE STEP

There are a few more things to keep in mind as you put together the training data:

Include enough labelled examples in each category: The minimum required by AutoML Vision is 100 examples per label. In general, the more labelled images you can bring to the training process, the better your model will be.

It's important to include roughly similar amounts of training examples for each category. If you have an abundance of data for one label, use only part of it to avoid having a widely different amount of examples per category.

Find images that are visually similar to what you're planning to ask the model to categorise. Ideally, your training examples are real-world data drawn from the same dataset you're planning to use the model to classify.

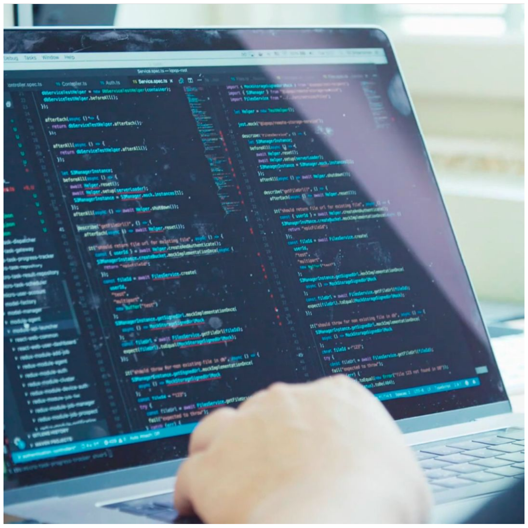
LESSON 5

Training your Machine Learning model

Import your data in AutoML Vision and start the training process

Lesson Overview

Prepare your data for import



It's time to go back to our Google Cloud account and continue the exercise by importing our training datasets to AutoML Vision.

The quickest way to add labelled images is to upload separate zipped folders containing examples for each label. In our case, we have two folders/labels: “positive” (images with examples of amber mining) and “negative” (without). You could also upload all the images together and label them manually inside the AutoML Vision interface but it would take much longer.

1	Import the data into AutoML (1)
2	Import the data into AutoML (2)
3	Training your Machine Learning model
4	Train, Validation and Test sets
5	Launch the training on AutoML

For more lessons, visit:

newsinitiative.withgoogle.com/training

Import the data into AutoML (1)

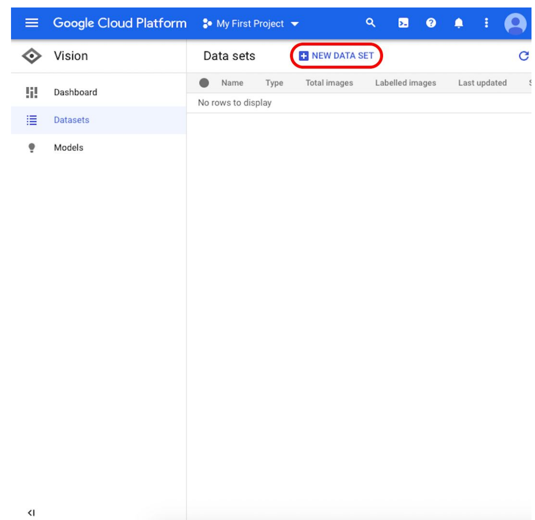
Download on your local disk the two zipped folders:

- Positive examples
- Negative examples

While they are being downloaded, re-open the Google Cloud platform via this link. Once the two folders have been downloaded to your local disk, follow these steps to upload them to AutoML Vision:

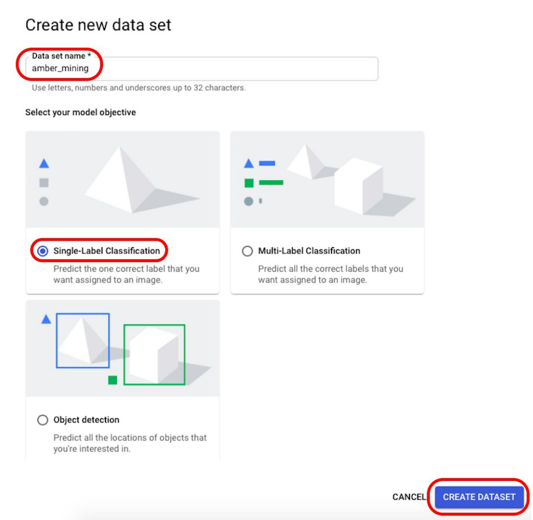
STEP 1 OF 5

From the interface, click on “New Dataset”.



