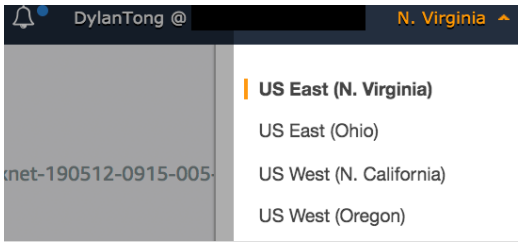
Lab 1: Creating a Quality Training Set using Amazon SageMaker Ground Truth

Most Computer Vision projects involve *supervised* machine learning, which requires a training data set that contains ground truth information. For instance, an object detection model is typically trained on images that are annotated with bounding boxes around objects that you wish to detect, and corresponding labels that specify the class of an object—is it a person, a car, or a bird?

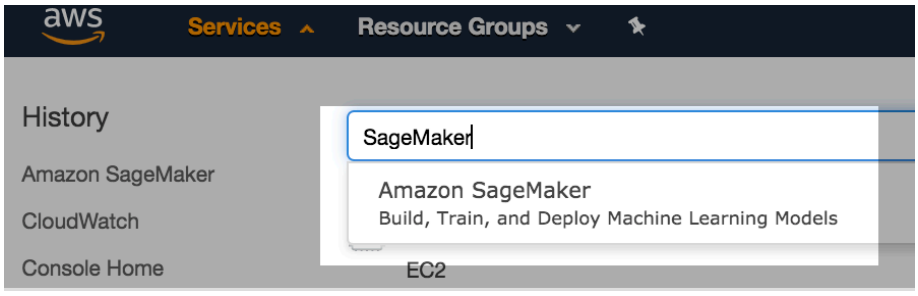
In this lab, we'll learn how to we can leverage Amazon SageMaker Ground Truth to create a training data set at scale. Challenges of annotating data at scale include dealing with the sheer volume of training data that needs to be produced on an ongoing basis and in a timely manner, managing the quality of the annotations, and integrating workflows with your machine learning tools and processes.

I. Prepare your Development Environment

1. Log into your AWS account and ensure you're in the right region designated for your workshop. The screenshot below indicates that I'm currently in us-east-1 (N. Virginia).

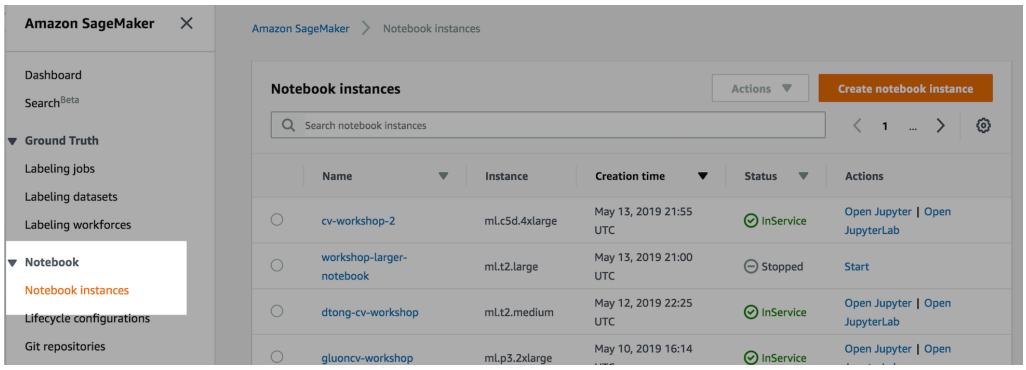


2. Navigate to the Amazon SageMaker console via the search bar.

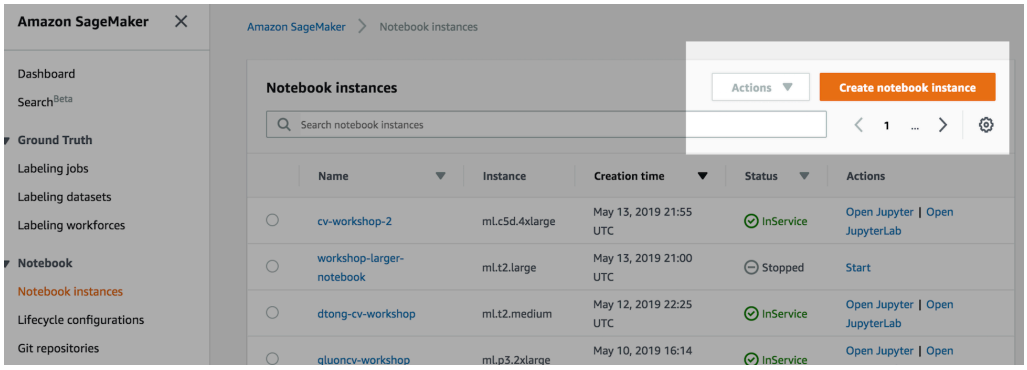


3. Next, launch a managed Amazon SageMaker notebook instance. We're going to use this notebook instance to run a number of labs. In this lab, the notebook will be used to stage some raw data that we will annotate.

Switch to the **Notebook Instances** page by using the navigation menu on the left hand side of the console.

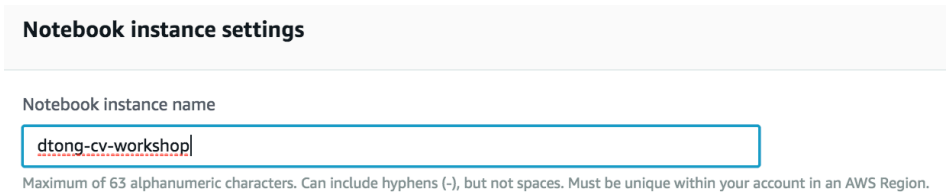


4. Click on the **Create notebook instance button**.

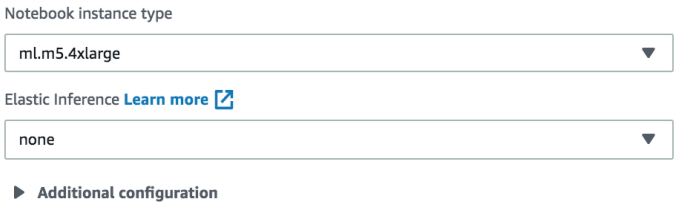


5. Next, we configure our notebook instance by working through the launch wizard. First, provide a **name** for your notebook.

   Utilize a unique prefix that you can remember, so you more easily find the resources that belong to you.

