Introducing Amazon SageMaker Ground Truth

Vikram Madan, SageMaker Ground Truth

December 10, 2018

Data Labeling

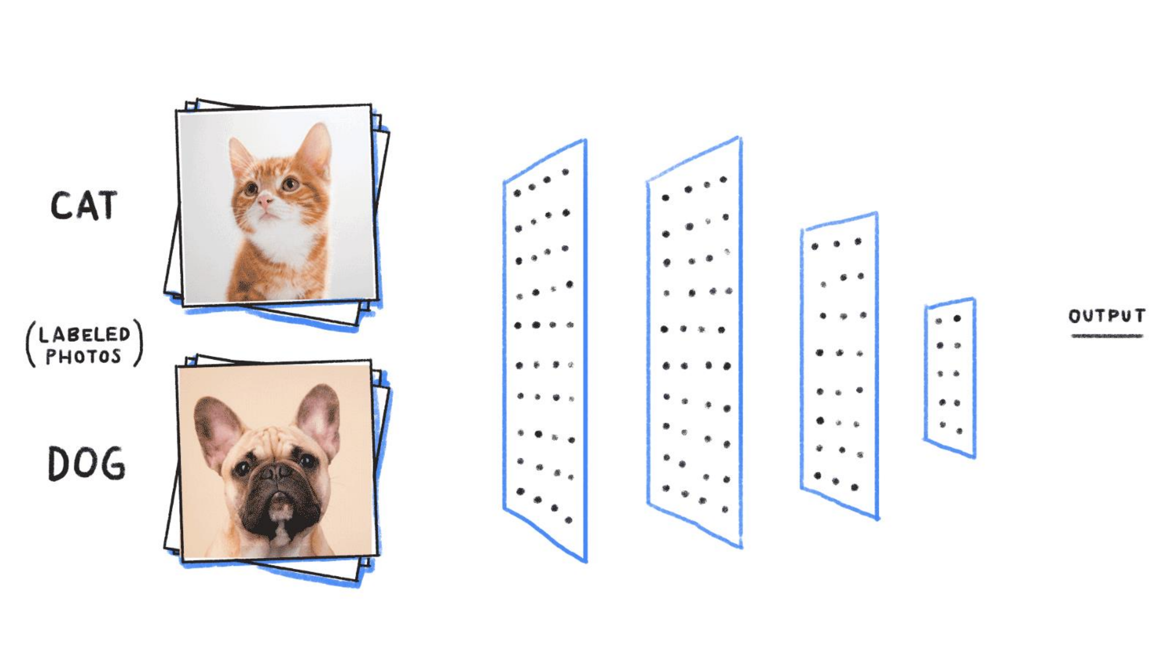
For training machine learning (ML) models

- Text analysis
- Precision agriculture
- Manufacturing efficiency
- Food safety
- Self-driving cars
- Inventory cataloging

and many more use cases…

Source: http://www.digitaljournal.com/tech-and-science/technology/john-deere-advancing-machine-learning-in-agriculture-sector/article/502194

Supervised learning algorithms



Source: https://becominghuman.ai/building-an-image-classifier-using-deep-learning-in-python-totally-from-a-beginners-perspective-be8dbaf22dd8

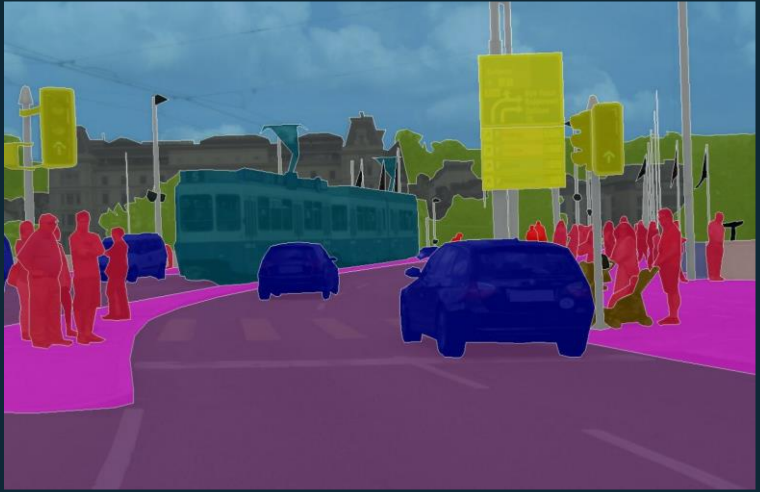
Why is data labeling difficult?

