

# Solution: Ground Truth

Before we get started with Ground Truth, we'll need data to use in the process. In order to do that you'll need to upload the `sentiment_data.jsonl` file that is in the Github repository under this lesson's starter files. Once uploaded into a bucket, make sure you know the name of the bucket you used. We'll need it later.

⌚ Upload succeeded  
View details below.

### Upload: status

The information below will no longer be available after you navigate away from this page.

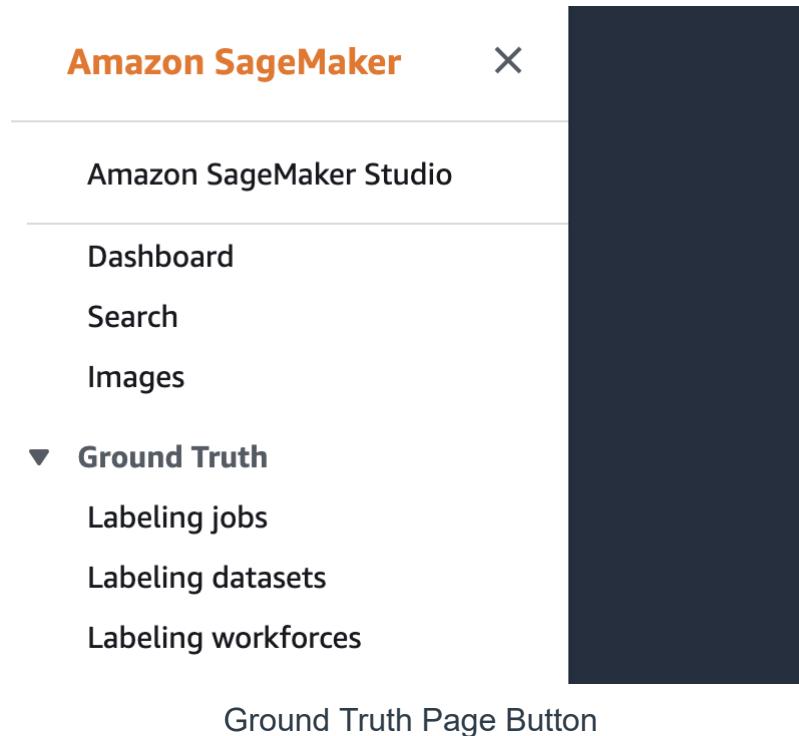
Summary	
Destination	Succeeded ⌚ 1 file, 449.0 B (100.00%)
	Failed ⌚ 0 files, 0 B (0%)

**Files and folders** **Configuration**

Files and folders (1 Total, 449.0 B)											
<input type="text"/> Find by name < 1 >											
Name	▲	Folder	▼	Type	▼	Size	▼	Status	▼	Error	▼
sentiment_data.jsonl	-	-	-	-	-	449.0 B	-	⌚ Succeeded	-	-	-

S3 Data Upload

Open up the Sagemaker console in the AWS Gateway and navigate to the **Ground Truth** - **Labeling jobs** tab.



Ground Truth Page Button

When you create a labeling job, there is a bit of setup required. You'll make sure to use the **Manual data setup** option for how to input the data, as well as the S3 location of the data file you just uploaded. For the dataset output, you can just use the same S3 bucket.

You'll need a proper Sagemaker role to execute this job. If you have not already created one from earlier in the course, you can create one now.

We'll also be using the full dataset.

## Job overview

### Job name

Maximum of 63 alphanumeric characters. Can include hyphens (-), but not spaces. Must be unique within your account in an AWS Region.

#### I want to specify a label attribute name different from the labeling job name.

Label attribute name is the key where your labels are stored in the augmented manifest. Ground Truth uses the labeling job name as the default label attribute name.

### Input data setup Info

Use the automated setup to have Ground Truth automatically identify your dataset in S3. Use the manual setup if you have an input manifest file.

#### Automated data setup

Provide the S3 location of the dataset you want labeled and let Ground Truth automatically connect to and use this dataset for your job.

#### Manual data setup

Provide the S3 location of a file (an input manifest file) that identifies the data objects you want labeled

### Input dataset location Info

Provide a path to the S3 location where your manifest file is stored.