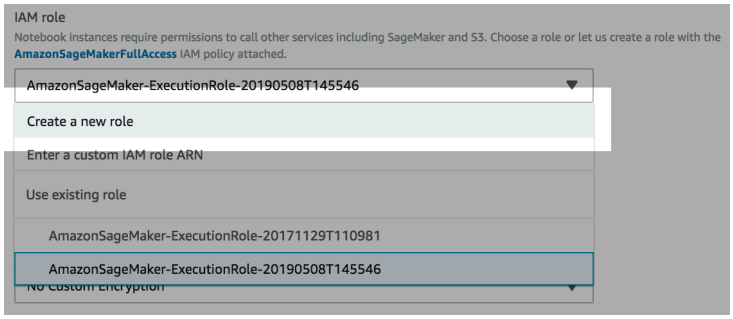


6. Select the ***ml.m5.4xlarge*** instance type.



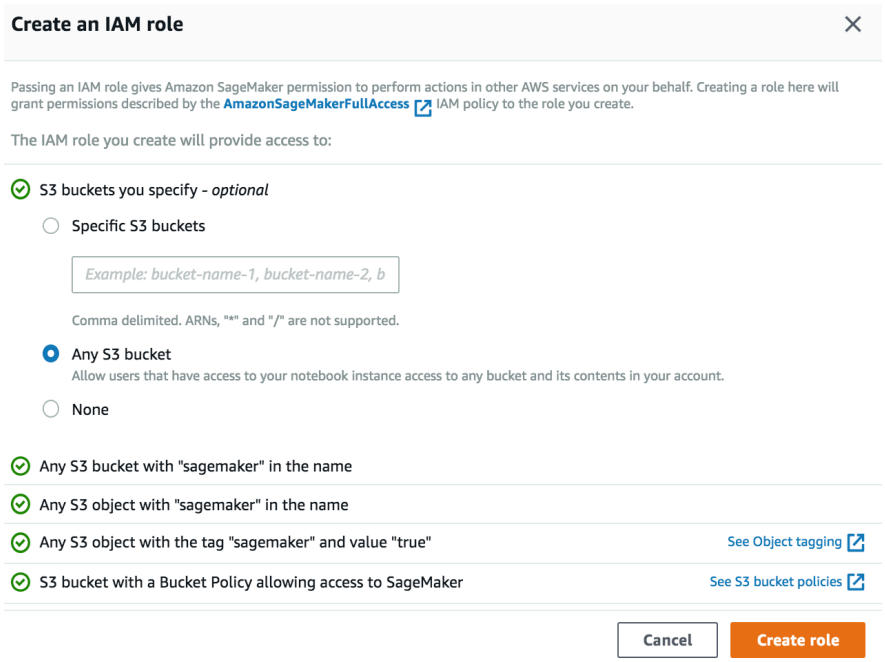
7. Your instance requires permissions to access data on S3, and execute SageMaker functionality required by this lab.

   Create a new role for your notebook. You can select an existing role if you have already done this previously.



8. For the purpose of simplicity, select **"Any S3 bucket."** If this wasn't a lab, you should practice the principles of least privilege.

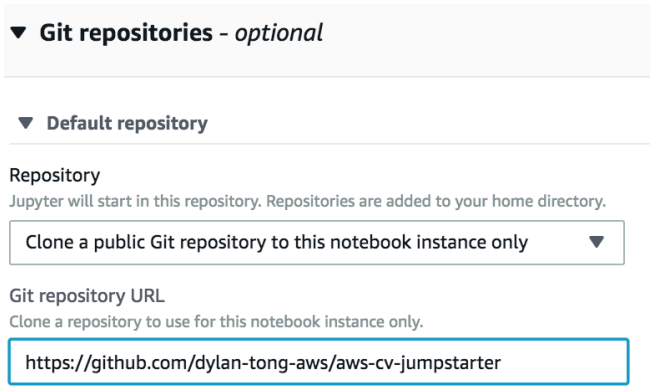
   Click on the **Create role** button. Ensure that the role that you just created is selected in your notebook configurations.



9. Next, configure the Git integration. We're going to launch the notebook and clone the lab repository over to your notebook instance.

   Select **"Clone a public Git repository to this notebook instance only."**

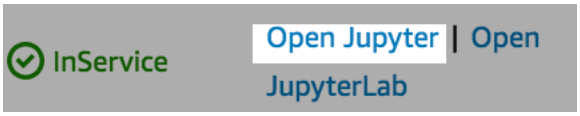
   Paste the following link into the text box under **"Git repository URL":** https://github.com/dylan-tong-aws/aws-cv-jumpstarter



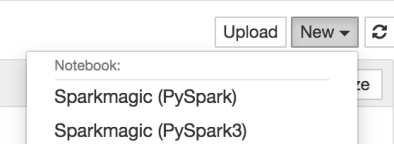
10. These basic configurations will suffice for the lab. In a production setting, you will likely want to launch this notebook into a VPC for better network security. Life-cycle configurations also come in handy if you like to automatically bootstrap your notebook instances with packages that aren't already pre-installed by default.

Click on **"Create notebook instance"** to launch your instance.

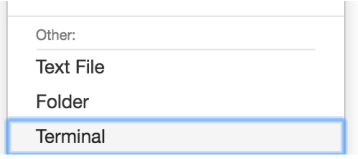
11. It will take about 5 minutes before your notebook is **InService**. Once it is, click on the **"Open Jupyter"** link.



12. Select the **"New"** drop down on the right hand side of the Jupyter admin console.



Scroll to the bottom and select **Terminal**.



13. We're going to create an S3 bucket from the terminal. The AWS CLI has been pre-installed, and it inherits the IAM permissions of the role that you attached to the instance.

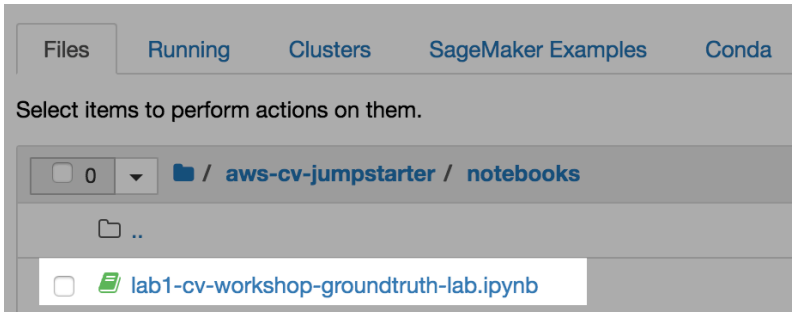
    Run the following command and replace the parts that are **high-lighted in red** with appropriate values. First, your bucket needs a **unique name**. Secondly, you need to create your bucket in the **same region as your notebook instance**. The example below will create the bucket in Oregon (us-west-2).

    *aws s3api create-bucket --bucket **dtong**-cv-jumpstarter-workshop --region **us-west-2** --create-bucket-configuration LocationConstraint=**us-west-2***

    The output should look like the following:

```
sh-4.2$ aws s3api create-bucket --bucket dtong-cv-jumpstarter-workshop --region us-west-2 --create-bu
cket-configuration LocationConstraint=us-west-2
{
    "Location": "http://dtong-cv-jumpstarter-workshop.s3.amazonaws.com/"
}
sh-4.2$
```

14. Return to the Jupyter admin console, and launch the Jupyter notebook for lab1 by clicking on the notebook as shown below:



15. Update the BUCKET variable with the name of the S3 bucket that you created previously from the terminal.

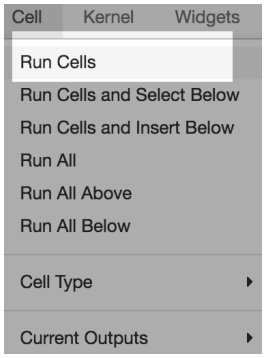
Replace << YOUR S3 BUCKET NAME >> with the name of your bucket.

```
BUCKET = '<< YOUR S3 BUCKET NAME >>'
S3_PREFIX = 'ground-truth-lab' # Any valid S3 prefix.
```

For instance, the following is the name of the bucket that I created.

```
BUCKET = 'dtong-cv-jumpstarter-workshop'
S3_PREFIX = 'ground-truth-lab' # Any valid S3 prefix.
```

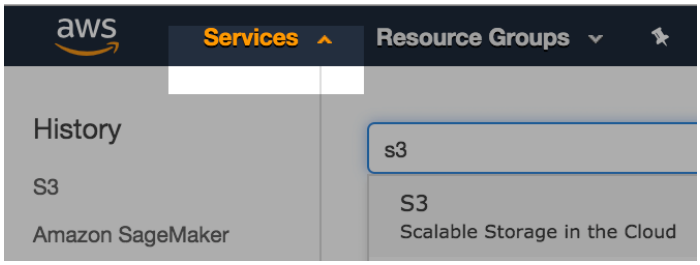
16. **Run through all the cells** that you created. This notebook will transfer raw data over to your S3 bucket.