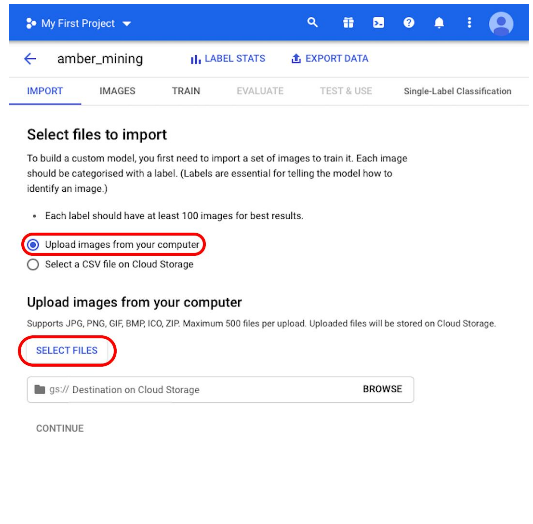
STEP 2 OF 5

Rename your dataset to something recognisable (for example, “amber_mining”), select “Single-Label Classification” as your model objective, and click on “Create dataset”.



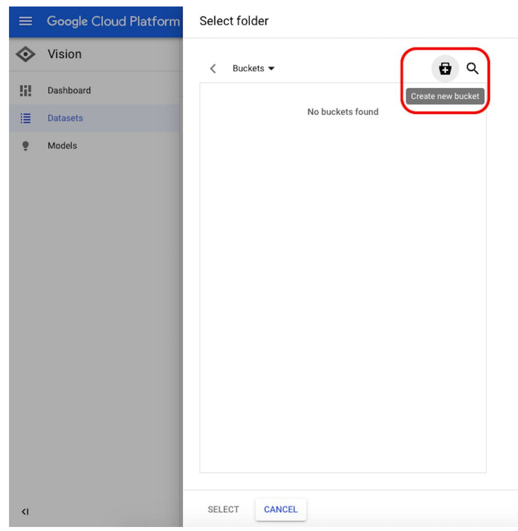
STEP 3 OF 5

Keep selected “Upload images from your computer” and click on “Select Files”. From the menu that will open, select both “positive.zip” and “negative.zip”. Confirm your selection.



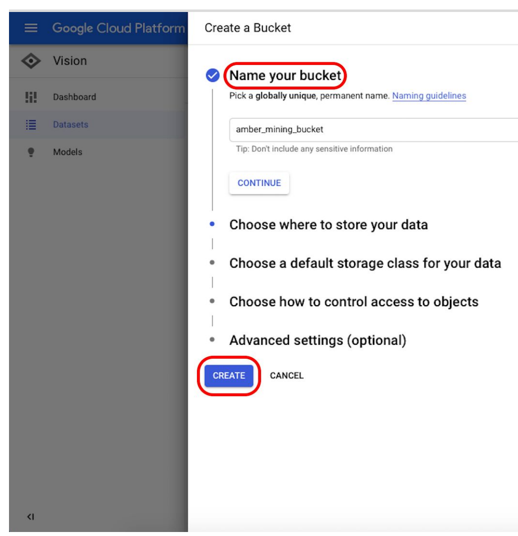
STEP 4 OF 5

Click on “Browse” to select a destination on Cloud Storage and in the window that will open, click on the icon in the top-right corner to “Create new bucket”.



STEP 5 OF 5

Give a name to your bucket. For the purpose of this exercise, it doesn’t matter what you select in the following options. Click on “Create” and then on “Select” in the next window.

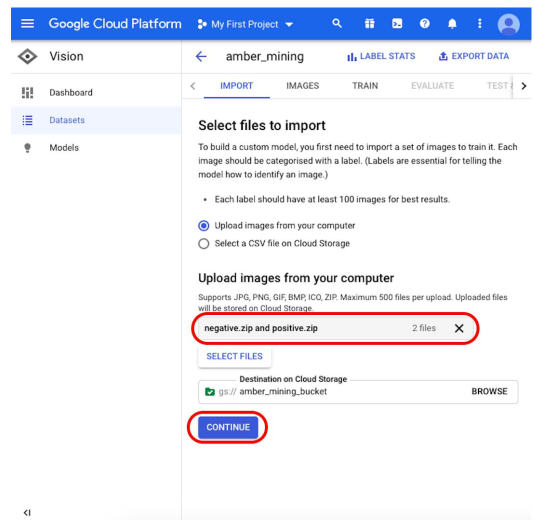


Import the data into AutoML (2)

