

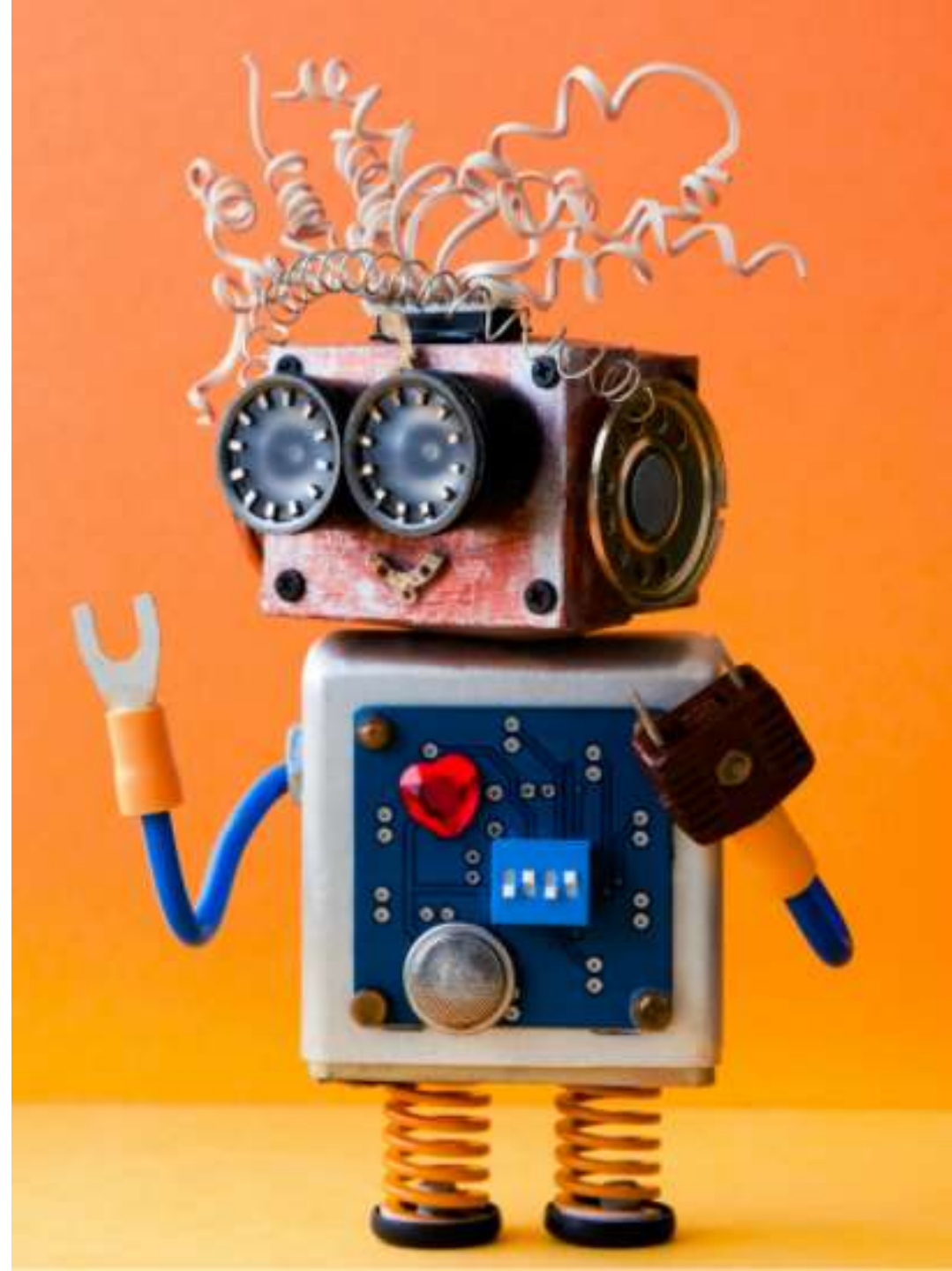
# PROJECT OVERVIEW & INTRO TO GROUNDTRUTH

---



EASY

ADVANCED



# PROJECT OVERVIEW

- In this project, we will learn how to label images using Amazon SageMaker GroundTruth.
- These are the key learning outcomes:
  1. The need for labelled datasets
  2. Applications of supervised learning
  3. Challenges of obtaining labelled datasets
  4. Learn how to define a labeling job using Amazon SageMaker Groundtruth
  5. Learn how to label data in Amazon SageMaker Groundtruth
  6. Understand the concept of JsonL and manifest files