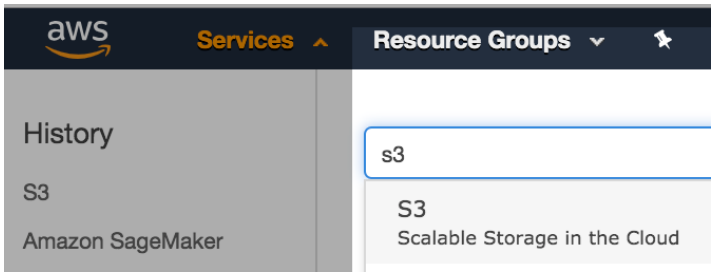
You should see a final output like the following. The process copies over 10 images, which we will annotate in the following steps.

```
bbox_labels_600_hie 100%[===================>]  84.27K  --.-KB/s    in 0.07s

2019-05-12 22:35:32 (1.25 MB/s) - 'bbox_labels_600_hierarchy.json.1' saved [86291/86291]

Copying image 0 / 10
Done!
```

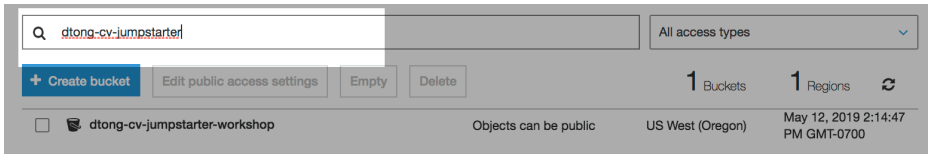
17. Let's validate what we've copied over into our S3 bucket. **Switch over to the S3 console**.



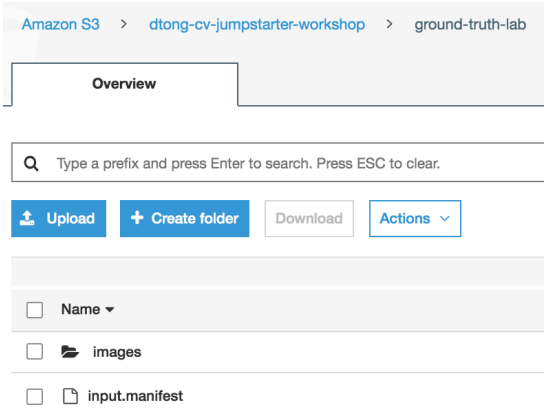
Use the search bar to link over to the S3 console.



18. **Select your bucket**. Use the search bar to filter the bucket list.



19. The contents of the bucket should appear as follows:



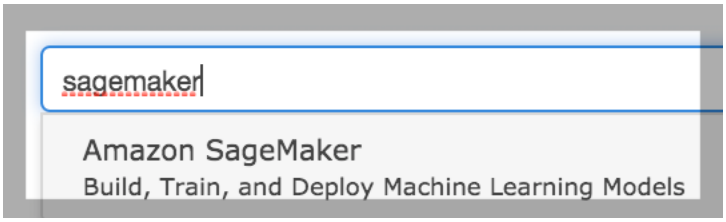
The images folder contains 10 images. If you download the input manifest, the contents should be like the following:

```
…
{"source-ref": "s3://dtong-cv-jumpstarter-workshop/ground-truth-lab/images/000062a39995e348.jpg"}
{"source-ref": "s3://dtong-cv-jumpstarter-workshop/ground-truth-lab/images/000411001ff7dd4f.jpg"}
…
```

The manifest file will be used by Amazon SageMaker GroundTruth to find the images that you wish to annotate. **Take note of the full path to this file**.

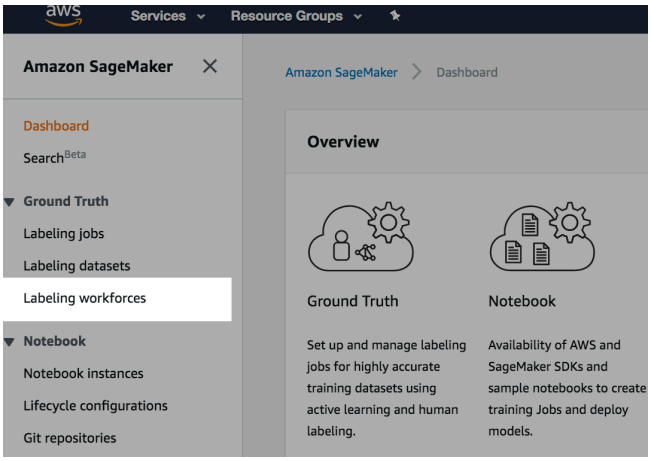
II. Prepare Your Labeling Workforce

20. Return to the **SageMaker console**.



21. Next, we're going to configure our a **Labeling workforce** to perform annotation on our behalf.

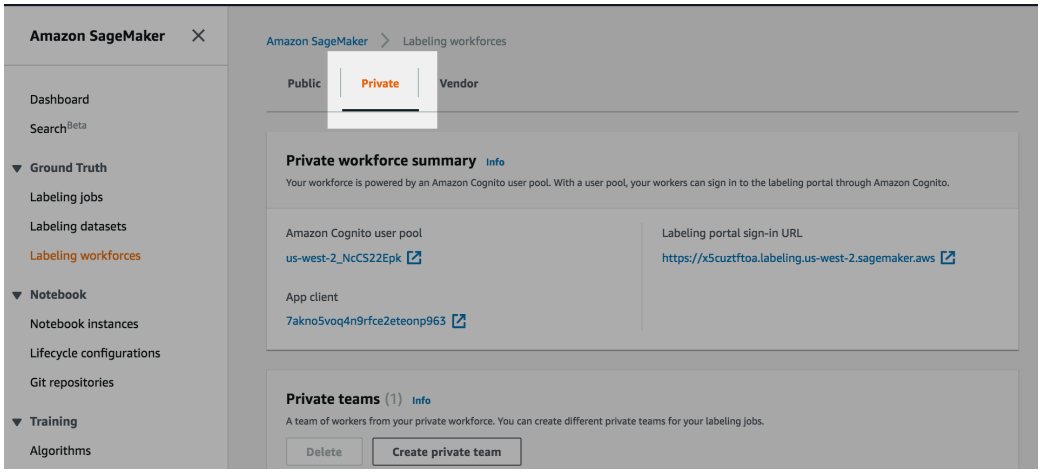
    Select "Labeling workforce" from the navigation pane on the left hand side of the console.



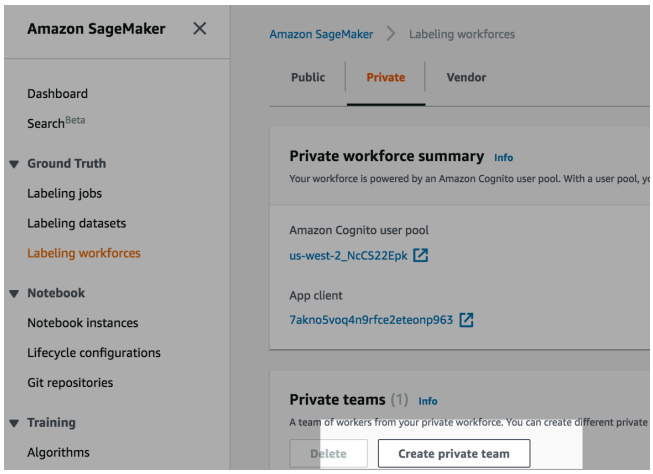
22. SageMaker GroundTruth allows you to register labeling workforces from Mechanical Turk (Public), 3rd-party vendors registered in the AWS Marketplace (Vendor) as well as from our organization (Private).

    We're going to create a private workforce in this lab consisting of yourself for the purpose of demonstration. This should be sufficient to give you a sense of how the other options would work in practice.

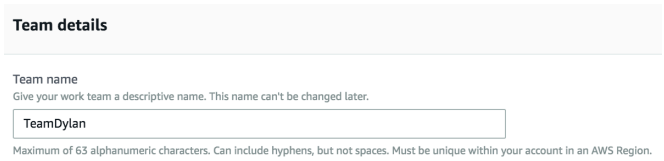
    Select the **Private** sub tab from the **Labeling workforce page**.