We are now ready to upload the training sets:

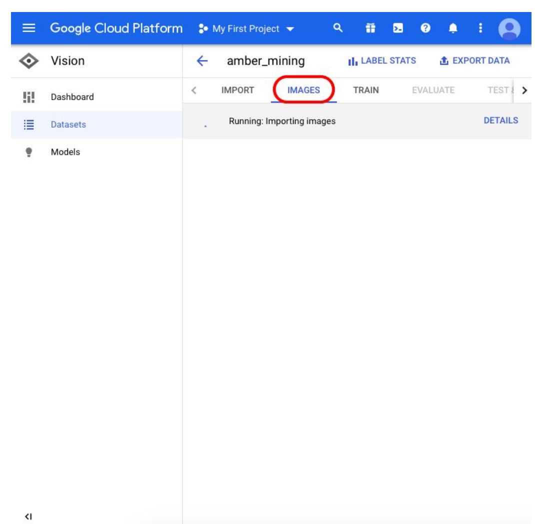
STEP 1 OF 3

Make sure that both “negative.zip” and “positive.zip” appear in the grey box and click on “Continue”. Wait a few seconds or a few minutes – depending on the speed of your connection – for the images to be uploaded.

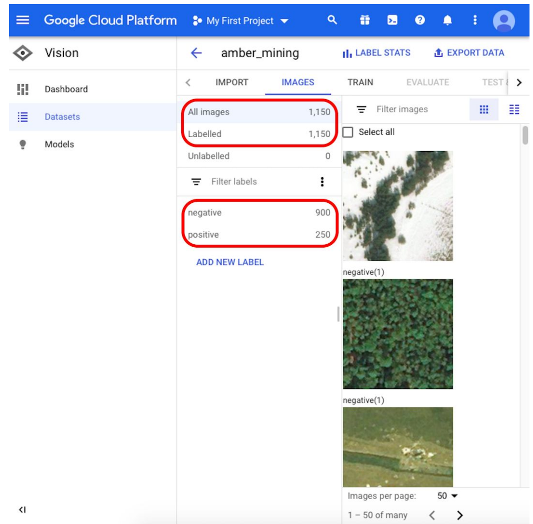


STEP 2 OF 3

When the upload is complete, click on “Images” from the menu on top of the page and wait for the import process to finish – it might take up to 30 minutes.



When the import process is done, you will be notified via email. Your Google Cloud Platform will show 1,150 images imported, 900 negative and 250 positive.

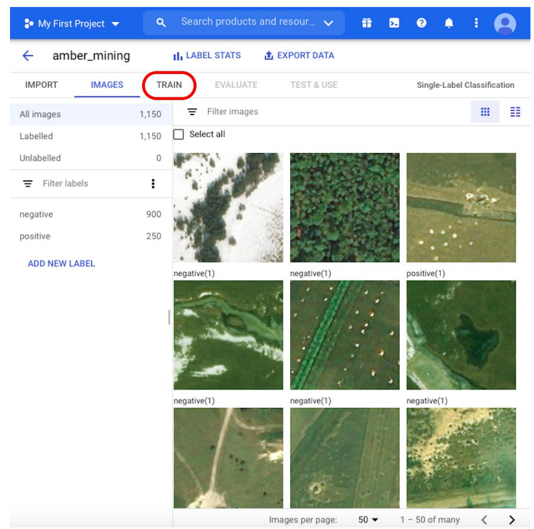


Training your Machine Learning model

We are now ready to start the training process. But first, browse through the images and learn more about our dataset. Check for example some of the "positive" images. Can you see the distinctive holes, trace of amber mining? If you can recognise it, then your model could do it, too.

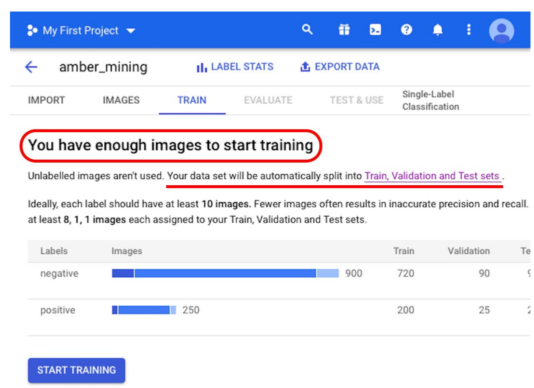
STEP # OF #

For some images, it might not be so easy even for yourself to tell if there are traces of amber mining or not. In the next lesson, we'll see how the model performs on those borderline examples. When you are ready to proceed, click on "Train"



STEP # OF #

At this point, the model tells you that "You have enough images to start training". It also informs you that "Your data set will be automatically split into Train, Validation and Test sets." Let's see what that means.