DL models need large labeled datasets

Large number of humans to perform labeling

Difficult to achieve high accuracy for labels

Consumes up to 80% of time to deploy ML



Source: https://medium.com/intro-to-artificial-intelligence/semantic-segmentation-udaitys-self-driving-car-engineer-nanodegree-c01eb6eaf9d

SageMaker Ground Truth

Amazon SageMaker: Build, train, and deploy ML

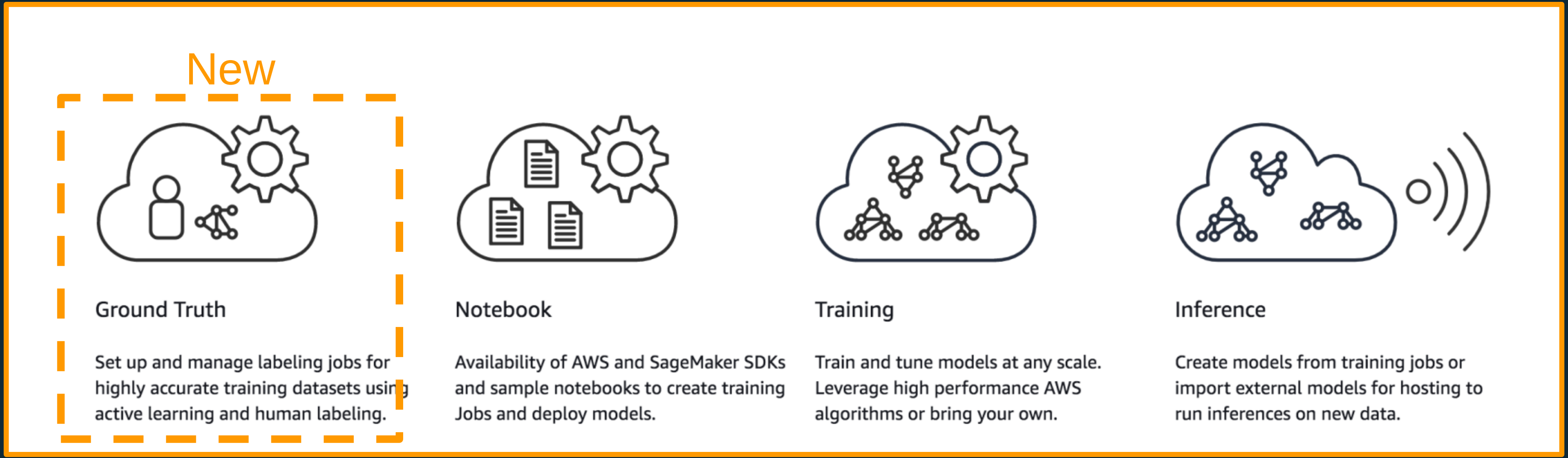
Pre-built notebooks for common problems	Built-in, high performance algorithms	One-click training	Optimization	One-click deployment	Fully managed with auto-scaling
Collect and prepare training data	Choose and optimize your ML algorithm	Set up and manage environments for training	Train and tune model (trial and error)	Deploy model in production	Scale and manage the production environment

intuit. F1 tinder SIEMENS NFL CONVOY

SIEMENS THOMSON REUTERS GE Healthcare Celgene Liberty Mutual. DOW JONES

aws

Extension of Amazon SageMaker



New

Ground Truth

Set up and manage labeling jobs for highly accurate training datasets using active learning and human labeling.

Notebook

Availability of AWS and SageMaker SDKs and sample notebooks to create training Jobs and deploy models.

Training

Train and tune models at any scale. Leverage high performance AWS algorithms or bring your own.

Inference

Create models from training jobs or import external models for hosting to run inferences on new data.