23. Select **Create private team**.

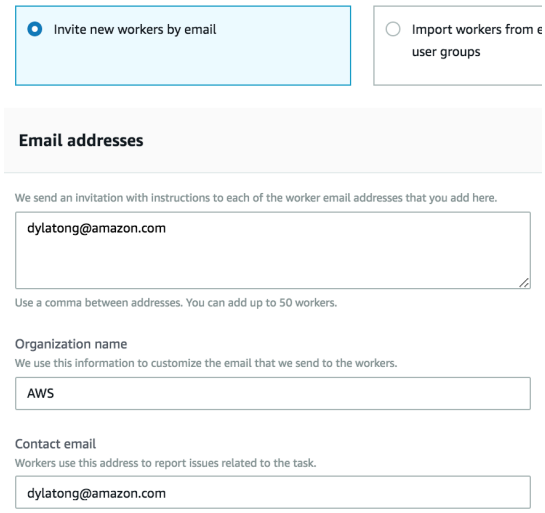


24. Enter a **unique name** for your team.

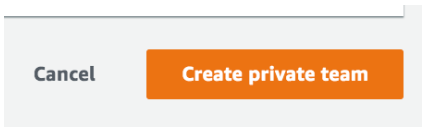


25. Select **“Invite new workers by email.”**

    Fill out the form as shown below with your email address in place of dylatong@amazon.com. **You need to provide an email address that you have access to.**

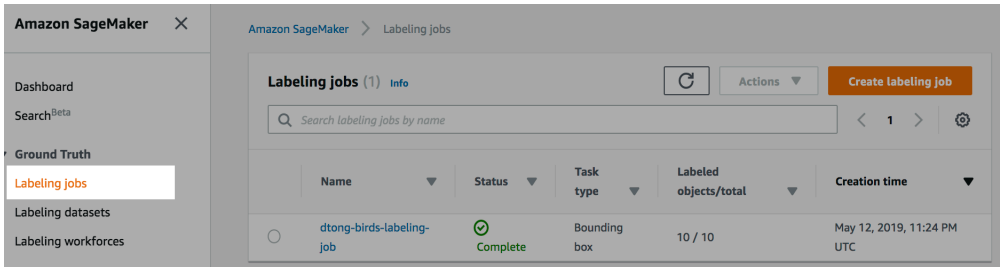


26. Leave the rest as default. You’re done with creating your private team. Click on the **“Create private team”** button.

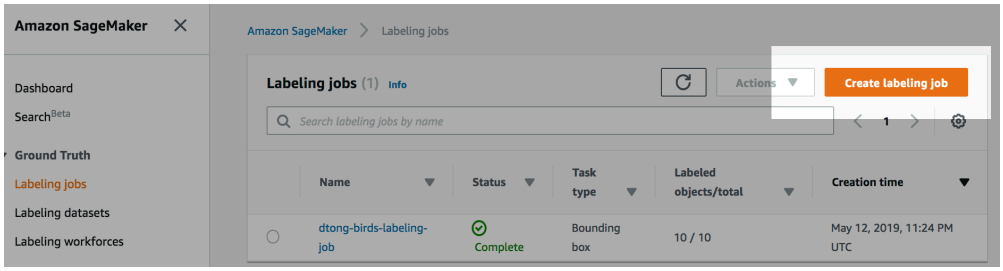


III. Define your Labeling Job

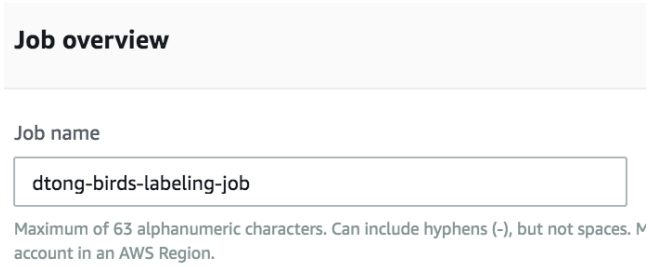
27. Select **Labeling jobs** from the navigation panel on the left.



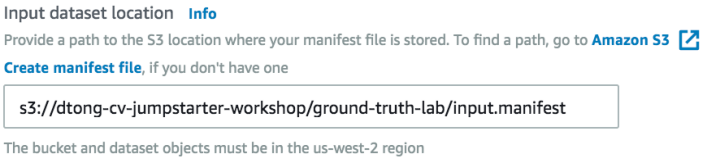
Click on the **"Create labeling job"** button.



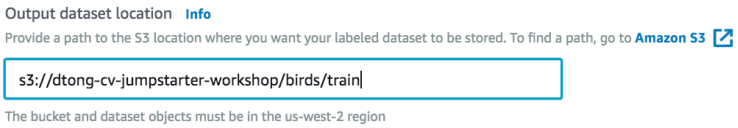
28. Complete the labeling job wizard. First, provide a **unique name** for your labeling job—one that you can remember.



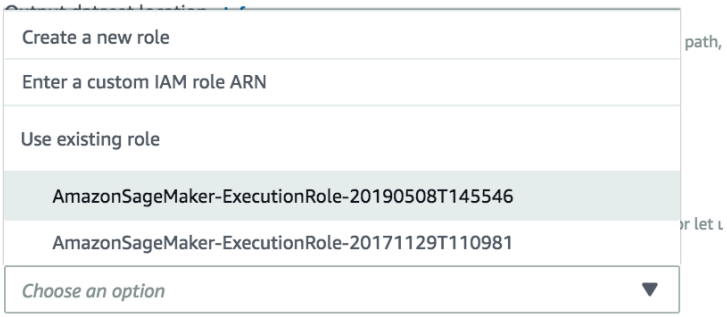
Provide the **full S3 URI** to the **input.manifest** file in your S3 bucket.



Specify an output path in your S3 bucket like the one below. Take note of this path as this is where the annotations created through SageMaker Ground Truth will be outputted.

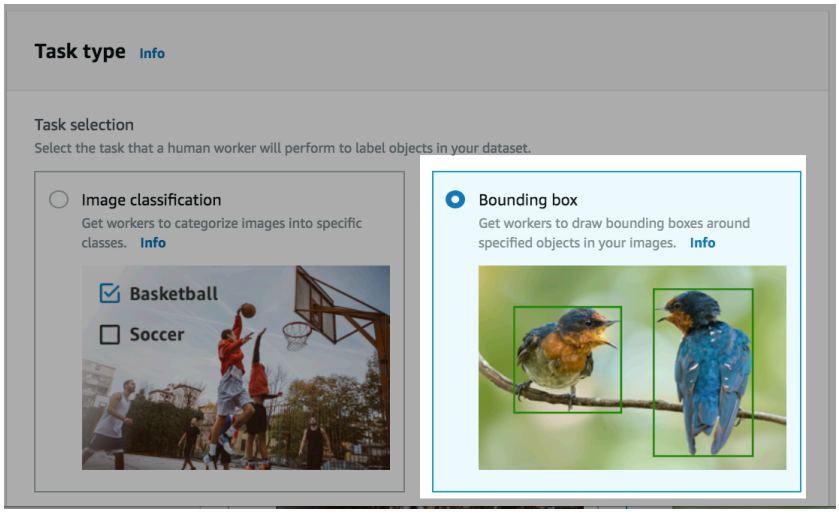


Select the SageMaker execution role that you created when you launched your notebook. The role will appear under **"Use existing role."**

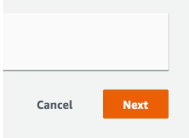


This role provides sufficient permissions for SageMaker Ground Truth. You should practice the least privileges principles outside of this lab. The role that you've created is overly permissive.

