
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

AIM209

Create & label high-quality datasets for training machine learning models

Aashmeet Kalra
Principal Solutions Architect
AWS

Geremy Cohen
Senior Solutions Architect
AWS

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

- Challenges with data labeling
- How these challenges are addressed with SageMaker Ground Truth
- Hands-on demos
  - Setting up a built-in workflow
  - Setting up a custom workflow
  - Setting up synthetic data
- How to choose between Ground Truth and Ground Truth Plus when implementing built-in and custom workflows
  - Do it yourself with Ground Truth
  - Turnkey solution with Ground Truth Plus
- Open discussion

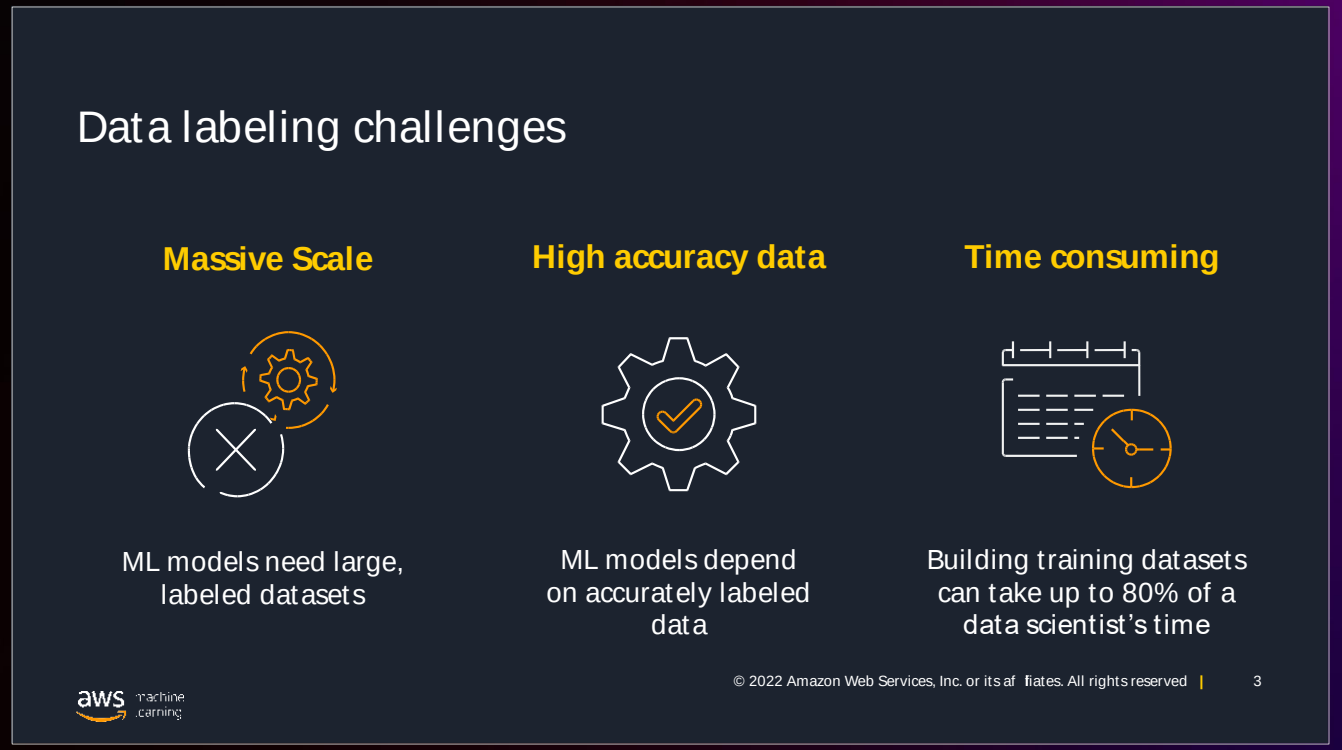
aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Challenges with data labeling

The most common challenges customers face around data labeling revolve around the following themes:

- Collecting data
- System logistics
- Annotator logistics
- Scheduling logistics

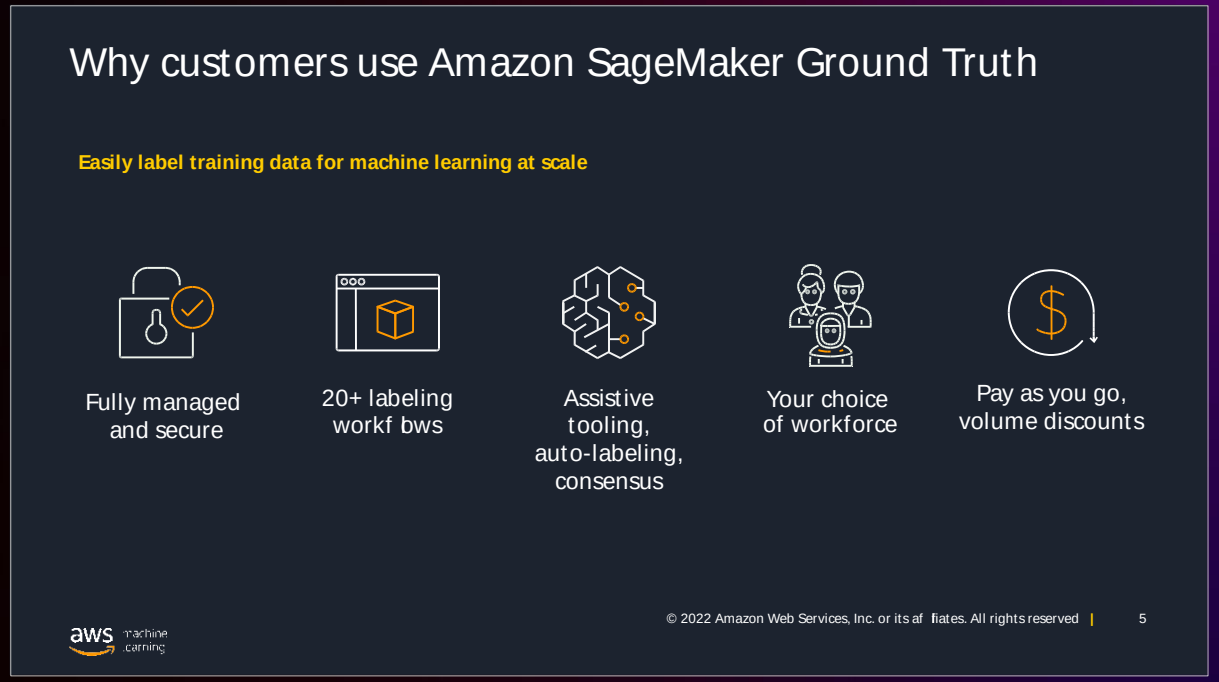


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Addressing labeling challenges with Ground Truth and Ground Truth Plus

To help customers overcome these challenges, AWS offers:

- Amazon SageMaker Ground Truth
- Amazon SageMaker Ground Truth Plus
- Amazon Ground Truth Synthetic Data Generation service
- Support for high volume labeling workloads, multiple modalities, and endless customization
- Annotator workforce options include your own team, Mturk, or SageMaker marketplace partners