Train, Validation, and Test sets

SINGLE STEP

The reason to split our dataset into three separate sets is that we keep some images to the side, so that, after the model is trained, we can evaluate its performance using data it wasn't trained on – but that we know the right label for.

If you do not specify how many images to keep in each set, then AutoML Vision uses 80% for training, 10% for validating, and 10% for testing:

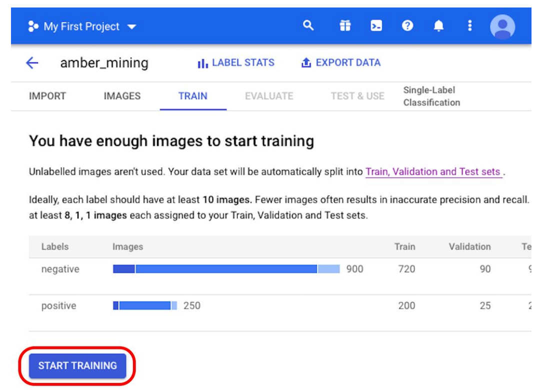
- The training set is what your model "sees" and initially learns from.
- The validation set is also part of the training process but it's kept separate to tune the model's hyperparameters, variables that specify the model's structure.
- The test set enters the stage only after the training process. We use it to test the performance of our model on data it has not yet seen.

Launch the training on AutoML

You can also split the dataset yourself. Manually splitting your data is a good choice when you want to exercise more control over the process or if there are specific examples that you're sure you want included in a certain part of your model training lifecycle. In our case, the default split – 80%, 10%, 10% – is perfectly fine, so we can proceed:

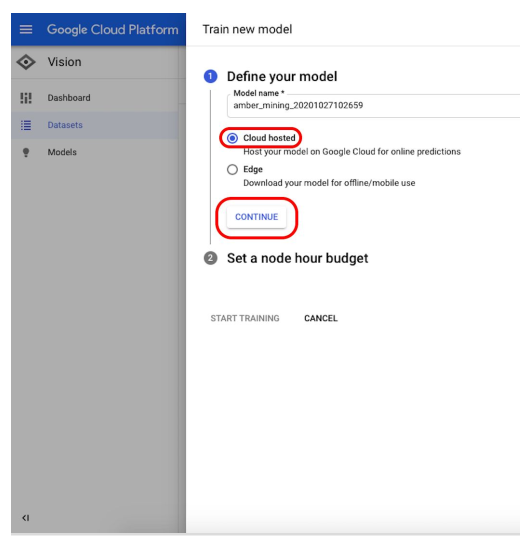
STEP 1 OF 5

Click on “Start Training”



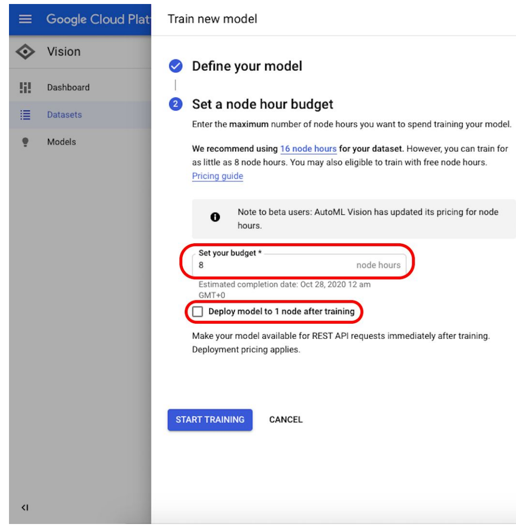
STEP 2 OF 5

Under “Define your model”, keep selected "Cloud hosted" and click "Continue".