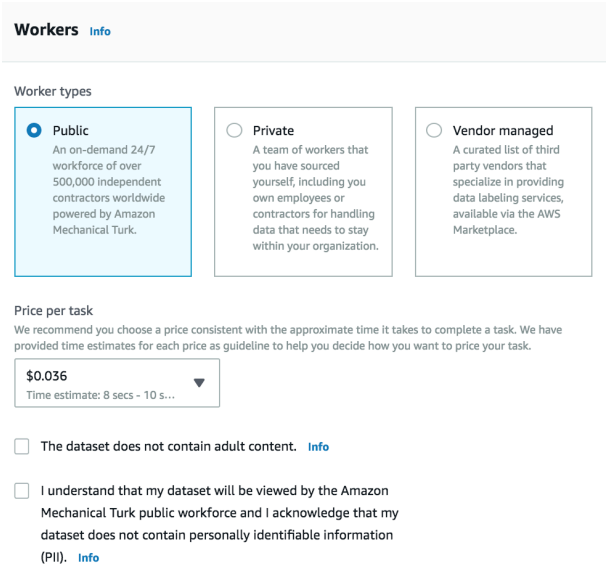
29. Under **Task type**, select **Bounding box**.



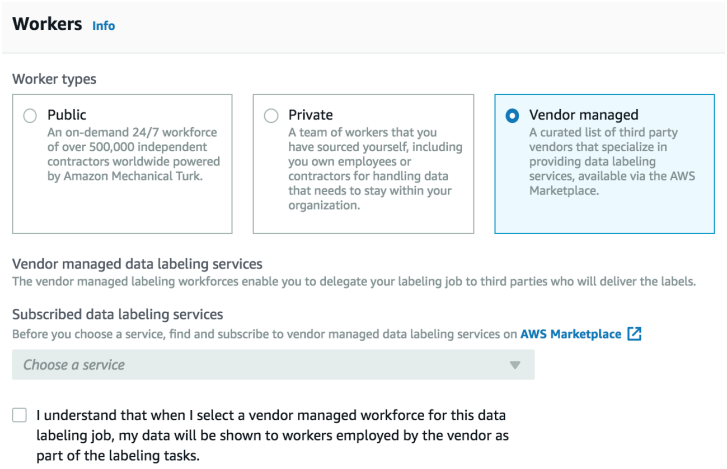
Click on **Next**.



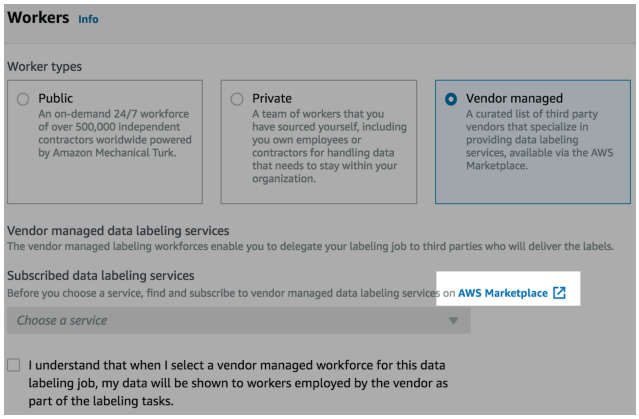
30. Next, we'll assign this job to a workforce. There are a few options. Selecting **public** would assign the job to a Mechanical Turk workforce.



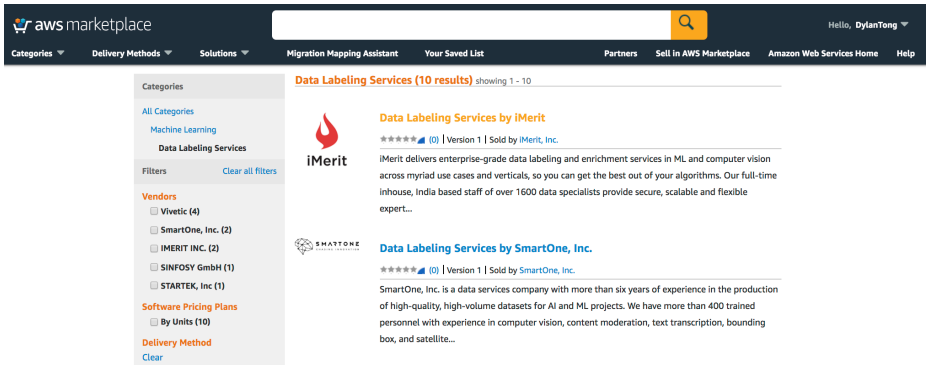
Selecting **Vendor managed** will allow you to assign the job to a 3rd-party vendor from the AWS Marketplace, which you have subscribed to. Billing needs to be setup in order for you to subscribe to a vendor.



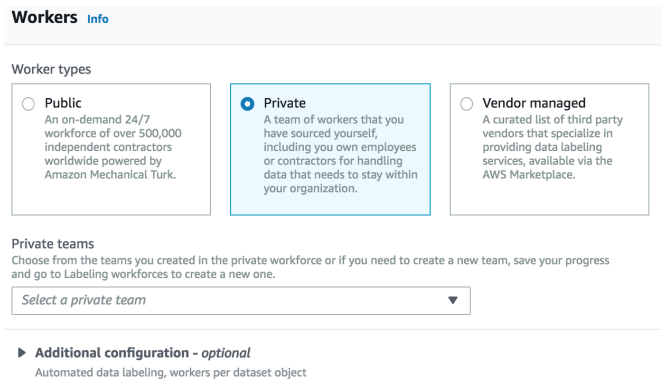
You can navigate to the **AWS Marketplace** from the **shortcut** displayed below:



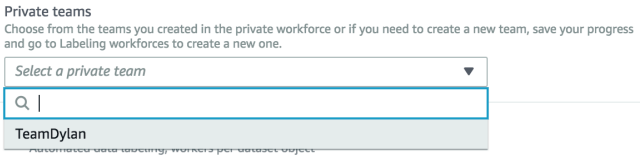
You can browse and subscribe to vendors from the AWS Marketplace as shown below.



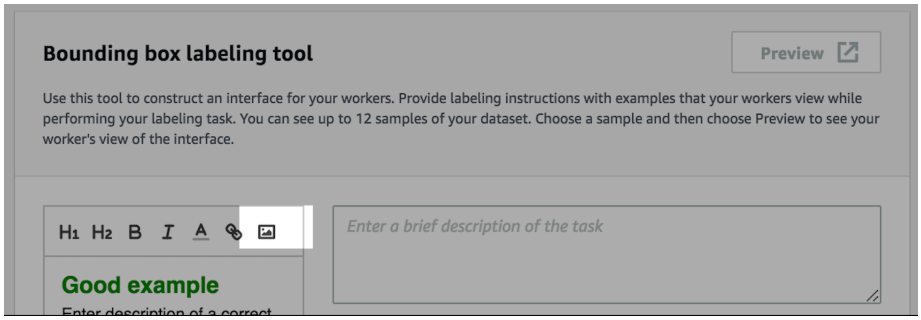
31. Select **Private** under **Worker types**.



32. Select the **private workforce** that you created earlier from the **Private teams** drop down.



33. We need to provide the labeling workforce with clear instructions. Providing a good and bad example is a good practice. **Select the icon** as shown below to link an image.

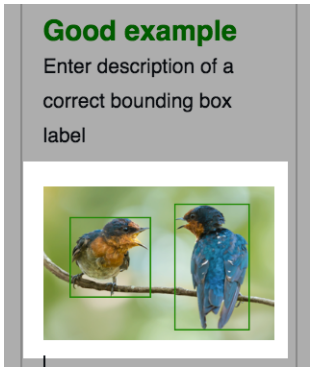


You can use the following image that is shared publicly from S3 through CloudFront:

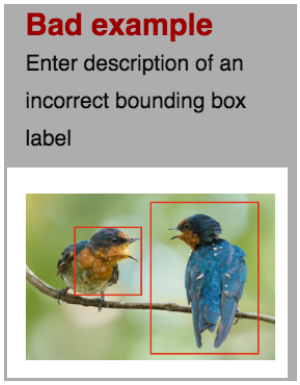
https://dvt7olt8euncl.cloudfront.net/41473cc4-ca5a-442f-9db9-cf116e59957f/src/images/bounding-box-good-example.png

Copy the link into dialog box and select **Save**.