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Security and governance

- AWS Identity and Access Management (IAM) managed and custom policies
- CORS configuration to the Amazon S3 buckets that contain your input data
- Run labeling jobs in VPC or use Amazon VPC mode from a private worker portal
- You can label highly sensitive data, stay in control of your data, and employ security best practices
  - Use a customer managed key to encrypt your output data
  - Use AWS KMS customer managed keys with your automated data labeling job to encrypt the storage volume attached to the compute instance used for model training and inference
- Workforce authentication and restrictions
- Observability and monitoring of labeling jobs

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

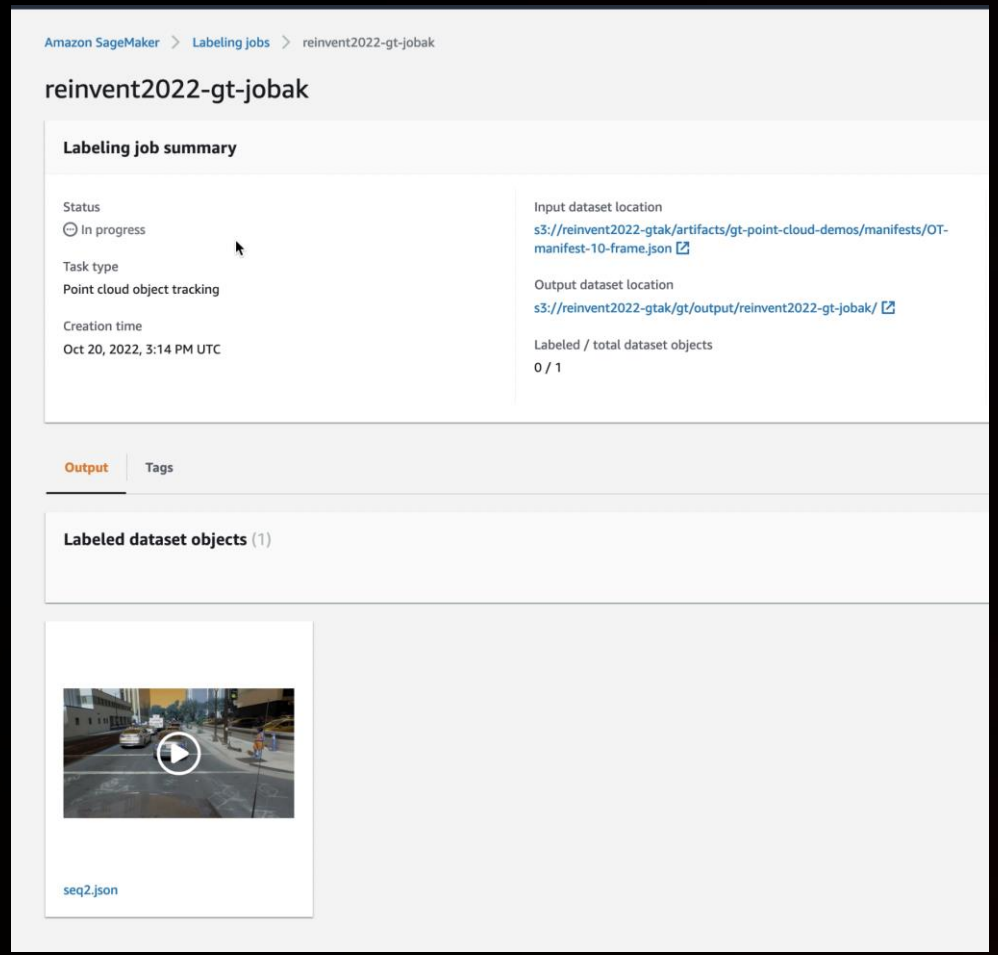
Hands-on demos

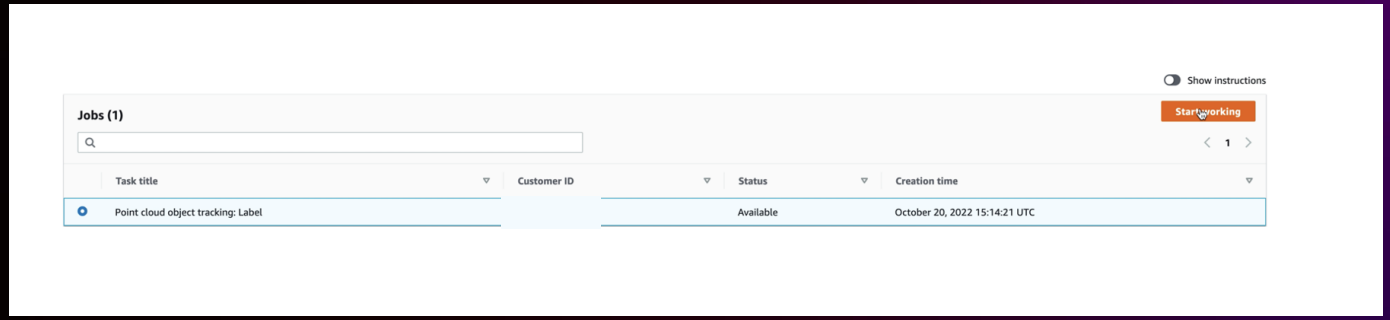
To provide a better idea of how easy it is to get started labeling with Ground Truth, we'll now demo the setup of:

- A built-in labeling workflow
- A customized labeling workflow
- Synthetic labeling data

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

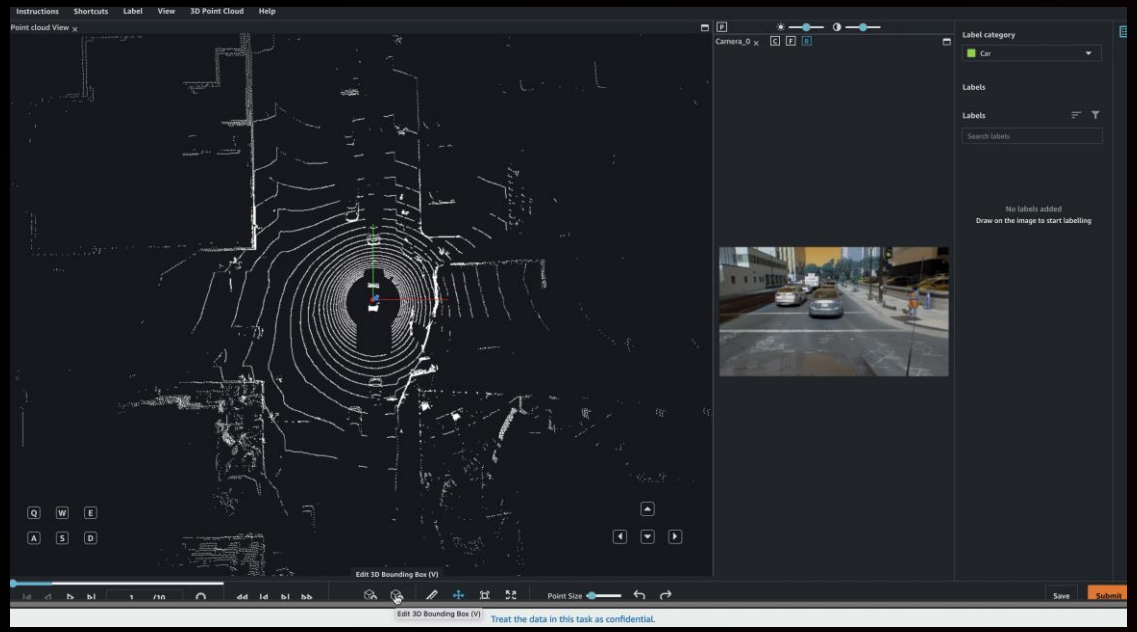
Built-in workflow walkthrough: Point cloud use case