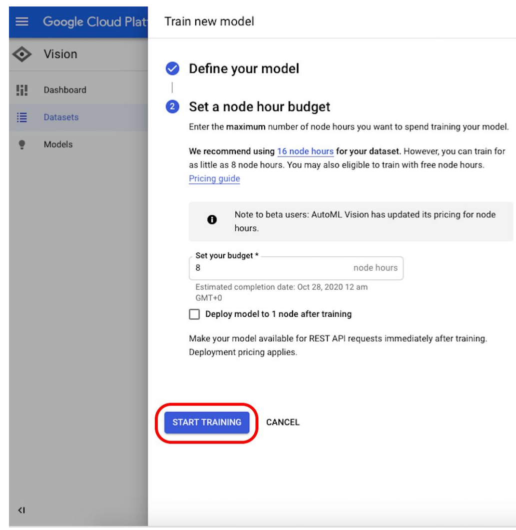
STEP 3 OF 5

Next, under "Set a node hour budget", set your budget to 8 node hours and leave the box "Deploy model to 1 node after training" unchecked.



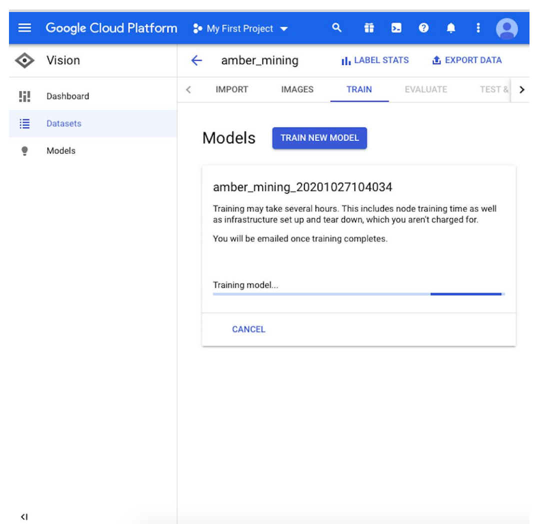
STEP 4 OF 5

Finally, click on "Start Training". Now, you have to wait a couple of hours while the model is trained.



STEP 5 OF 5

The machine is learning. You will be notified via email once training completes.



LESSON 6

Evaluate and Test

How to interpret the output of your model and evaluate its performance

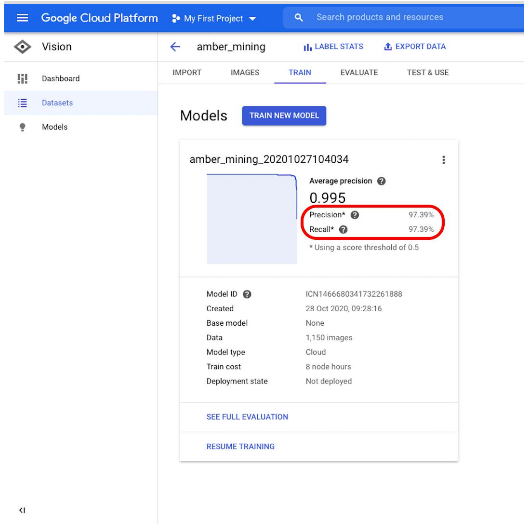
Lesson Overview

Precision and Recall

Once the model is trained, you will see a summary of the model performance with scores for "Precision" and "Recall".

Precision tells us what proportion of the images identified by the model as positive should indeed have been categorised as such. Recall instead tells us what proportion of actual positive images were correctly identified.

Our model performed very well in both categories, with scores above 97%. Let's see what that means in more detail.



1 Evaluate the model performance

2 False positives and False negatives

3 Test and train again

For more lessons, visit:

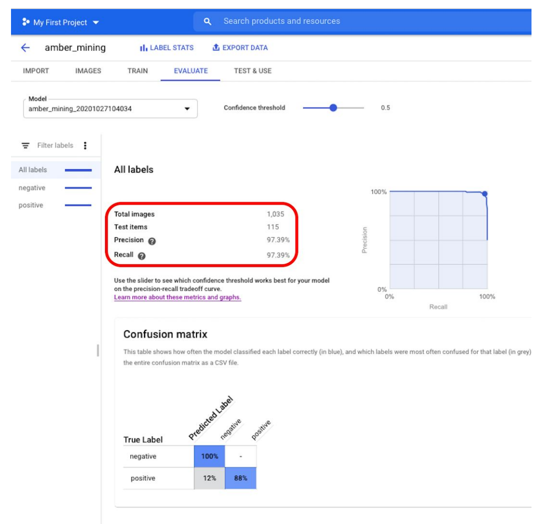
newsinitiative.withgoogle.com/training

Evaluate the model performance

Click on "Evaluate" on the top menu and let's explore the interface. First, it shows us again the scores on precision and recall. In our case, the precision score tells us that 97% of the test images that the model identified as examples of amber mining were indeed showing traces of amber mining.