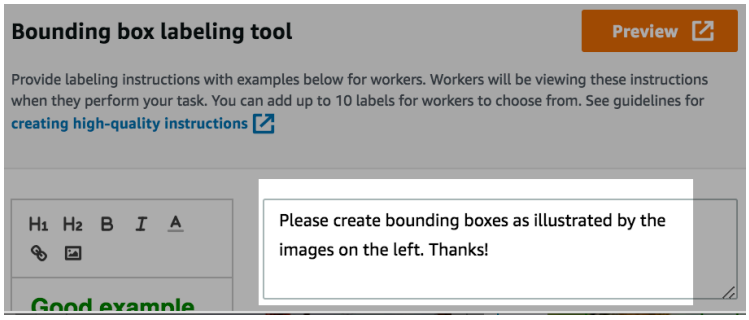
You should see the following image. Delete the placeholder, and leave this new image in its place.



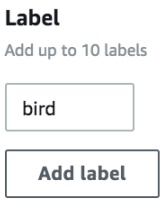
Do the same with the following link to provide a "bad example:"
https://dvt7olt8euncl.cloudfront.net/41473cc4-ca5a-442f-9db9-cf116e59957f/src/images/bounding-box-bad-example.png



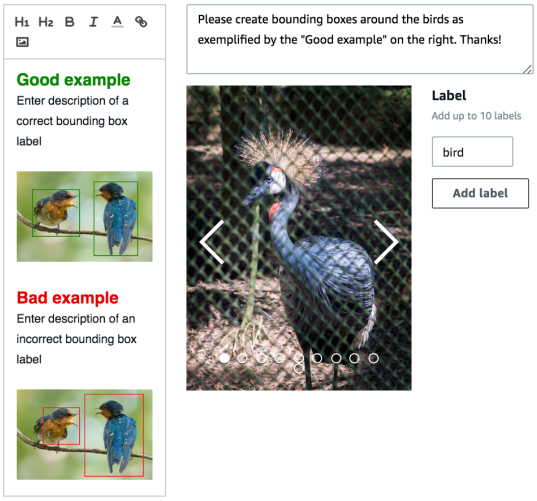
34. Provide instructions in the text box as shown below:



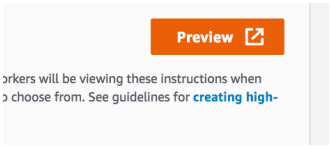
35. Enter **"bird"** into the **Label** textbox. In this lab we create a simple bird detector that doesn't differentiate between different types of birds. In practice, you can add multiple labels for each object class you wish to classify.



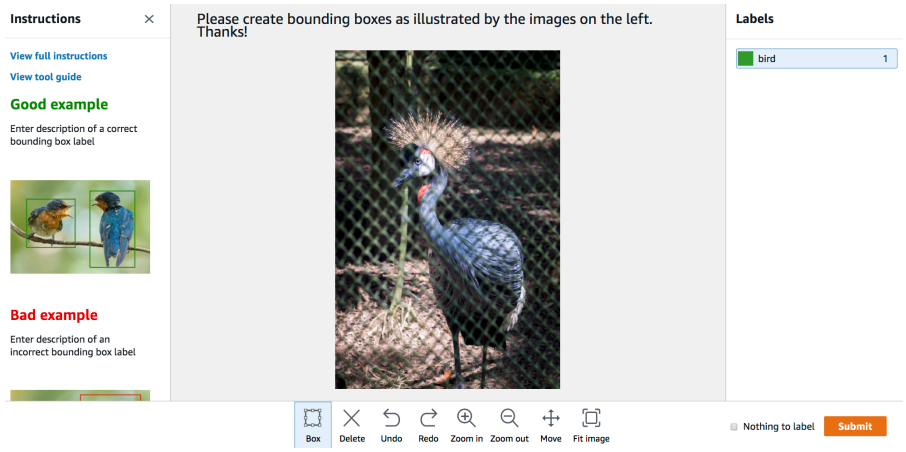
36. The completed template should appear like the following:



You can select the Preview button to get a glimpse of the interface that the labelers will see.

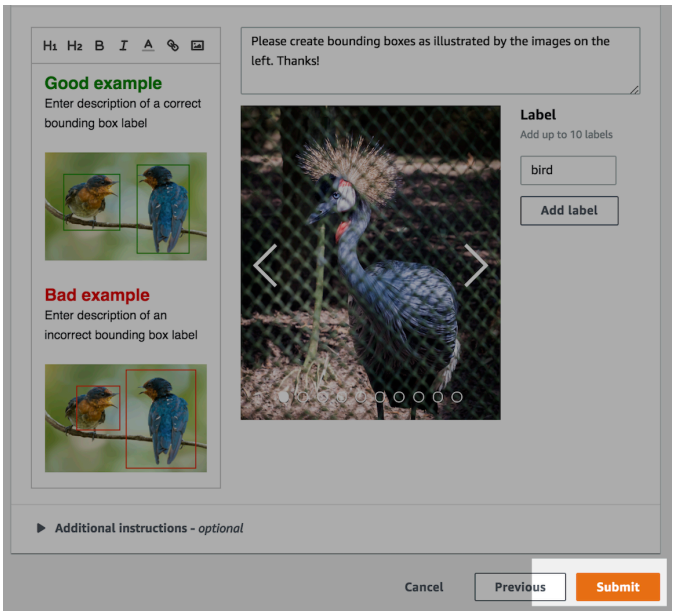


You should see something like this:



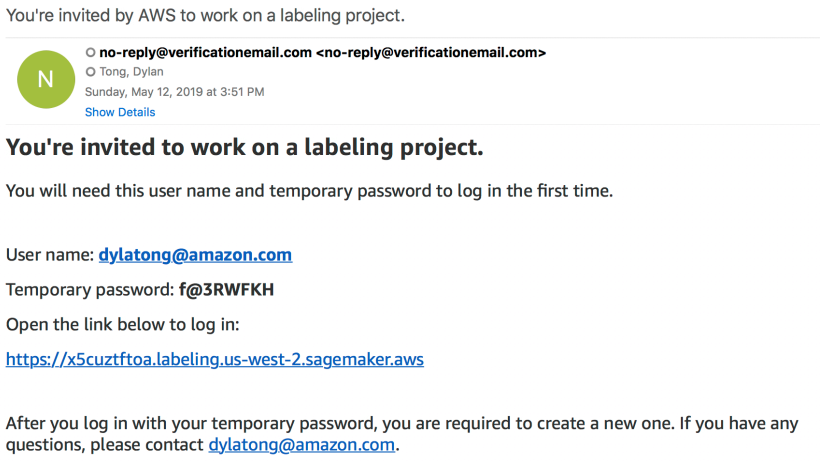
Close the window.

37. Click on the **Submit** button to complete the **Labeling job** description.

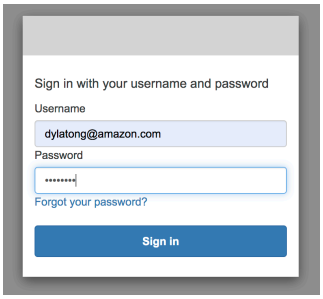


IV. Annotate your Data

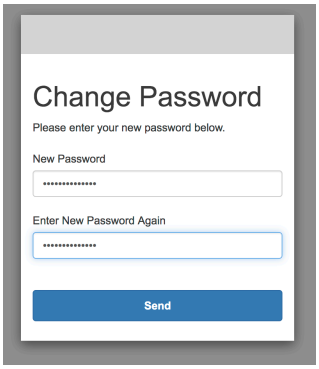
38. Login to the email account, which you had an invitation sent out to earlier. The invitation should appear similar to the one below:



Follow the instructions, and click on the link to login. Login with the temporary credentials that have been provided.