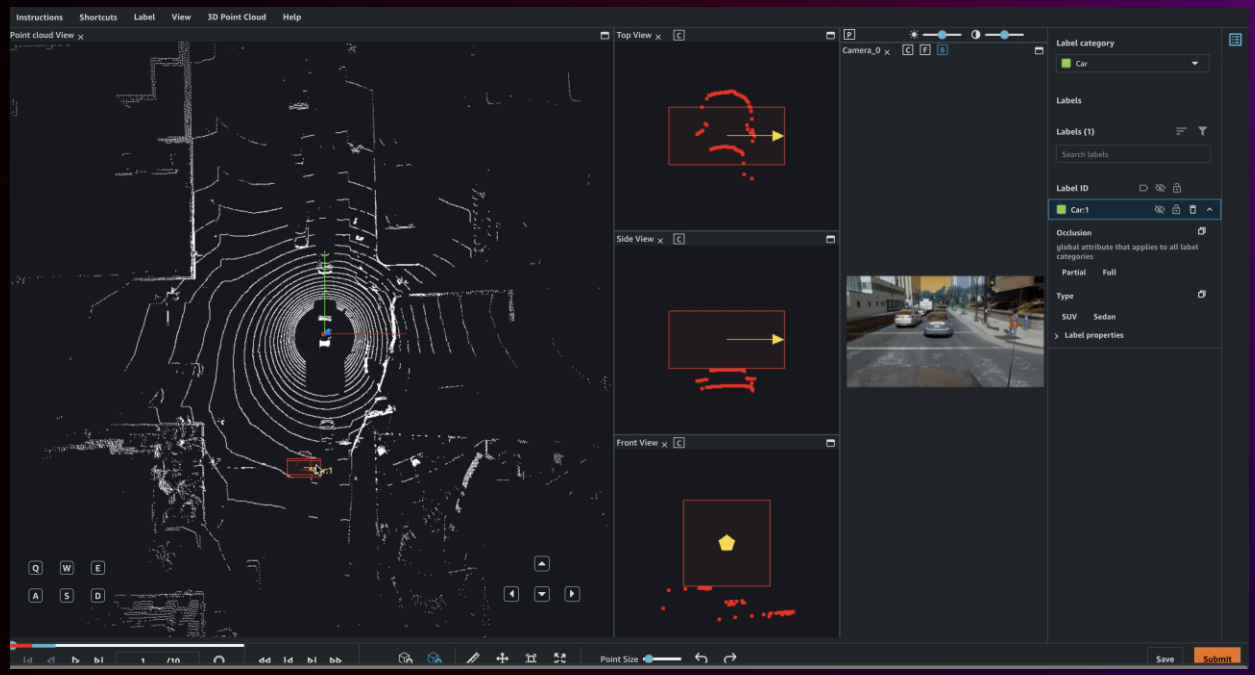




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

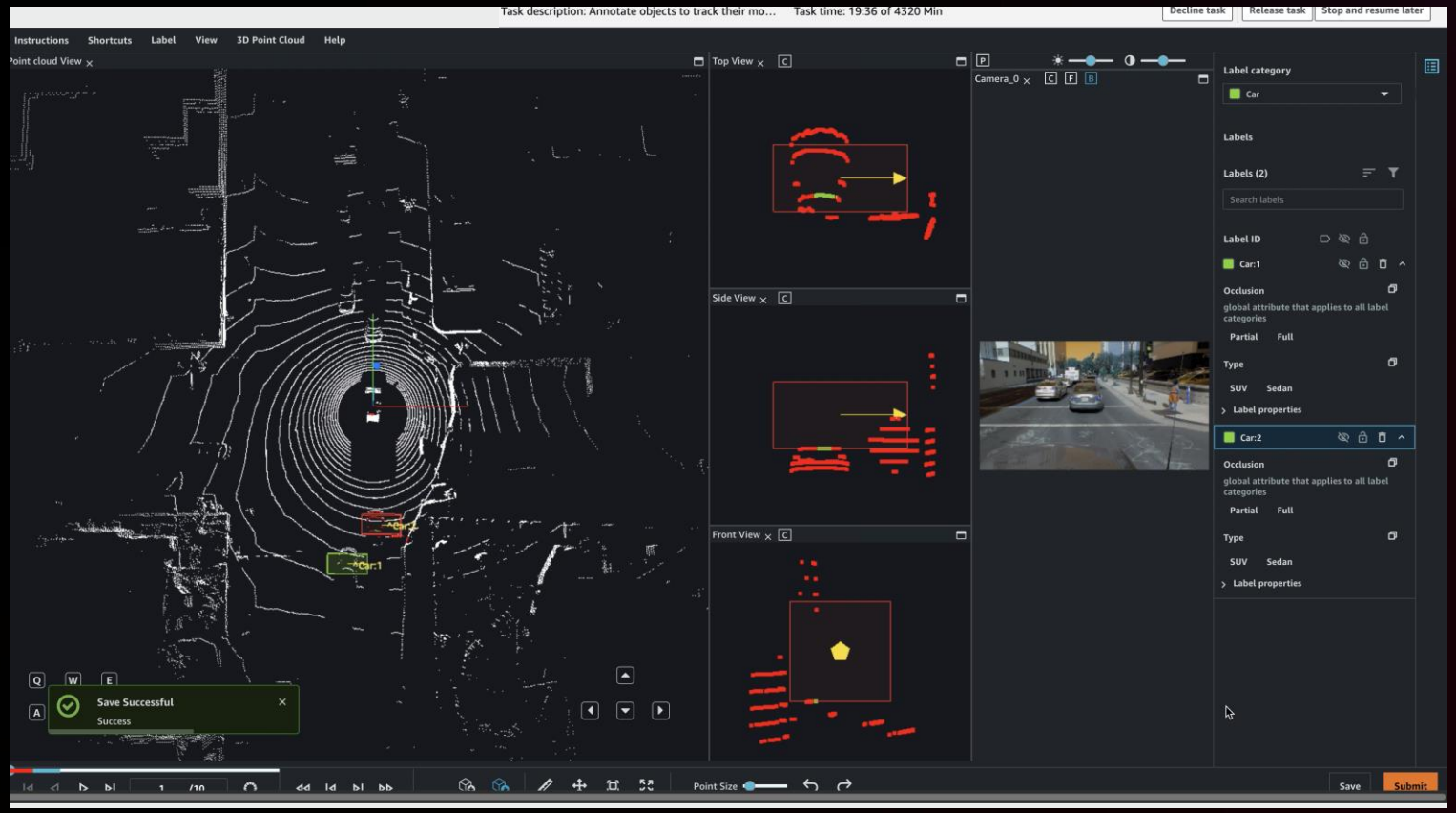
Point cloud – Labeling

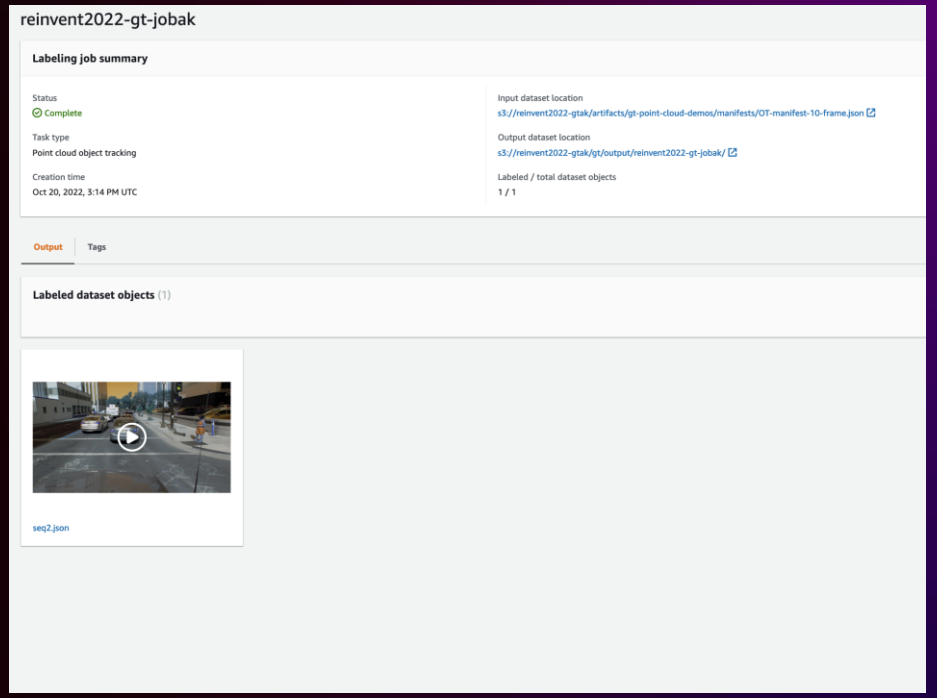




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Point cloud – Labeling

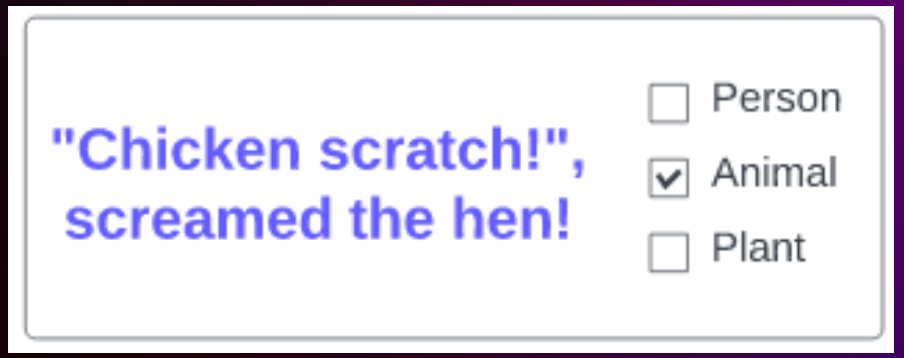




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Customized workflow: Text classification use case

If the built-in workflow types are not sufficient for your use case, Ground Truth makes it easy for you to create an entirely new workflow or customize a built-in one.

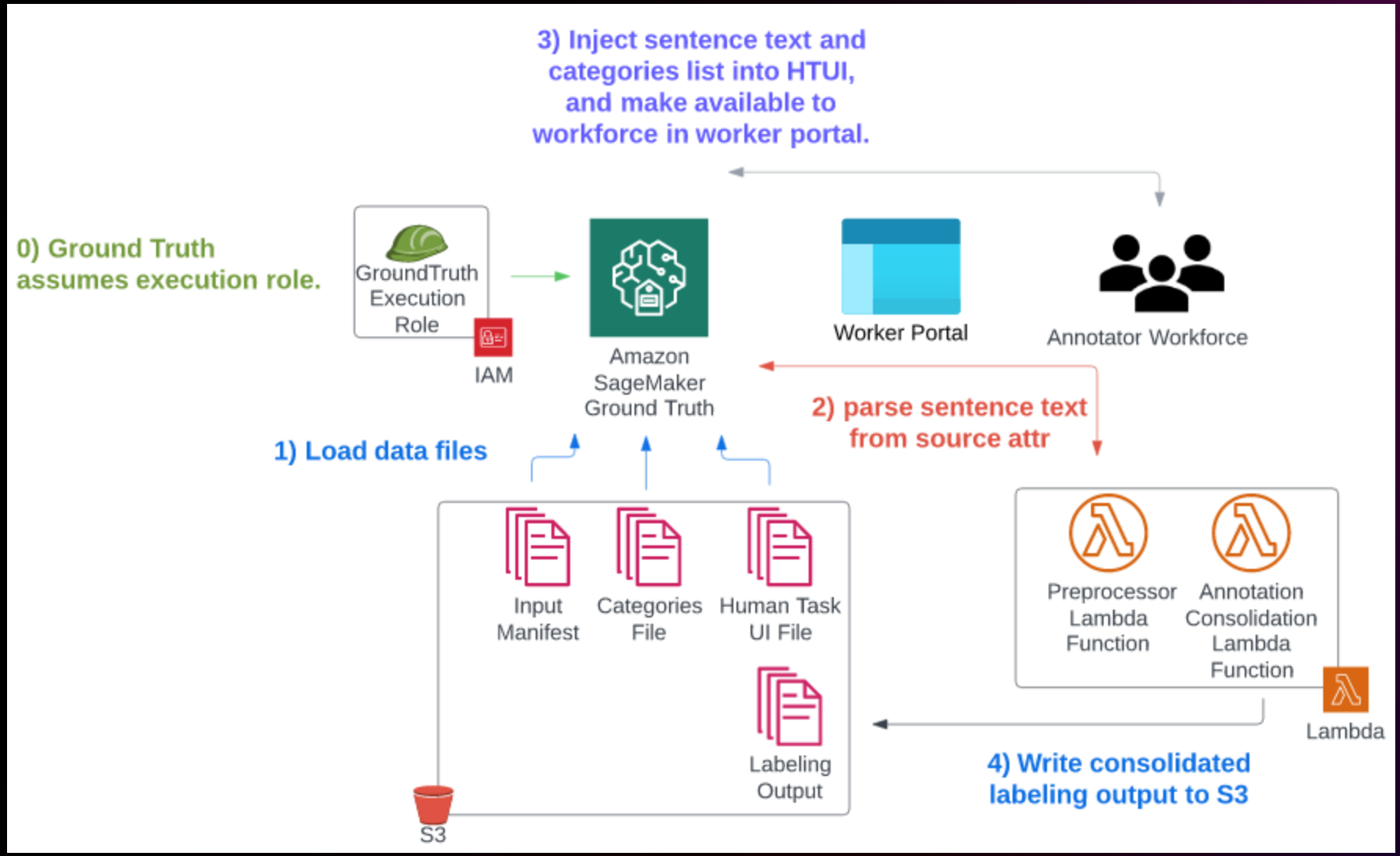


We'll demonstrate customizing a built-in text classification workflow. The use case surrounding our customization:

- Annotators have been labeling data using the built-in Ground Truth text classification workflow.
- The labeling UI that is provided with the built-in workflow works fine, but based on annotator feedback, we'd like to make the following modifications:
  - Increase submit button size
  - Enable keyboard hotkeys: Enter to Submit, 'r' to reset selections

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Customized Workflow Architecture



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

```
{"source":"His nose could detect over 1 trillion odors!"}
{"source":"Why do fish live in salt water? Because pepper makes them sneeze!"}
{"source":"What did the buffalo say to his son when he went away on a trip? Bison!"}
{"source":"Why do plants go to therapy? To get to the roots of their problems!"}
{"source":"What do you call a nervous tree? A sweaty palm!"}
{"source":"Some kids in my family really like birthday cakes and stars!"}
{"source":"A small portion of the human population carries a fabella bone."}
```