## 20 IMAGES BELONGS TO 4 CLASSES (BALANCED DATASET)

### BAGS



001



002



003



004



005

### EYEWEAR



006



007



008



009



010

### FLIPFLOPS



011



012



013



014



015

### WATCHES



016



017



018



019



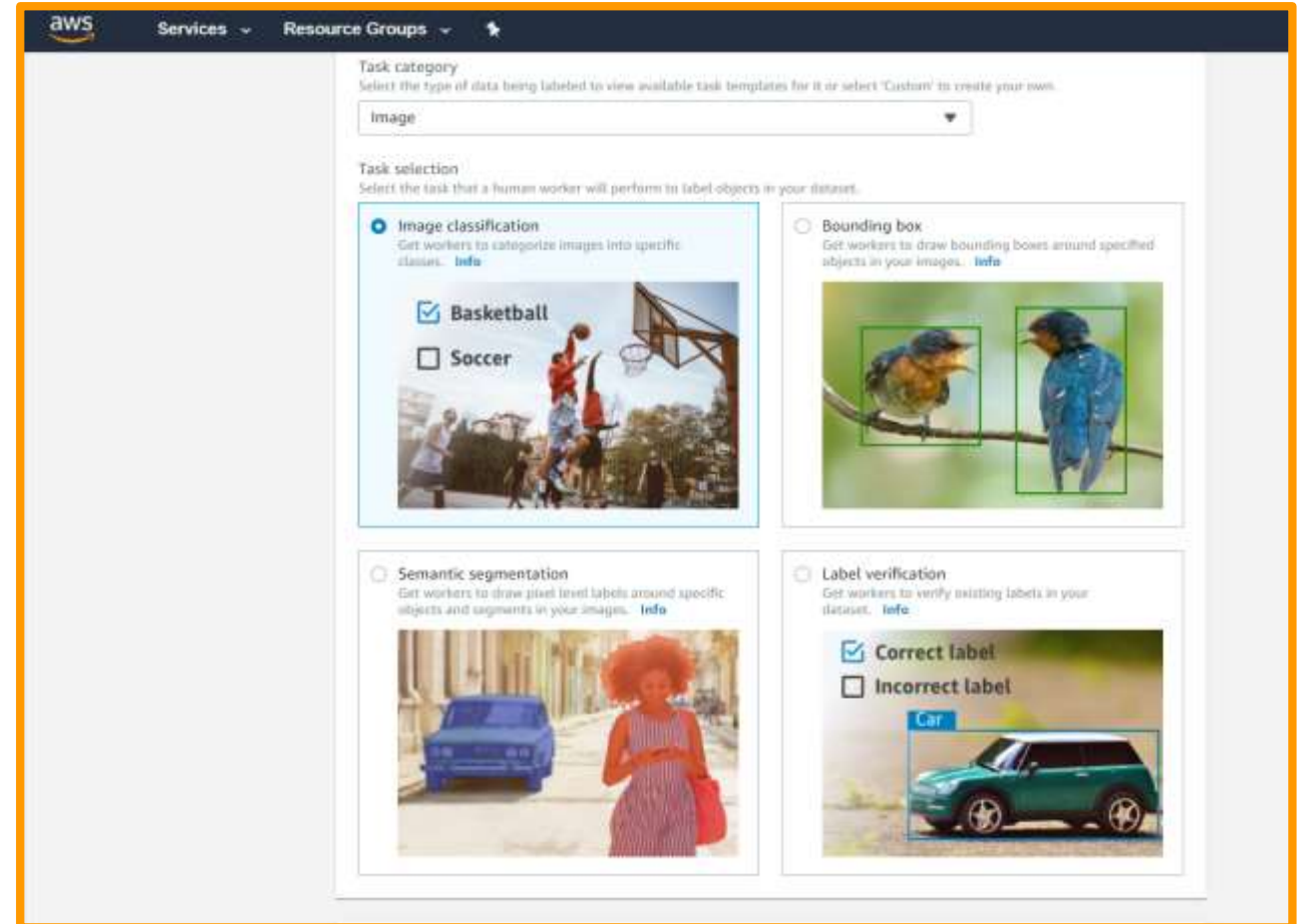
020

# SAGEMAKER GROUNDTRUTH 101

- AWS SageMaker GroundTruth is a service offered by AWS to label data.
- In machine learning terminology, Ground truth means “gold standard”!
- Ground Truth indicates the “true” or “real” class that you would like your model to learn how to predict.



**AMAZON  
SAGEMAKER  
GROUNDTRUTH**



# GROUNDTRUTH KEY LABELING TASKS AND FEATURES

- Several labeling tasks are available in SageMaker GroundTruth:
  - Bounding boxes
  - Image Classification
  - Semantic Segmentation
  - Text Classification
  - Custom Tasks
- Check this out: <https://aws.amazon.com/sagemaker/data-labeling/features/>