### Output dataset location Info

Provide a path to the S3 location where you want your labeled dataset to be stored.

### IAM Role Info

Provide the ID or ARN for your own AWS KMS encryption key for Amazon SageMaker to access your S3 bucket. Choose a role or let us create a role with the [AmazonSageMakerFullAccess](#) IAM policy attached.

### ▼ Additional configuration - optional

Dataset object selection, encryption

#### Dataset object selection Info

You can use the full dataset or create a subset of your data.

#### Full dataset

Use all the dataset objects

#### Random sample

Create a subset by specifying a sample size

#### Filtered subset

Create a subset by specifying a query

## Job Overview Initial

The task type we are interested in is **Text Classification (Single Label)**. We are just interested in sentiment that is positive, negative, or neutral.

## Task type [Info](#)

### Task category

Select the type of data being labeled to view available task templates for it or select 'Custom' to create your own.

Text



### Task selection

Select the task that a human worker will perform to label objects in your dataset.

#### Text Classification (Single Label)

Get workers to categorize text into individual classes. [Info](#)

- Positive
- Negative

*'The movie tells a lovely and wise story with honesty and has been acted out with unassuming grace.'*

#### Text Classification (Multi-label)

Get workers to categorize text into one or more classes. [Info](#)

- Positive
- Inspiring
- Jargon

*'Every day is a fresh start. Always start your day with a cup of positivitea.'*

## Job Overview Task

For the **Workers** section, we'll be keeping this private. More importantly, we'll just be emailing the labeling job to ourselves so we can test it out. Most of the defaults in this section are great the way they are.

# Select workers and configure tool

Workers [Info](#)

Worker types

Amazon Mechanical Turk  
An on-demand 24/7 workforce of over 500,000 independent contractors worldwide powered by Amazon Mechanical Turk.

Private  
A team of workers that you have sourced yourself, including your own employees or contractors for handling data that needs to stay within your organization.

Vendor managed  
A curated list of third party vendors that specialize in providing data labeling services, available via the AWS Marketplace.

Team name

sentiment

Maximum of 63 alphanumeric characters. Can include hyphens, but not spaces. Must be unique within your account in an AWS Region. The name can't be changed later.

Invite private annotators

Enter email addresses of workers that will work on this job.

*Add email addresses, separated by commas to add annotators.*

⚠ This field is required.

Enter up to 20 addresses and use a comma between each one.

Task timeout

The maximum time a worker can work in a single task. If you want to use values beyond 8 hours, contact AWS Support.

0 hours 5 mins 0 secs

Task expiration time

The amount of time that a task remains available to workers before expiring. If you want to use values beyond 10 days, contact AWS Support.

10 days 0 hours 0 mins 0 secs

## Select Private Worker

Some additional detail is needed, for **Organization** you can make up your own. Try and be creative! The **email** can be the same email you are using for annotation above.

**Organization**

We use this information to customize the worker invitation.

 This field is required.

**Contact email**

Workers can use this to report issues related to the job.

 This field is required.

Enter one email address only.

**Enable automated data labeling** [Info](#)

Amazon SageMaker will automatically label a portion of your dataset. It will train a model in your AWS account using Built-in Algorithm and your dataset. When you enable this, training jobs use new computing resources on your behalf. For cost information, See SageMaker [pricing](#) 

► **Additional configuration - optional**

Workers per dataset object

## Email filling

We will define how we want to label our data in this section. The most important part is creating the **positive**, **negative**, and **neutral** labels on the right-hand side. Since we are just sending the job to ourselves, everything else in this section is optional. In the real world, we would need to be as descriptive as possible in order to not confuse the annotator.

## Text classification labeling tool (Single Label)

[Preview](#)

Provide labeling instructions with examples below for workers. Workers will be viewing these instructions when they perform your task. Workers can choose up to 30 labels. See guidelines for [See guidelines for creating high-quality instructions](#)

H<sub>1</sub> H<sub>2</sub> B I A

Enter positive, negative, or neutral.

label sentiment

The movie was great, lots of action and top notch movie stars. I'll be watching this one on repeat for months.



● ● ● ●

### Select an option

Add up to 30 labels

positive

negative

neutral

Add new label

You can add 27 more labels.