Input Manifest

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Labeling Categories File

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

```
<script src="https://assets.crowd.aws/crowd-html-elements.js">
</script>

<crowd-form>
  <crowd-classifier
    name="crowd-classifier"
    categories="{{ task.input.labels | to_json | escape }}"
    header="Please classify"
  >

    <classification-target>
      <strong>{{ task.input.taskObject }}</strong>
    </classification-target>
    <full-instructions header="Full Instructions">
      <div>
        <p>Based on the general subject or topic of each
          sentence presented, please classify it as
          person, animal, or plant. </p>
      </div>
    </full-instructions>

    <short-instructions>
      Complete tasks
    </short-instructions>
  </crowd-classifier>
</crowd-form>
```

Human Task UI

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

```
<script src="https://assets.crowd.aws/crowd-html-elements.js">
</script>

<crowd-form>
  <crowd-classifier
    name="crowd-classi
    categories="{{ tas
    header="Please cla
  >

    <classification-ta
      <strong>{{ task.
    </classification-t
    <full-instructions
      <div>
        <p>Based on th
          sentence pre
          person, anim
      </div>
    </full-instruction

    <short-instruction
      Complete tasks
    </short-instructio
  </crowd-classifier>
</crowd-form>
```

```
<script>
  document.addEventListener('all-crowd-elements-ready', () => {
    // Creating new button to inject in label pane
    const button = document.createElement('button');
    button.textContent = 'Submit';
    button.classList.add('awsui-button', 'awsui-button-variant-primary',
                         'awsui-hover-child-icons');

    // Editing styling to make it larger
    button.style.height = '60px';
    button.style.width = '100px';
    button.style.margin = '15px';

    // Adding onclick for submission
    const crowdForm = document.querySelector('crowd-form');
    button.onclick = () => crowdForm.submit();

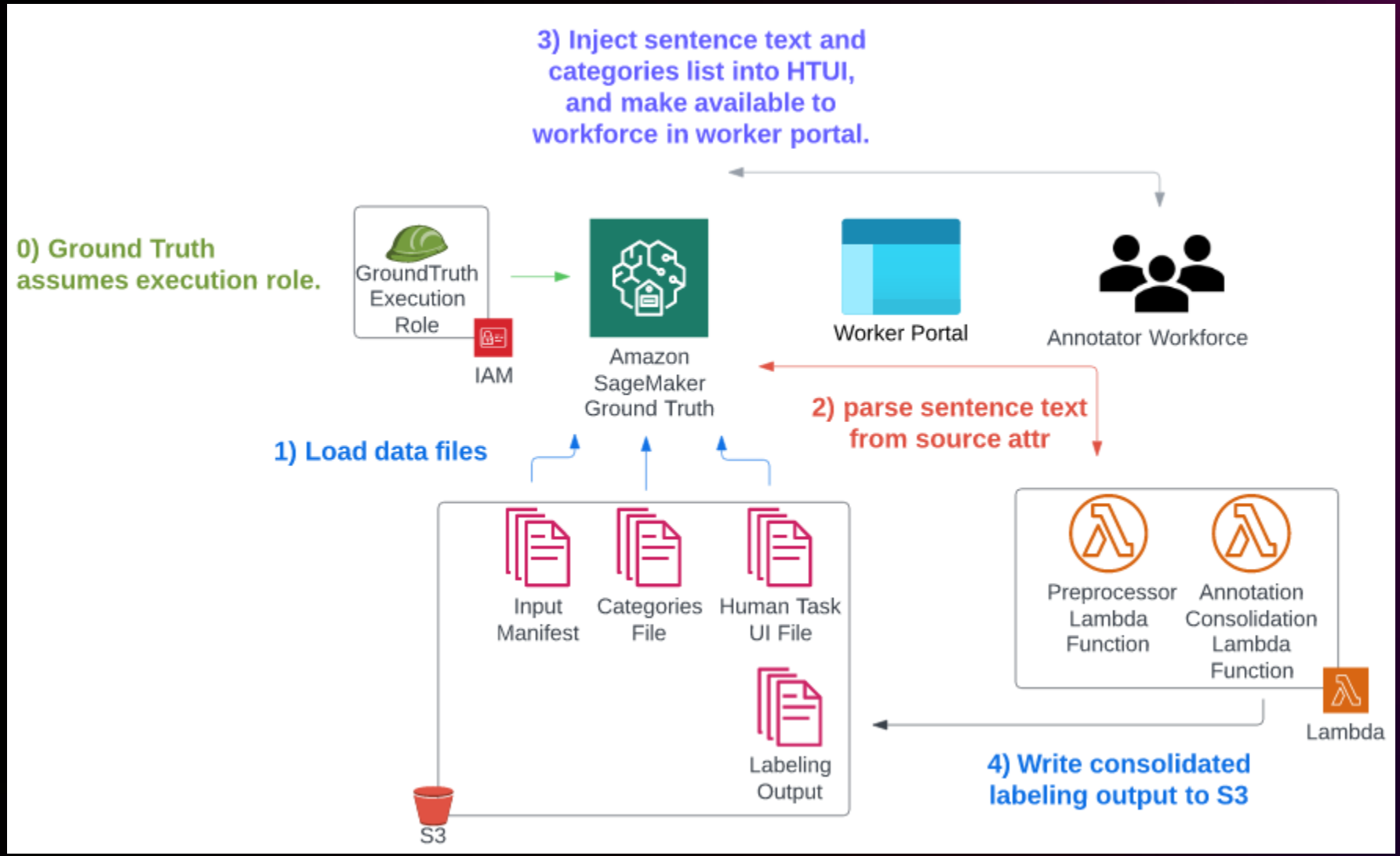
    // Injecting
    const crowdClassifier = document.querySelector('crowd-classifier').shadowRoot;
    const labelPane = crowdClassifier.querySelector('.category-picker-wrapper');
    labelPane.appendChild(button);

    // Adding a Enter hotkey
    document.addEventListener('keydown', e => {
      if (e.key === 'Enter') {
        crowdForm.submit();
      }
      if (e.key === 'r') {
        crowdForm.reset();
      }})});
</script>
```