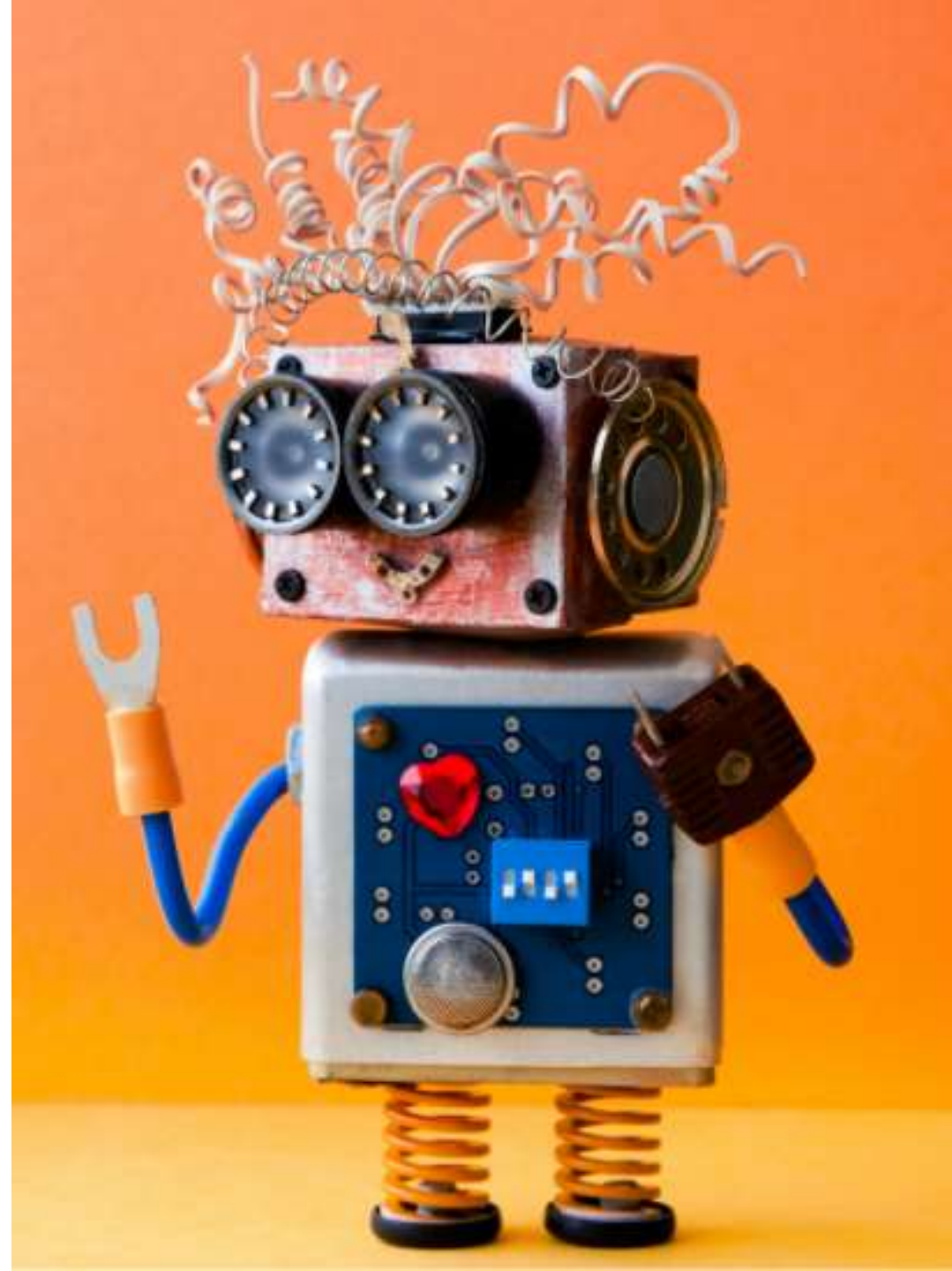
# AVAILABLE DATA LABELING WORKFORCES

**Public Mechanical Turks:** Amazon SageMaker facilitates the interaction between customers who require data labeling and an on-demand 24x7 global workforce of 500K contractors globally.

**Private:** A team of labelers can be specified by the customer including their own private labelers. Everything is managed through SageMaker Ground Truth. No need for labelers to have IAM or Amazon account.

**Vendors:** SageMaker Groundtruth provides a curated third-party vendors who can offer data labeling services.

# 21<sup>st</sup> CENTURY NEW GOLD





# WHY DATA IS CONSIDERED THE NEW GOLD OF THE 21st CENTURY?

- The data revolution is here! Data is the new gold of the 21st Century.
- Companies nowadays have access to a massive amount of data and their competitive advantage lies in their ability to gain valuable insights from this data.
- Data can empower companies to boost their revenues, improve processes and reduce costs.
- Data could be leveraged in many industries such as finance, banking, healthcare, transportation, and technology sectors.



Image Source: [https://www.flickr.com/photos/tao\\_zhyn/442965594](https://www.flickr.com/photos/tao_zhyn/442965594)

# DATA-DRIVEN SUCCESS STORIES

- Netflix leverages customer data to recommend new content to users using its data-driven recommender system which earns it ~\$1B USD in customer retention.