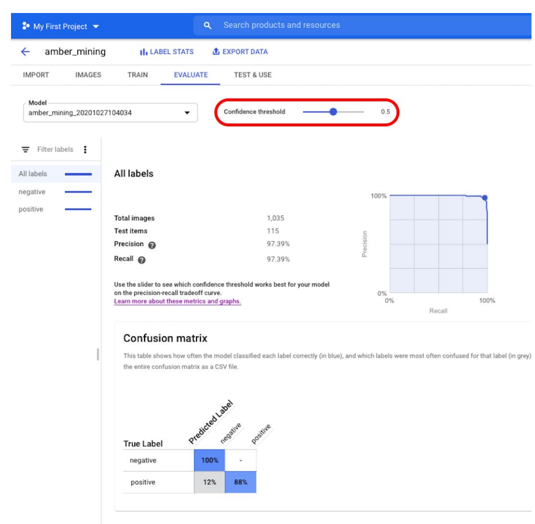
STEP 1 OF 3

The recall score instead tells us that 97% of the test images showing examples of amber mining were correctly labelled as such by the model.

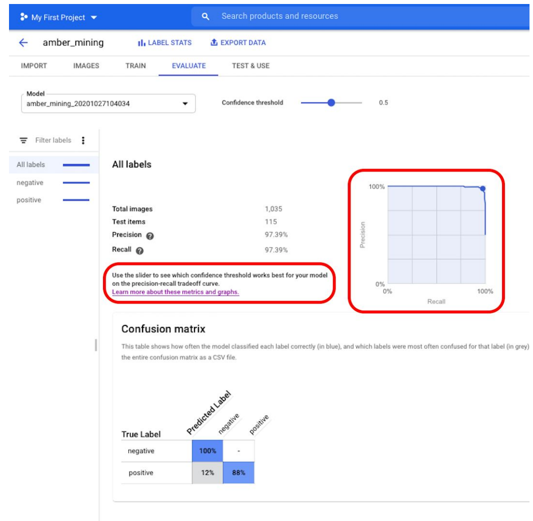


STEP 2 OF 3

Confidence threshold is the level of confidence the model must have to assign a label. The lower it is, the more images the model will classify, but the higher the risk of misclassifying some images.



If you want to dig deeper and also explore the precision-recall curves, follow the link on the interface to learn more.

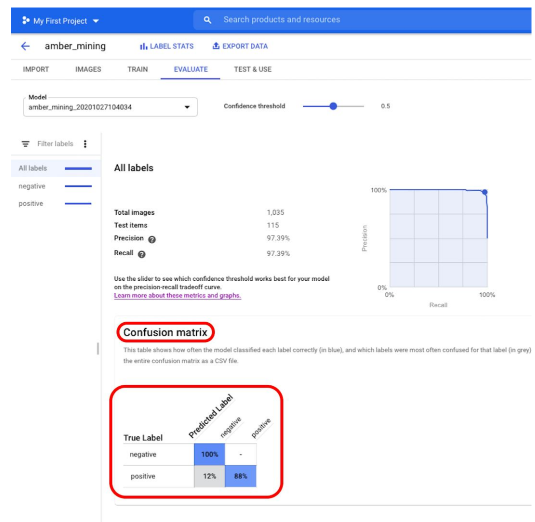


False positives and False negatives

Next, let's look at the Confusion Matrix. The higher the scores on blue background, the better the model performed. In this example, the scores are very good.

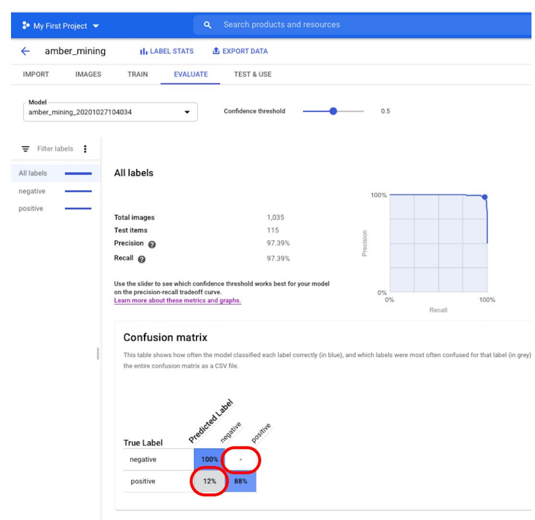
STEP 1 OF 5

All images that should have been labelled as negative (no amber mining) were recognised by the model and 82% of the images that included traces of amber mining were correctly labelled as such.