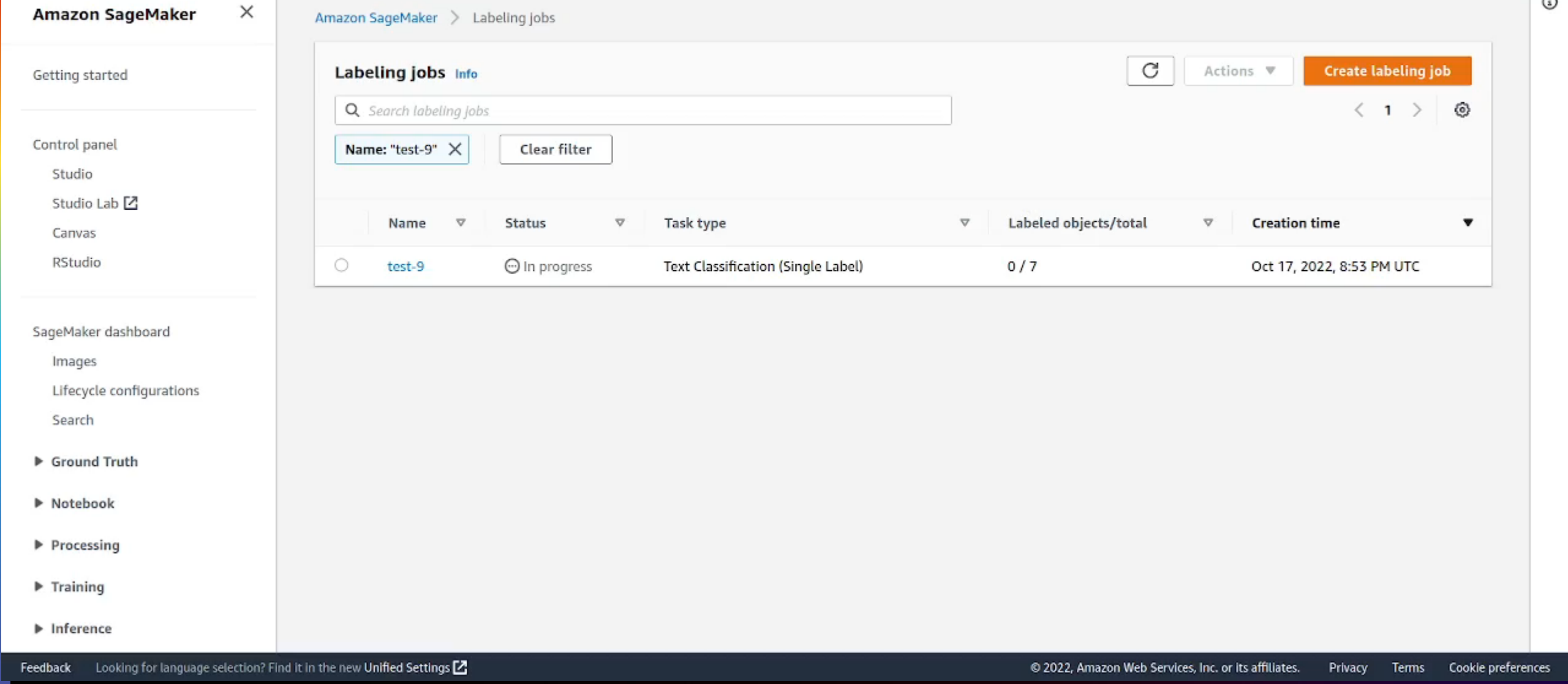
Human Task UI –Tweaks for button and hotkeys

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Customized Text Classification Use Case



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Job Labeling and Review

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Choosing between Ground Truth and Ground Truth Plus

WHEN IMPLEMENTING BUILT-IN AND CUSTOM WORKFLOWS

Now that you've seen some examples of built-in and customized labeling workflows with Ground Truth, we can discuss factors to consider when choosing between

- Creating the workflow on your own with Amazon SageMaker Ground Truth
- Having the workflow delivered as a turnkey solution with Ground Truth Plus

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Running the Workflow on your own

When considering running a labeling workflow on your own, ask yourself these questions:

1. Will you be implementing a custom workflow?

2. How much data pre or post-processing does your workflow require?

3. Are you able to actively hire, train, and support your annotation workforce to keep the workflow moving and make your SLAs?

For lower-volume, built-in workflow types that require minimal data pre and post-processing, if you have the capacity to manage your annotation workforce, running the workflow on your own is recommended (and can actually save you time and money.

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Delivery as a turnkey solution (with Ground Truth Plus)



Turnkey solution



Guaranteed throughput



Higher quality and feedback loop



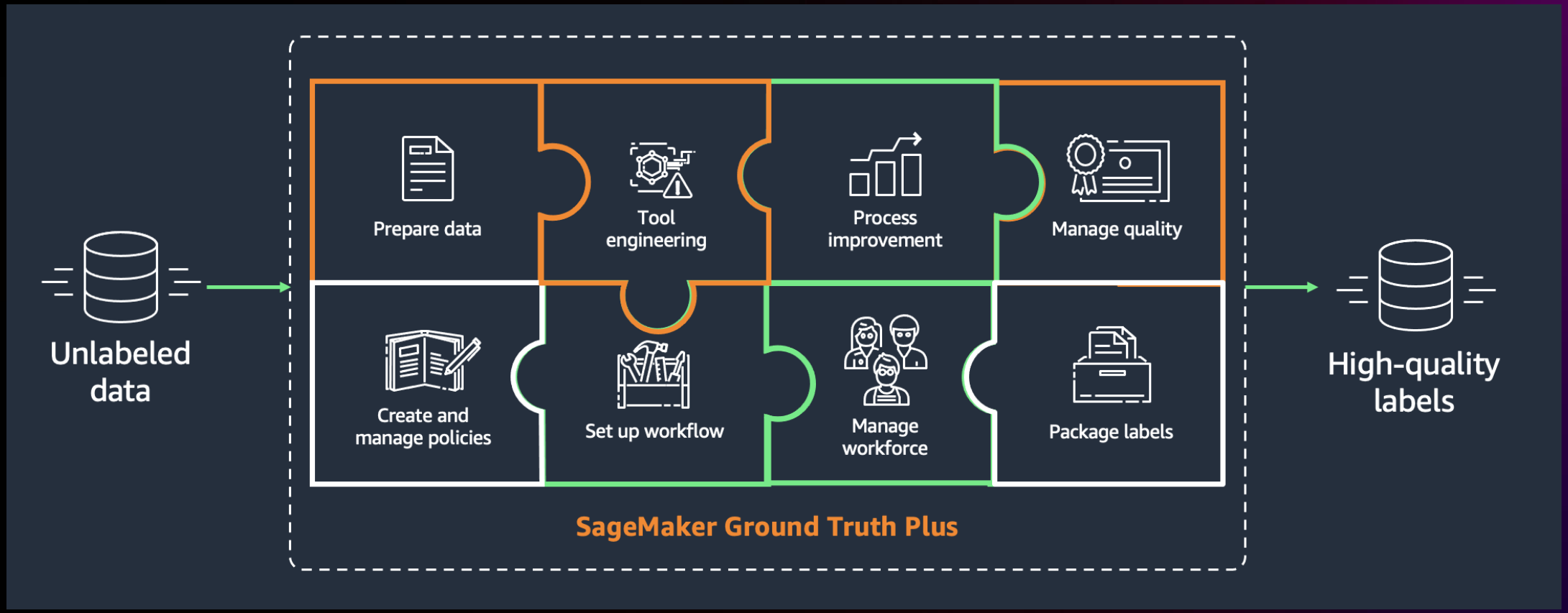
Predictable and lower cost



Tracking & visibility

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Turnkey data labeling service



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Getting started

Ground Truth Plus

Getting started Info

Ground Truth Plus is a turnkey data labeling service that enables you to easily create high-quality training datasets without having to build labeling applications and manage the labeling workforce on your own. Ground Truth Plus sets up data labeling workflows and manages them on your behalf. Learn more



Step 1. Request a pilot

Fill out the project requirement form and our team of AWS experts will schedule a call with you to discuss and set up your data labeling project.

Request a pilot



Step 2. Share data

After the consulation call, share raw data in an Amazon S3 bucket.



Step 3. Create project team

Create a project team after AWS experts set up your data labeling project. The project team helps you ensure that the right users have access to the project portal.

Create project team



Step 4. Access project portal

The project portal allows you and your team to monitor project progress, track metrics, and provide feedback on the output data.