- Amazon integrated data-driven recommendations at every stage of the purchasing process which resulted in 29% sales increase to \$12.83 billion during its second fiscal quarter 2020, up from \$9.9 billion during same time last year”.



Source: <https://medium.com/eleks-labs/4-powerful-use-cases-for-data-science-in-finance-35d50075ff80>

Source: <https://emerj.com/ai-sector-overviews/ai-in-banking-analysis/>

Source: <https://www.precisely.com/blog/big-data/beyond-big-data-examples-success>

Photo: <https://commons.wikimedia.org/wiki/File:Logonfx.png>

Photo: [https://commons.wikimedia.org/wiki/File:Amazon\\_PNG6.png](https://commons.wikimedia.org/wiki/File:Amazon_PNG6.png)



# DATA-DRIVEN SUCCESS STORIES

- “JP Morgan invests \$11.5 billion/year in new data driven technologies. Its machine learning-based Contract Intelligence (COiN) platform reviews 12,000 commercial loan agreements in few hours compared to 360,000 man-hours it would take to do so manually.”
- “Banking institutions prevented \$22 billion worth of fraudulent transactions in 2018 with the power of AI/ML”.
- “Bank of America introduced Erica chatbot that served 6 million users as of March 2019.”
- Electronic trades account for almost 45% of revenues in cash equities trading” U.K. research firm Coalition Report.