Label machine learning training data easily and accurately

Key Features

Data Labeling Jobs

- Use pre-built templates for image and text labeling tasks
- Create customized tasks for your specific image and text labeling requirements

Automated Data Labeling

- Prioritize which data goes to humans first (“not all data is created equal”)
- Get part of your data labeled automatically (reduces redundant / unnecessary labeling)

High Accuracy Labeling

- Improve accuracy with annotation consolidation and UI templates with built-in labeling UX best practices

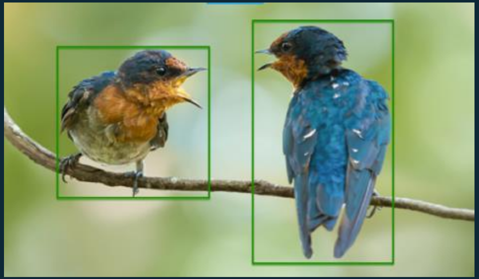
Dataset and Label Management

- Query and analyze the results of your labeling jobs
- Track and manage your datasets and enable easy integration with your data lake

Multiple Workforce Options

- Scale out labeling easily with the public Mechanical Turk workforce
- Direct work to your own workers or use vendor workforces listed on AWS Marketplace

Supported Data Labeling Tasks



Bounding boxes

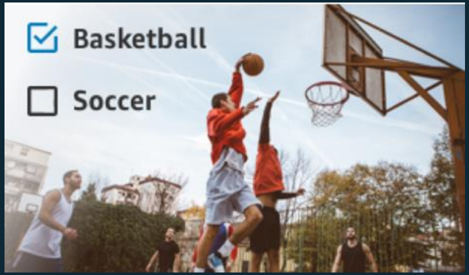
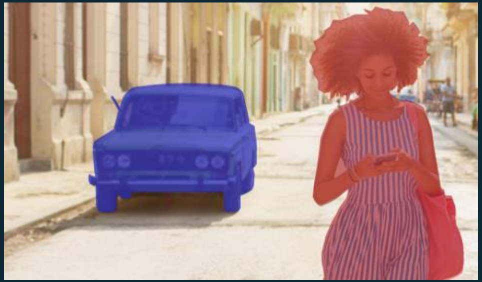
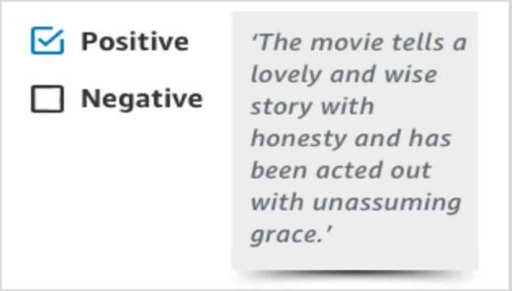


Image Classification



Semantic Segmentation



☑ Positive

□ Negative

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

Text Classification

Custom tasks

Supported Workforce Options

Public

An on-demand 24 x7 workforce of over 500,000 independent Contractors worldwide, powered by Amazon Mechanical Turk

Private

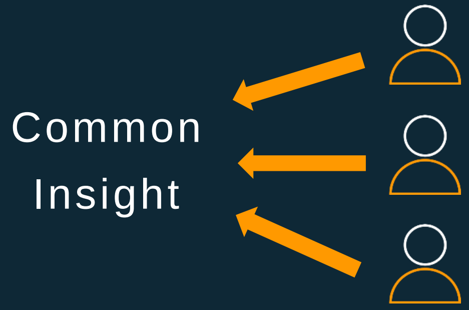
A team of workers that you have sourced yourself, including your own employees or contractors for handling data that needs to stay within your organization

Vendors

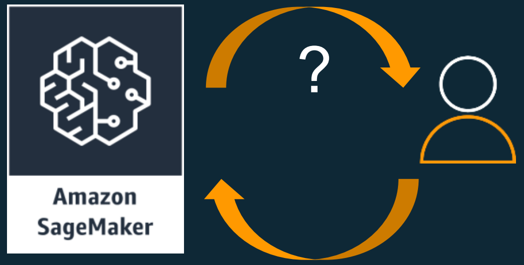
A curated list of third party vendors that specialize in providing data labeling services, available via the AWS Marketplace