Open project portal

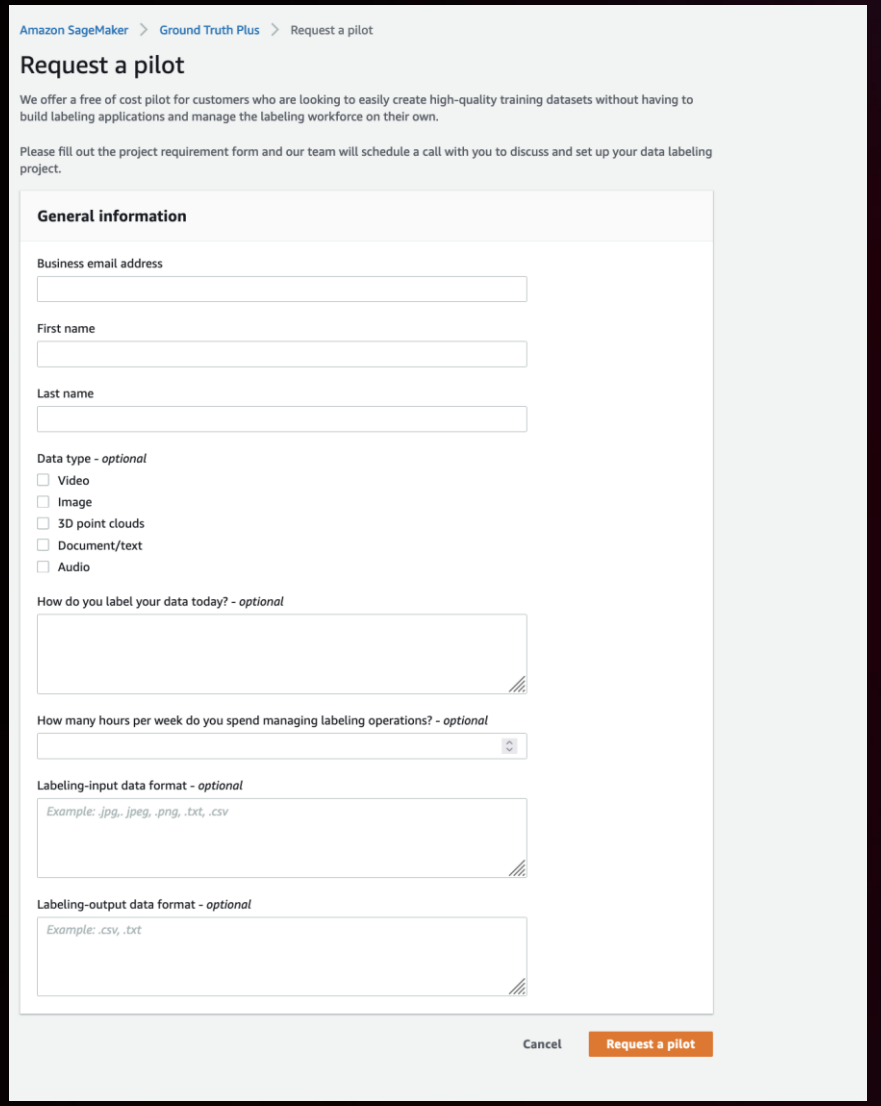


Step 5. Recieve labeled data

Collect high-quality labeled data securely in an Amazon S3 bucket.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Request a pilot



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

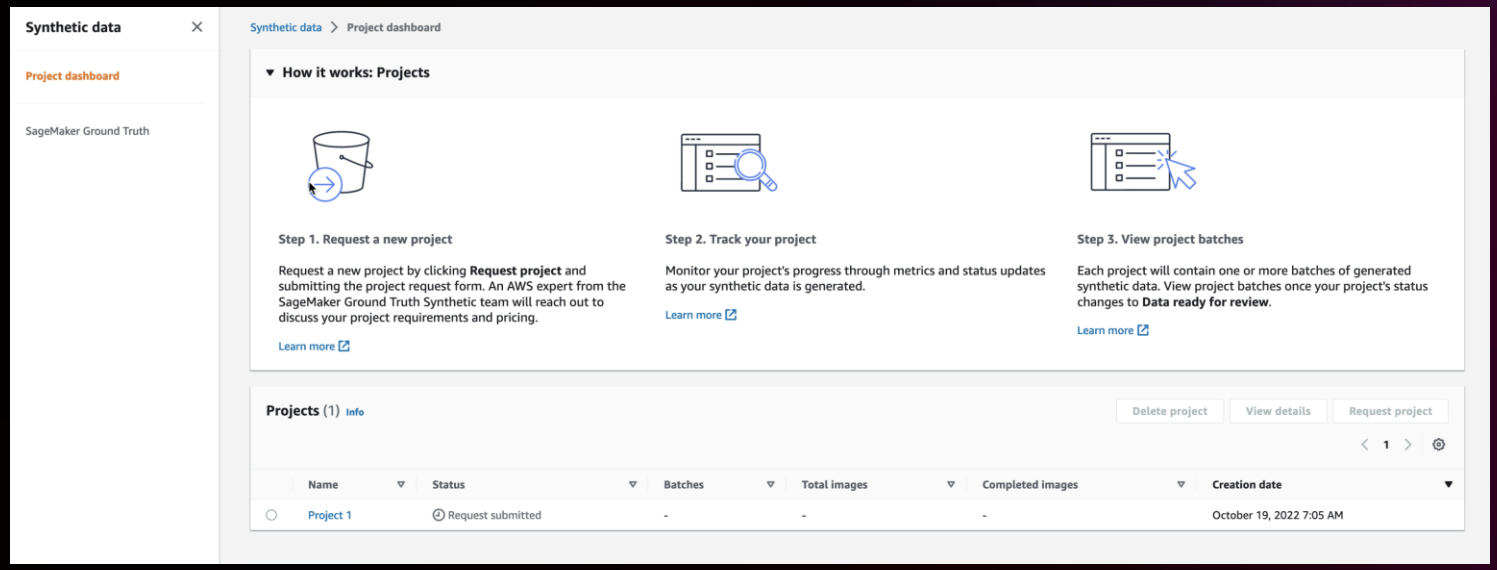
Built-in workflow walkthrough: Synthetic labeling data

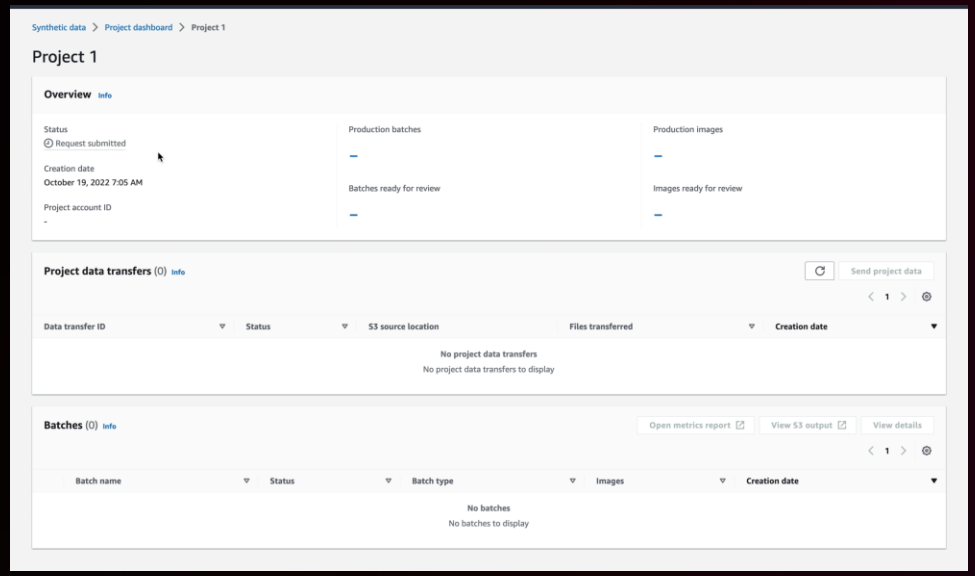
Demo has the following steps:

1. Create project/request a project
2. Track projects
3. View project batches
4. Fidelity and diversity report

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Request a new project





© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

https://us-east-1.console.aws.amazon.com/sagemaker-groundtruth-synthetic/home#project-portal

Services | Search for services, features, blogs, docs, and more [Option+S] | Select a Region | Admin/aashmee-Isengard @ myaccountaash

Resource Groups & Tag Editor | IAM

Synthetic data

- Project dashboard
- SageMaker Ground Truth

Synthetic data > Project dashboard

How it works: Projects



Step 1. Request a new project

Request a new project by clicking **Request project** and submitting the project request form. An AWS expert from the SageMaker Ground Truth Synthetic team will reach out to discuss your project requirements and pricing.

Learn more



Step 2. Track your project

Monitor your project's progress through metrics and status updates as your synthetic data is generated.

Learn more



Step 3. View project batches

Each project will contain one or more batches of generated synthetic data. View project batches once your project's status changes to **Data ready for review**.

Learn more

Projects (1) Info

Delete project | View details | Request project

1

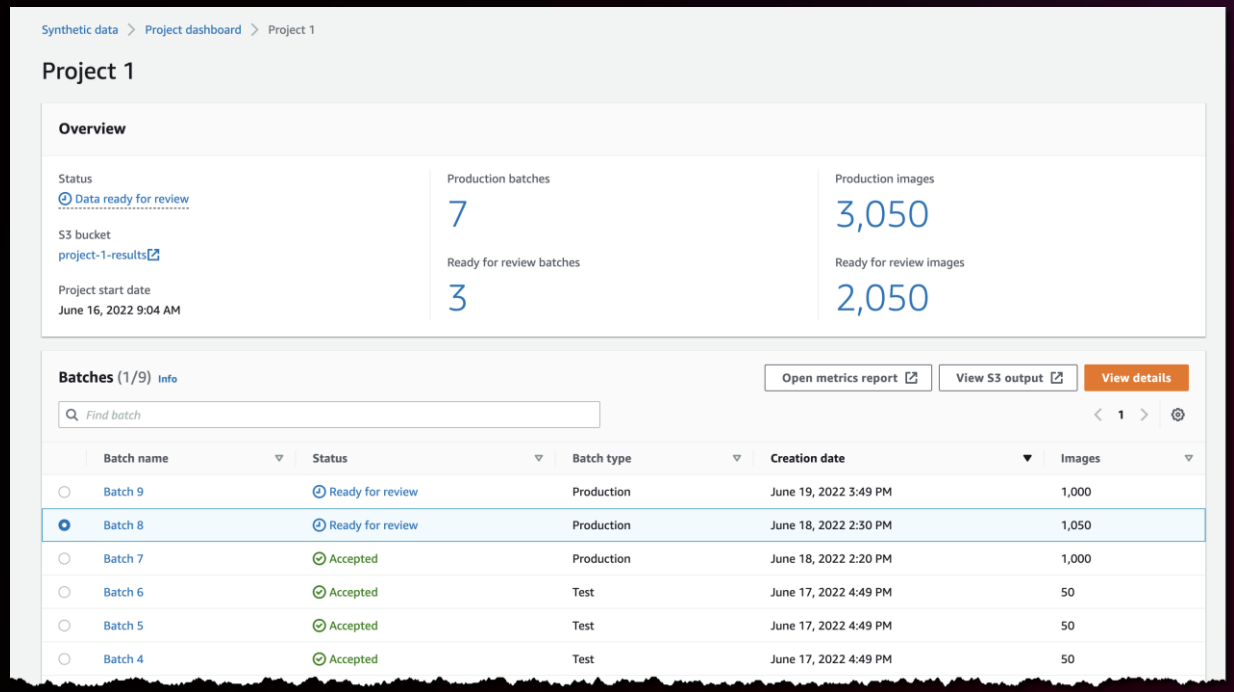
Name	Status	Batches	Total images	Completed images	Creation date
Project 1	Request submitted	-	-	-	October 19, 2022 7:05 AM

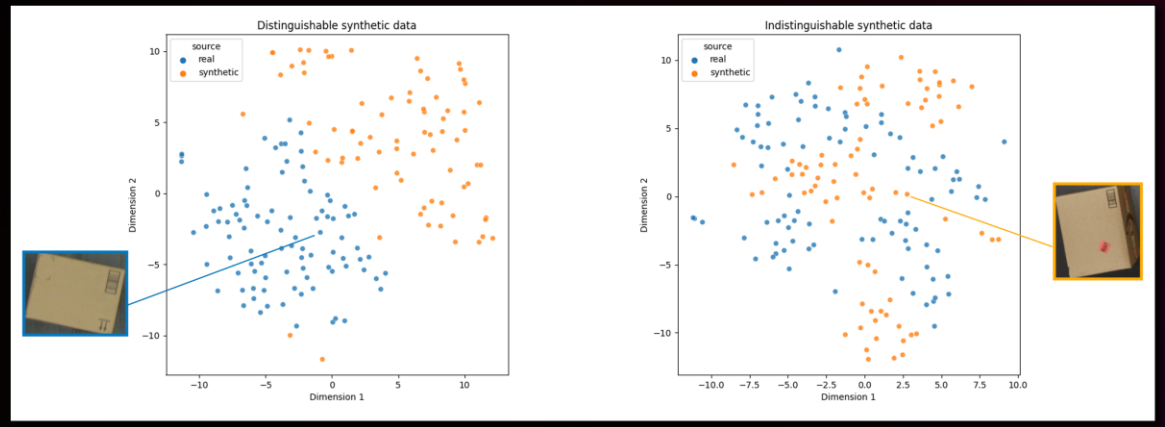
Feedback | Looking for language selection? Find it in the new Unified Settings

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Outcome – Synthetic data





© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Conclusion

- Dive deep into data labeling challenges
- How these challenges are addressed with SageMaker Ground Truth
- Choosing between Ground Truth and Ground Truth Plus when implementing built-in and custom workflows
- How to create synthetic datasets with Amazon SageMaker Ground Truth

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Additional resources

- Amazon SageMaker – https://aws.amazon.com/sagemaker/
- Data labeling with Amazon SageMaker – https://aws.amazon.com/sagemaker/data-labeling
- Amazon SageMaker GroundTruth Plus – https://aws.amazon.com/about-aws/whats-new/2021/12/amazon-sagemaker-ground-truth-plus/
- Blog – https://aws.amazon.com/blogs/aws/amazon-sagemaker-ground-truth-build-highly-accurate-datasets-and-reduce-labeling-costs-by-up-to-70/

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Aashmeet Kalra
@aashmee

Geremy Cohen
@gcohen



Please complete the session survey in the **mobile app**

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.