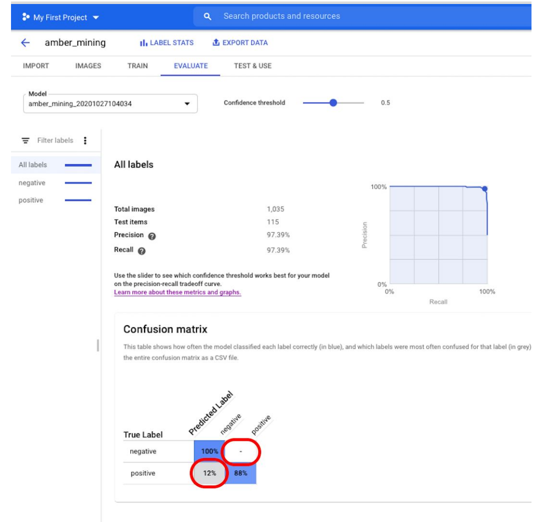
STEP 2 OF 5

We have no false positives – no images were wrongly labelled as examples of amber mining – and only 12% of false negatives: images showing traces of amber mining that the model failed to recognise.



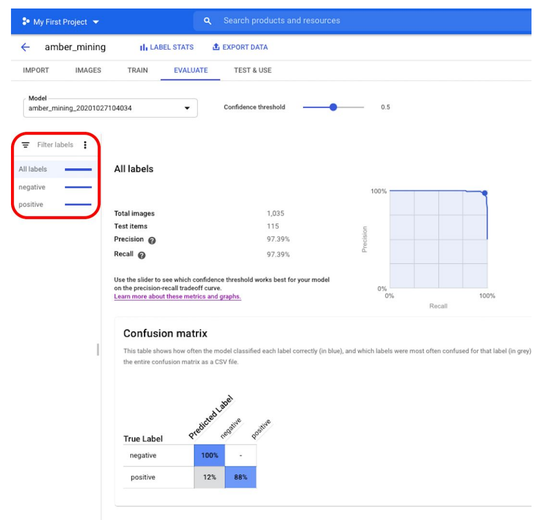
STEP 3 OF 5

This is good for the purpose of our investigation into illegal amber mining: it's better to miss some positive examples than to bring as proof of amber mining images that do not actually show that.



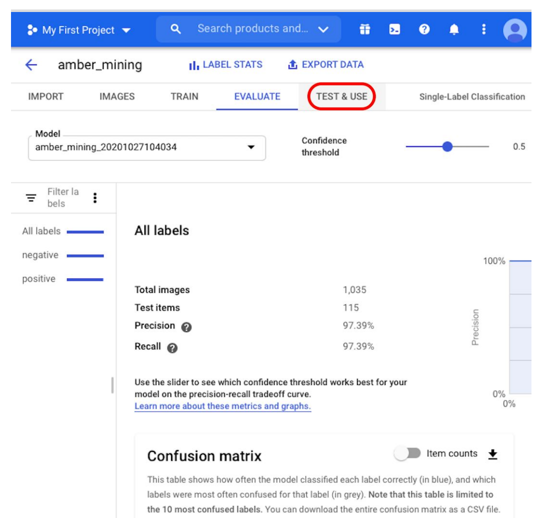
STEP 4 OF 5

Click on the left filters if you want to see which test images were correctly or wrongly classified by the model.



STEP 5 OF 5

Not yet sure if you can trust the model? By clicking on “Test & Use”, you can upload brand-new satellite images – with or without traces of amber mining – to see if the model labels them correctly.



Test and train again

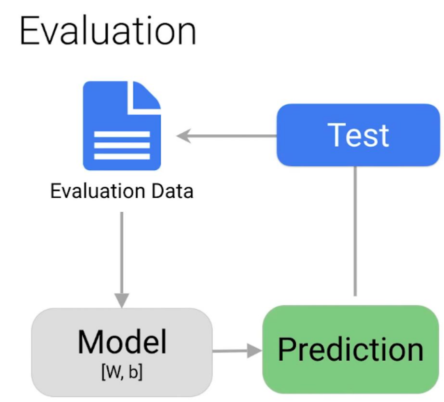
SINGLE STEP

A few final considerations before we wrap up:

You might be wondering how the model is getting some wrong answers when we told it all the right answers to begin with. If you are, you might want to review the split into training, validation, and test sets described in the previous lesson.