Improving accuracy and efficiency of data labeling

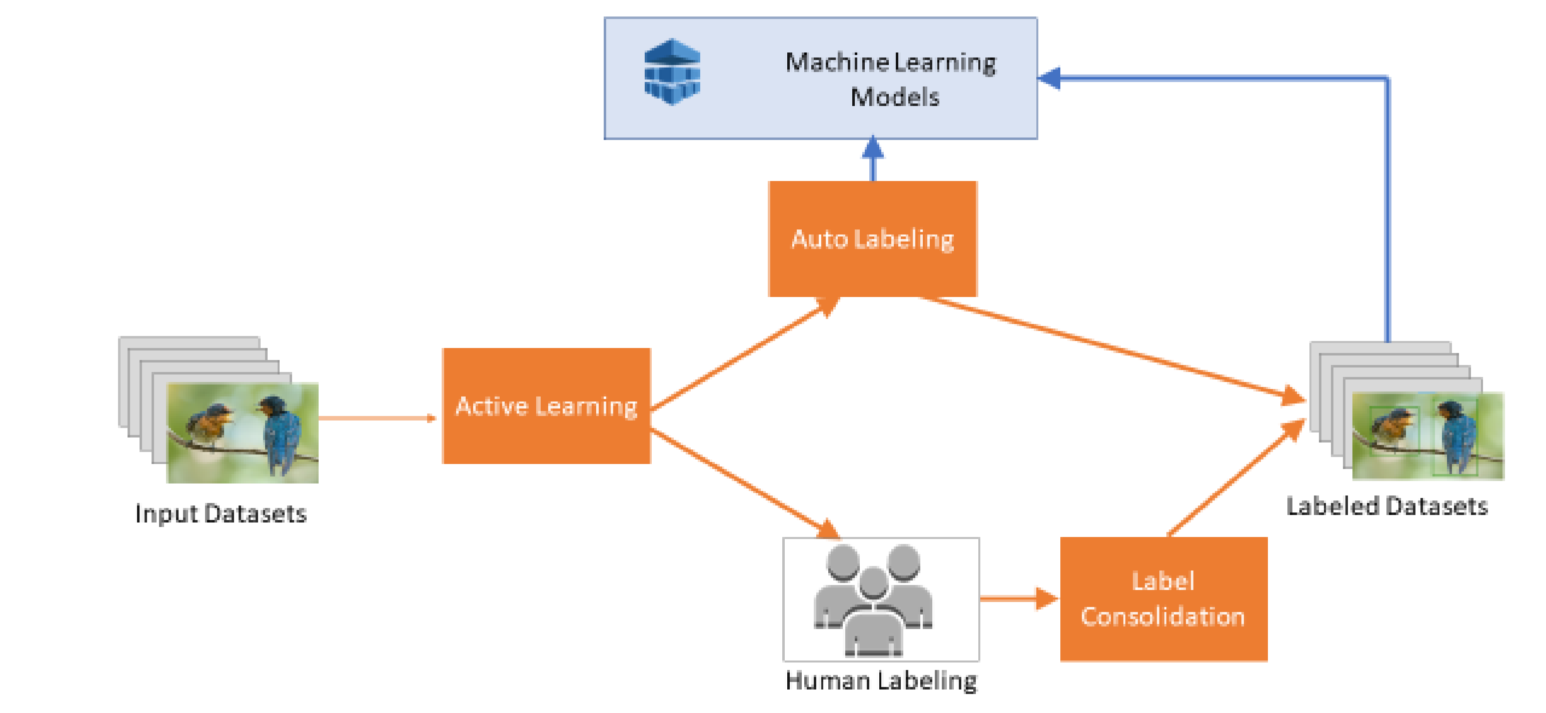
- Consolidate annotations from multiple workers



- Only send to humans examples which are hard for the machines to label well



Automated data labeling



Pricing

Automated data labeling

Base Pricing

For each labeling job that you run with Ground Truth, you are billed per labeled object:

Data Labeling Jobs	
Pricing Tier	**Price per Object**
Less than 50,000 objects	$0.08 per object
50,000 to 1,000,000 objects	$0.04 per object
Over 1,000,000 objects	$0.02 per object

Free Tier: Up to a total of 1,000 objects for the first 2 months after SageMaker sign-up

Public Workforce (*Optional*)