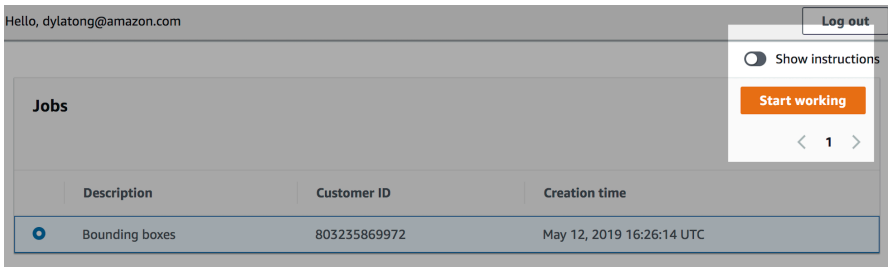


Change your password and remember it.



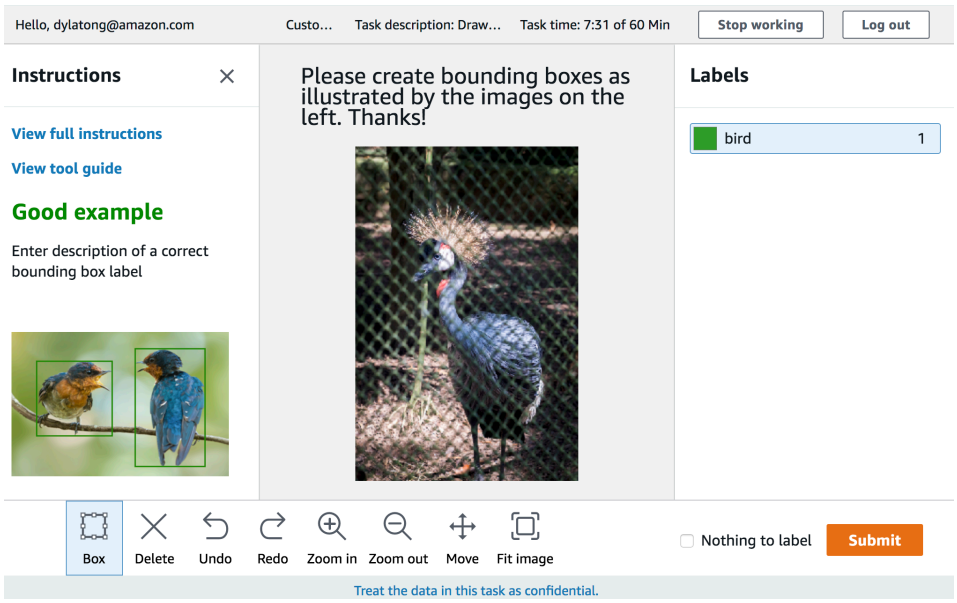
39. You should see the labeling job that you created earlier once you've logged into the portal. If not, try waiting a moment and refreshing your browser.

    Select the job, and click **Start working.**

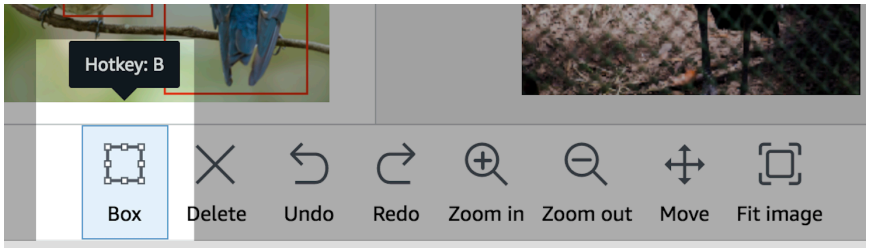


40. This labeling job consists of 10 images. You will need to annotate them, so that we can use them in the next lab.

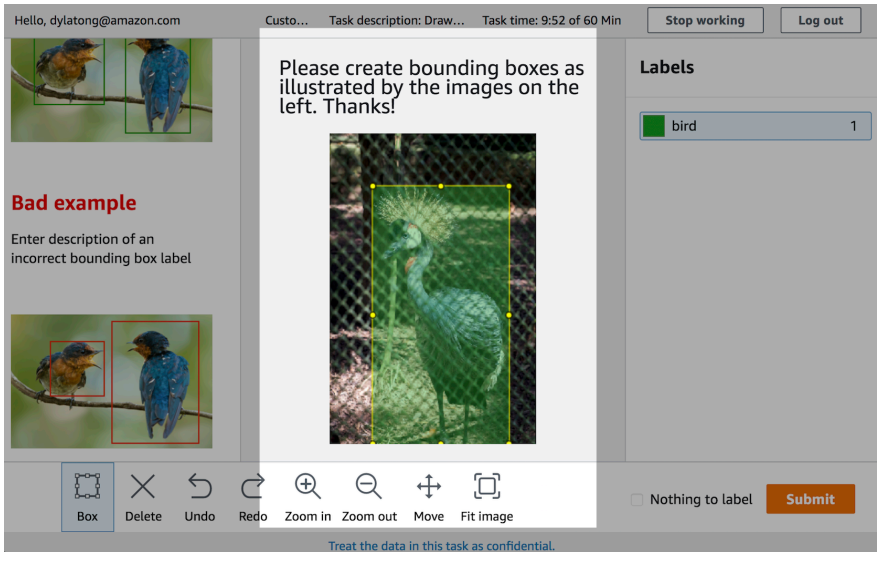
    Create bounding boxes around each of the birds in the images.



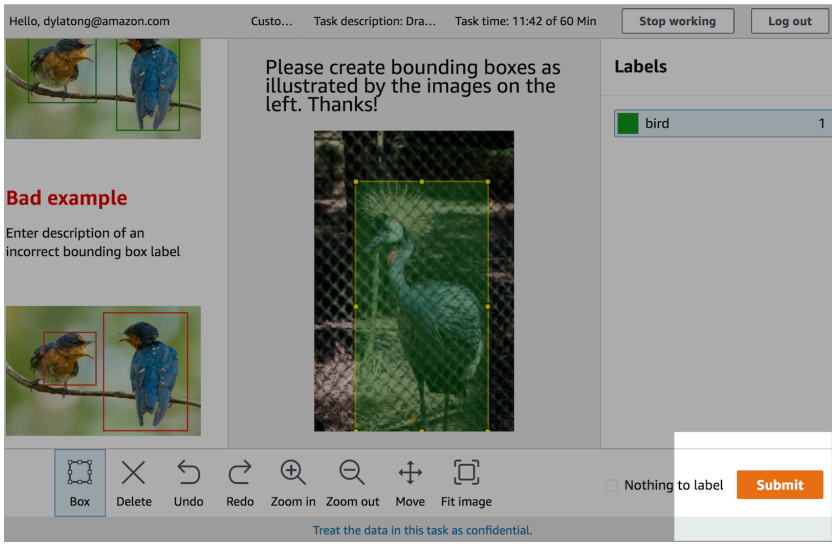
Select the box icon in the toolbar.