► Additional instructions - *optional*

Cancel

Previous

Create

### Label Options

After you create the labeling job, you will be presented with the overall summary. AWS will need to provide the necessary tooling around the annotation process, so wait a few minutes for it to set up.

## sentiment

**Labeling job summary**

[View labeling tool](#)

Status ⌚ In progress	Input dataset location <a href="#">s3://sagemaker-studio-466356262471-xwyfq86g7hr/sentiment_data.jsonl</a>	ARN <a href="#">arn:aws:sagemaker:us-east-1:466356262471:labeling-job:sentiment</a>
Task type Text Classification (Single Label)	Output dataset location <a href="#">s3://sagemaker-studio-466356262471-xwyfq86g7hr/sentiment/</a>	Worker type Private
Creation time Jun 13, 2021, 11:21 PM UTC	Labeled / total dataset objects 4 / 4	Workteam ARN <a href="#">arn:aws:sagemaker:us-east-1:466356262471:workteam/private-crowd/sentiment</a>

**Output** | Tags

**Labeled dataset objects (4)**

[Query output](#)

Text	Label	Source	Confidence	Label creation time
The movie was great, lots of action and ...	-	-	-	-
This is the 3rd sequel to the franchise ...	-	-	-	-
It was the best movie ever!...	-	-	-	-
It was the worse movie ever!...	-	-	-	-

## Labeling Job Summary

Shortly after creating the labeling job, you should receive an email from AWS regarding contributing to the job. You'll click through the email and use the necessary credentials to log into your Jobs panel. If you don't see your recently created labeling job, wait a few more minutes till it's ready. Eventually, you'll see something similar to below. When you do, **Start working** on the task and finish it.

**Jobs (1)**

[Start working](#)

Task title	Customer ID	Status	Creation time
Text Classification (Single Label): label sentiment	466356262471	Available	June 13, 2021 23:23:51 UTC

## Private Job

After you finished the task, you can **complete** the labeling job by stopping it from the **Actions** dropdown on the top right side of the summary page. After the job finishes, you'll be able to see results. In the **Output dataset location** navigate to S3 bucket and download the **output.manifest** file.

## sentiment

Actions ▾

## Labeling job summary

View labeling tool

Status	Input dataset location	ARN
COMPLETE	s3://sagemaker-studio-466356262471-xwyfq86g7hr/sentiment_data.jsonl	arn:aws:sagemaker:us-east-1:466356262471:labeling-job/sentiment
Task type	Output dataset location	Worker type
Text Classification (Single Label)	s3://sagemaker-studio-466356262471-xwyfq86g7hr/sentiment/	Private
Creation time	Labeled / total dataset objects	Workteam ARN
Jun 13, 2021, 11:21 PM UTC	4 / 4	arn:aws:sagemaker:us-east-1:466356262471:workteam/private-crowd/sentiment

Output

Tags

## Labeled dataset objects (4)

Query output

&lt; 1 ... &gt; ⌂

Text	Label	Source	Confidence	Label creation time
The movie was great, lots of action and ...	positive	Human annotated	0	Jun 14, 2021, 6:28 AM UTC
This is the 3rd sequel to the franchise ...	negative	Human annotated	0	Jun 14, 2021, 6:28 AM UTC
It was the best movie ever!...	positive	Human annotated	0	Jun 14, 2021, 6:28 AM UTC
It was the worse movie ever!...	negative	Human annotated	0	Jun 14, 2021, 6:28 AM UTC

The output file should look like this. If it does not, make sure you received the labeling task in your email and you completed it before stopping the labeling job.

**output.manifest File**

```
{"source": "The movie was great, lots of action and top notch movie stars. I\u0027ll be watching again.", "sentiment": 0, "sentiment-metadata": {"class-name": "positive"}, "confidence": 0}, {"source": "This is the 3rd sequel to the franchise and you can definitely tell. It\u0027s been a while since the first one.", "sentiment": 0, "sentiment-metadata": {"class-name": "neutral"}, "confidence": 0}, {"source": "It was the best movie ever!", "sentiment": 1, "sentiment-metadata": {"class-name": "positive"}, "confidence": 0}, {"source": "It was the worse movie ever!", "sentiment": 0, "sentiment-metadata": {"class-name": "neutral"}, "confidence": 0}
```

**QUIZ QUESTION**