Task Type	Price per Task
Task 1 (under 5 seconds)	$0.012
Task 2 (5 - 7 seconds)	$0.024
Task 3 (8 - 10 seconds)	$0.036
Task 4 (11 - 13 seconds)	$0.048
Task 5 (14 - 16 seconds)	$0.060
Task 6 (17 – 24 seconds)	$0.072
Task 7 (25 - 30 seconds)	$0.120
Task 8 (45 - 60 seconds)	$0.240
Task 9 (60 - 90 seconds)	$0.360
Task 10 (1.5 - 2 minutes)	$0.480
Task 11 (2 - 2.5 minutes)	$0.600
Task 12 (2.5 - 3 minutes)	$0.720
Task 13 (3 - 3.5 minutes)	$0.840
Task 14 (3.5 - 4 minutes)	$0.960
Task 15 (4 - 4.5 minutes)	$1.080
Task 16 (4.5 - 5 minutes)	$1.200

Auto Labeling (*Optional*)

Passed Through:

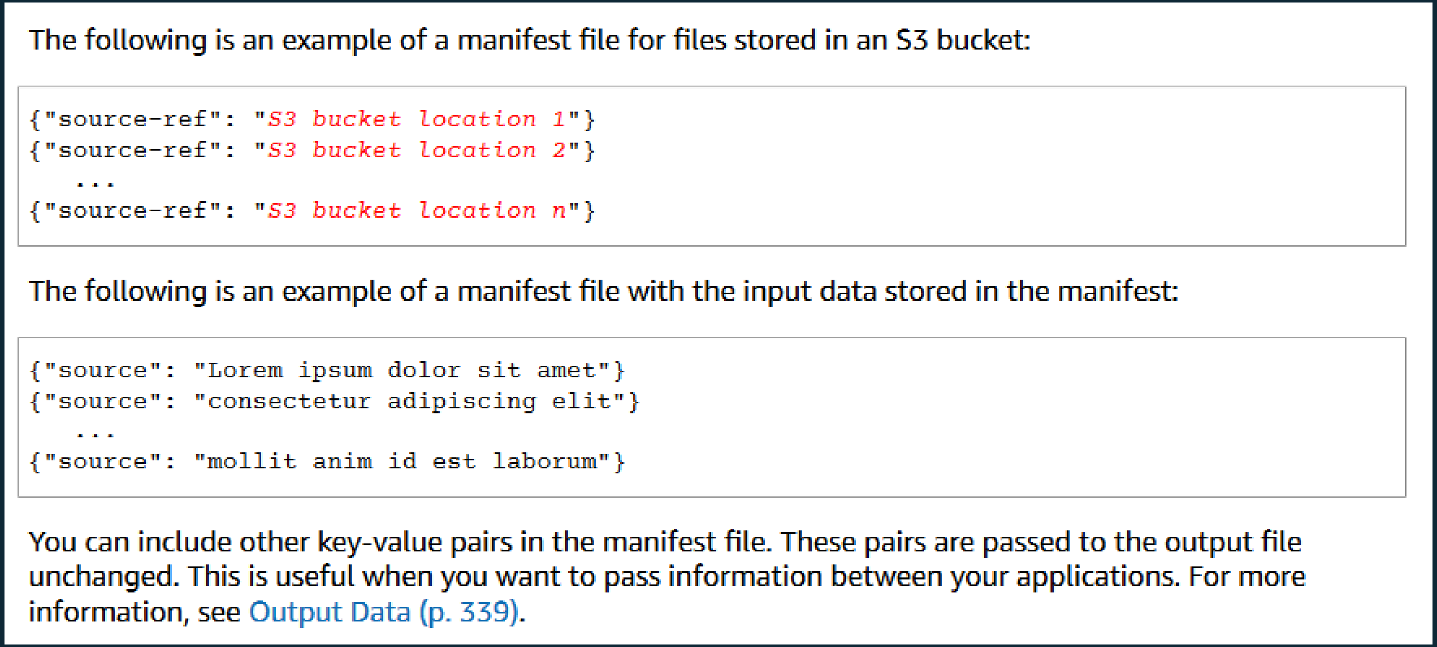
Vendor Pricing

Pricing for the vendor workforces are set by the vendor. At GA, vendors charge by the hour, rounded to the nearest 10 second increments. Pricing info is displayed on the product detail page.