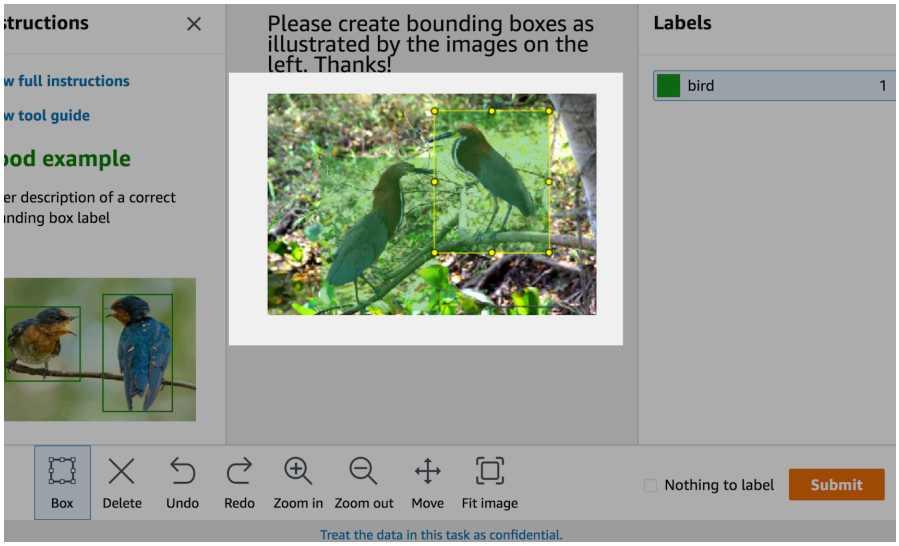
Click, drag and release to create the boxes.



Click on the **Submit** button once you're done with creating the bounding boxes.

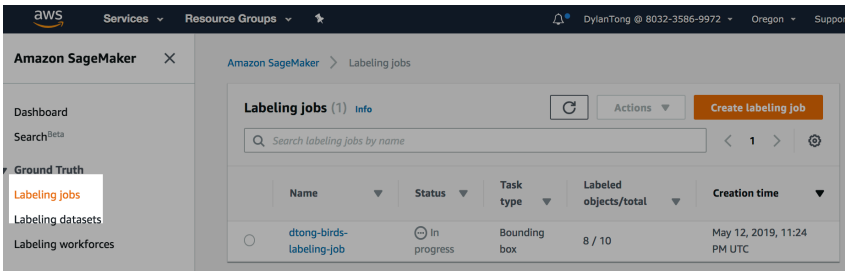


Some of the images have more than one bird. Create boxes around all the birds like in the example below.

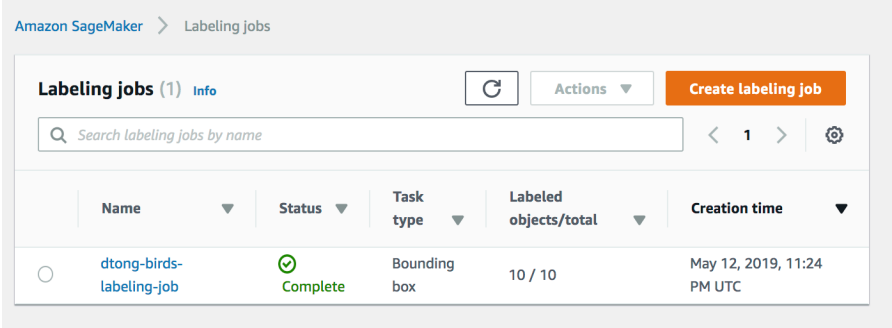


41. Once you've annotated the 10 images, you'll be logged out of the annotation interface. You can view the status of your labeling job. Expect a short delay.

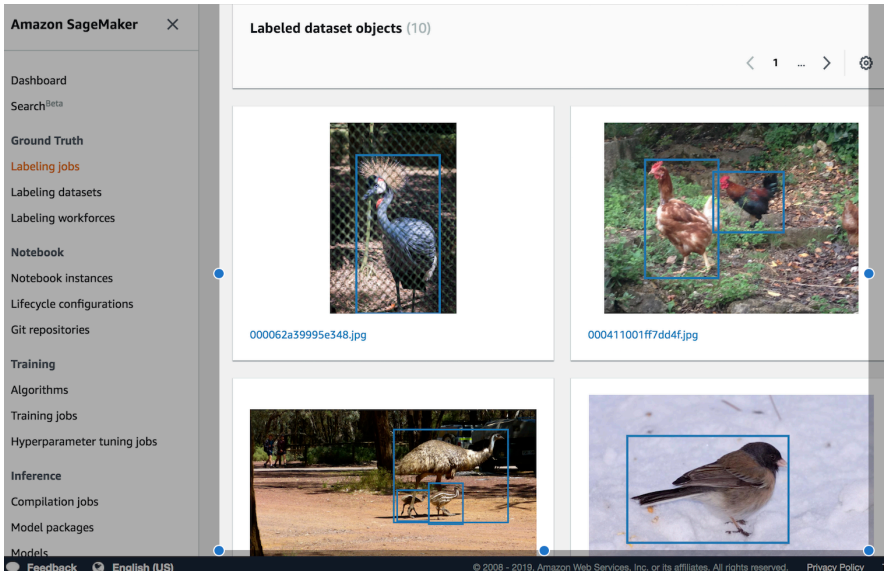
    For manual (human) labeling jobs, SageMaker Ground Truth processes your annotations via Annotation Consolidation to improve the accuracy of labels. In the lab, you're the only worker in your private workforce, so Annotation Consolidation doesn't have any effect.



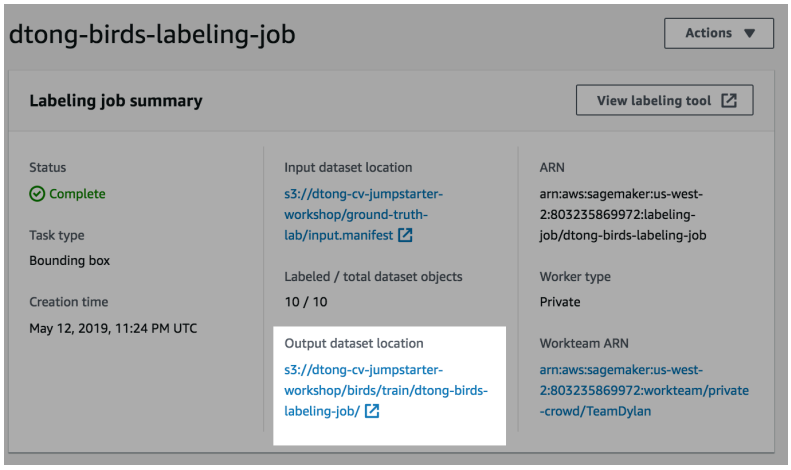
42. Once your labeling job shows up as **Complete**, click on the labeling job.



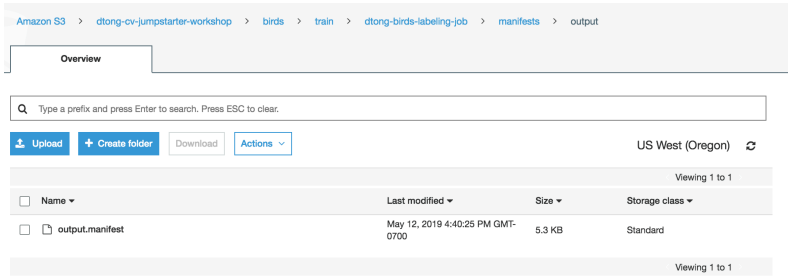
Scroll down. Within the details of the job, you can preview your annotations.



43. Click on the link **“Output dataset location.”**



This will take you to the location of the **output.manifest** file that was created by your labeling job. It contains meta-data for your image annotations.



The contents of the **output.manifest** should look something like this. Take note of the location of this file. In the next lab we'll run a training job, which can take this file as direct input without further modifications.


```
{
    "source-ref": "s3://dtong-cv-jumpstarter-workshop/ground-truth-lab/images/0027f99a032cca4a.jpg",
    "dtong-birds-labeling-job": {
        "annotations": [
            {
                "class_id": 0,
                "width": 578,
                "top": 147,
                "height": 382,
                "left": 139
            }
        ],
        "image_size": [
            {
                "width": 1024,
                "depth": 3,
                "height": 683
            }
        ]
    },
```