For this example, almost all of the images were classified correctly. But that will not always be the case. If you are not satisfied with your model's performance, you can always update and improve your dataset and train the model again. You could carefully analyse what went wrong in the first iteration and, for example, add to your training set more images similar to those that were misclassified by the model.

As for humans, learning is an iterative process.



Google News Initiative

LESSON 7

Looking ahead to ML-powered journalism

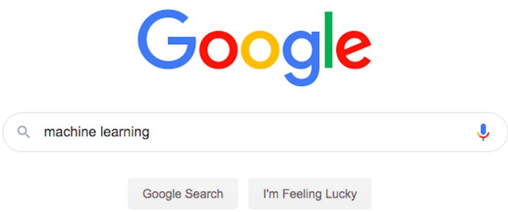
A step-by-step summary and where to learn more

Lesson Overview

Moving the next steps

You have reached the end of this crash course on classifying images with machine learning. You should now have a better understanding of how machine learning works in practice, as well as what its power and limitations are.

If you have previously taken our [Introduction to Machine Learning](), you now have the perfect toolkit to understand how to enhance your reporting with machine learning and how you can make the most, in a responsible way, of the power offered by these technologies.



#	Summary of the training steps
#	Where to learn more
#	Credits

For more lessons, visit:

newsinitiative.withgoogle.com/training

Summary of the training steps

SINGLE STEP

Let's recap all the steps involved in training a machine learning model:

1. Reflect on your use case and consider whether ML can be (part of) the solution to the problem you are trying to solve.
2. Source the data you need and take every possible step to minimise the potential impact of biases in the data.
3. Clean and prepare the data so that it contains the right balance of information the model needs to learn from.
4. Choose the algorithm that is best suited to meet the goals of your use case and the features of your training data.
5. Upload the dataset you prepared to the algorithm of your choice and wait for it to learn.
6. Evaluate the results and decide if they are good enough to use the model for your journalistic purpose. If not, repeat.

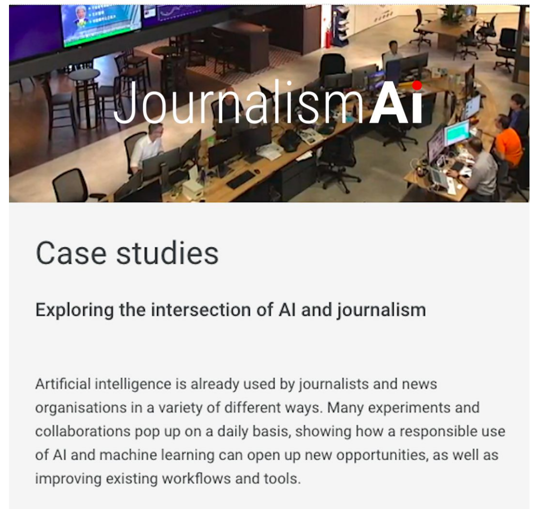
Where to learn more

SINGLE STEP

In this crash course, we learned how machine learning can be used to classify images in the context of an investigation, but there are many other journalistic use cases. Find a few examples in this handy collection by the Quartz AI Studio that highlights instances of "how you might feel when machine learning can help".

If you want to know more about how news organisations use machine learning and other AI-powered technologies, browse the ever-growing library of case studies curated by the JournalismAI team.

And never forget to carefully reflect on how to identify potential biases in your data and make sure that they are not replicated and multiplied by your ML model. Check out this Google Cloud guide on Inclusive ML to learn more.



Credits

SINGLE STEP

This course was developed by JournalismAI in collaboration with Anatoliy Bondarenko and his team at Texty.

Texty is a Ukrainian data journalism agency that promotes transparency and accountability by developing high-quality journalism and data journalism, which includes analysis and presentation of big data in an interesting and comprehensive way.

JournalismAI is a project of POLIS – the journalism think-tank at the London School of Economics and Political Science – and it's funded by the Google News Initiative.

Special thanks to Agnes Stenbom, Fabienne Meijer, Florencia Coelho, and Jarno Koponen, for their precious feedback during the development of the course.

Sign up to the JournalismAI newsletter to stay informed about project activities.

JournalismAi

Google News Initiative

TEXTY.ORG.UA