Vendor	Location	Price
iMerit	India	$5/hour
iMerit	US	$25/hour
Smart One Group	Madagascar	$6/hour

Demos

Input and output dataset



The following is an example of a manifest file for files stored in an S3 bucket:

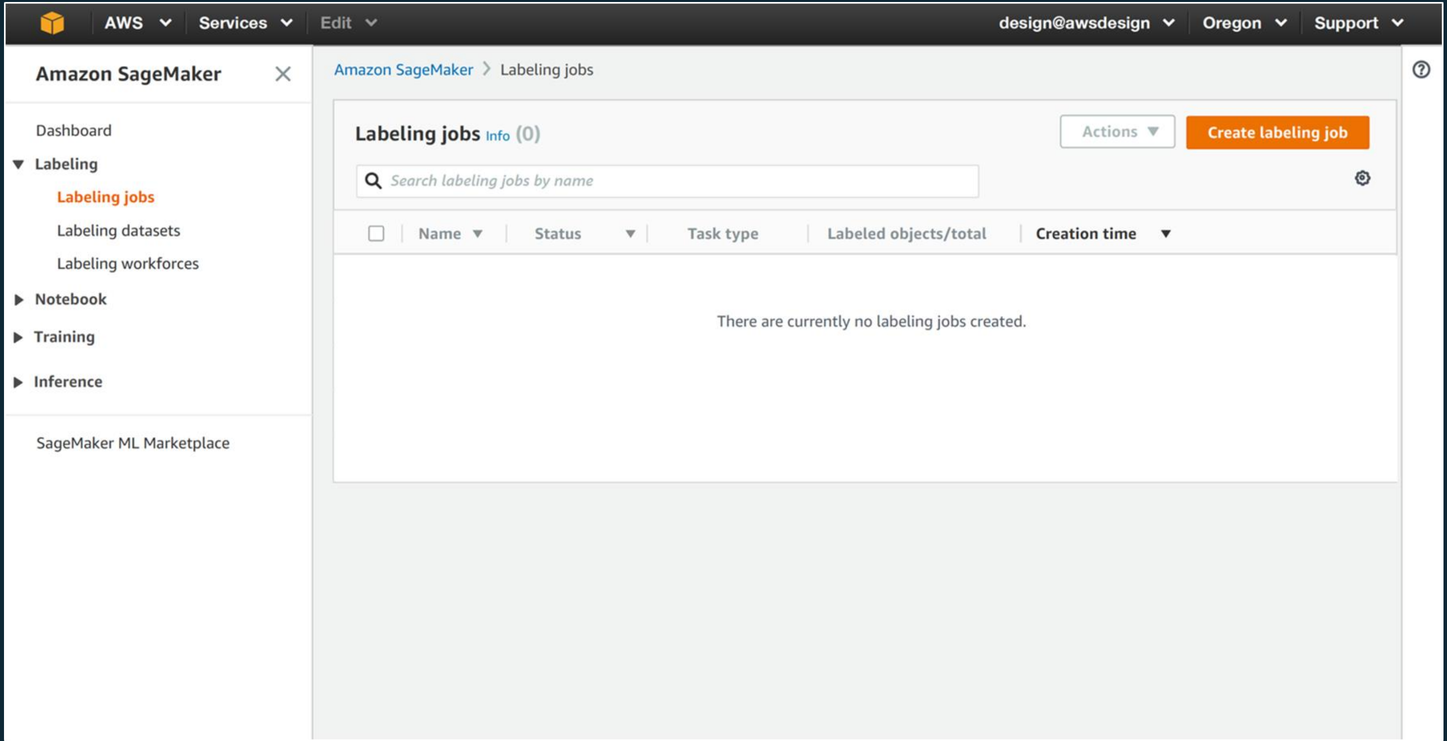
```
{"source-ref": "S3 bucket location 1"}
{"source-ref": "S3 bucket location 2"}
   ...
{"source-ref": "S3 bucket location n"}
```

The following is an example of a manifest file with the input data stored in the manifest:

```
{"source": "Lorem ipsum dolor sit amet"}
{"source": "consectetur adipiscing elit"}
   ...
{"source": "mollit anim id est laborum"}
```

You can include other key-value pairs in the manifest file. These pairs are passed to the output file unchanged. This is useful when you want to pass information between your applications. For more information, see Output Data (p. 339).

Create labeling job



There are currently no labeling jobs created.