Source: <https://medium.com/eleks-labs/4-powerful-use-cases-for-data-science-in-finance-35d50075ff80>

Source: <https://emerj.com/ai-sector-overviews/ai-in-banking-analysis/>

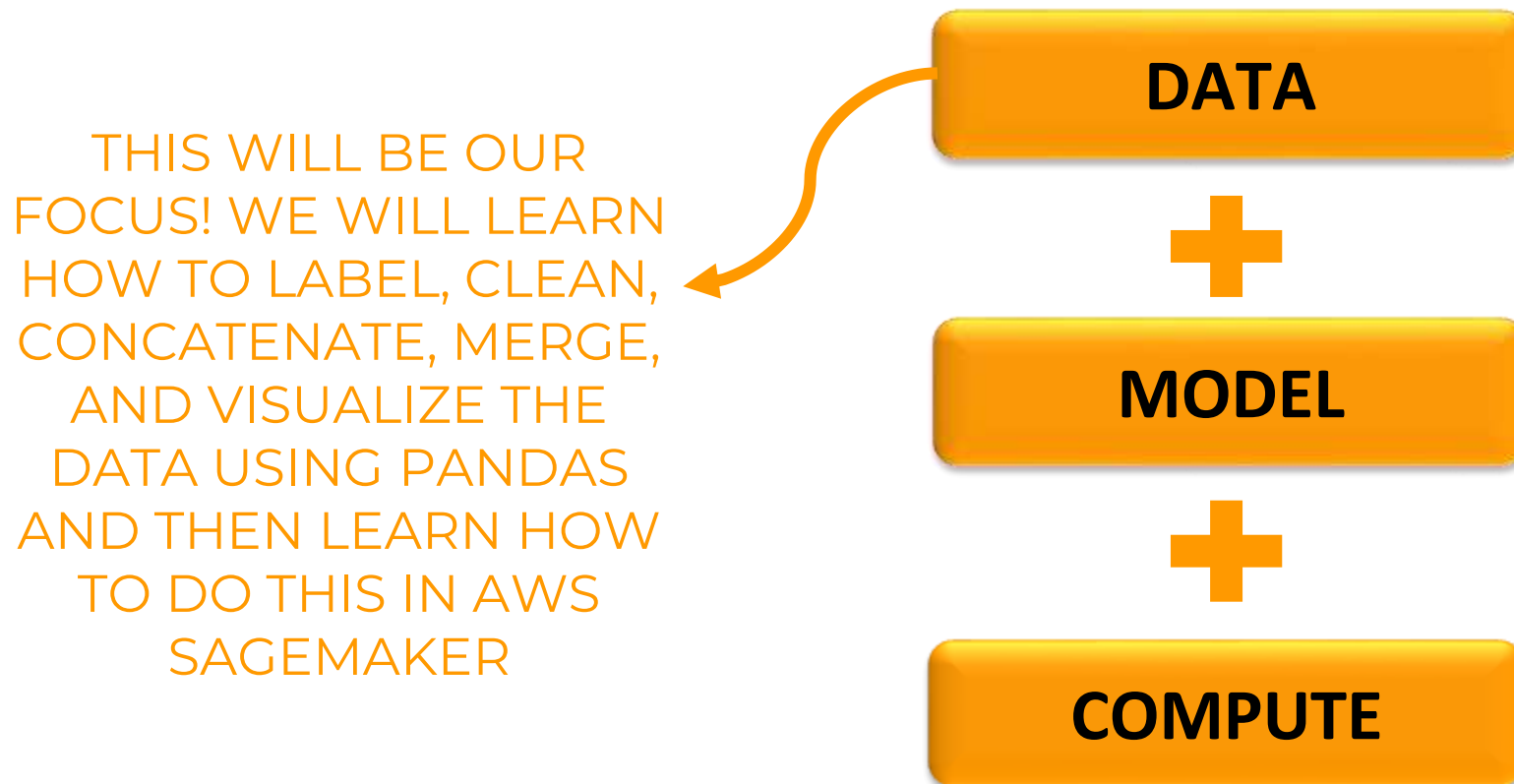
Source: <https://www.precisely.com/blog/big-data/beyond-big-data-examples-success>

Photo: <https://www.flickr.com/photos/bensutherland/178395814>

Photo: <https://www.flickr.com/photos/moneyblognewz/5280927344>

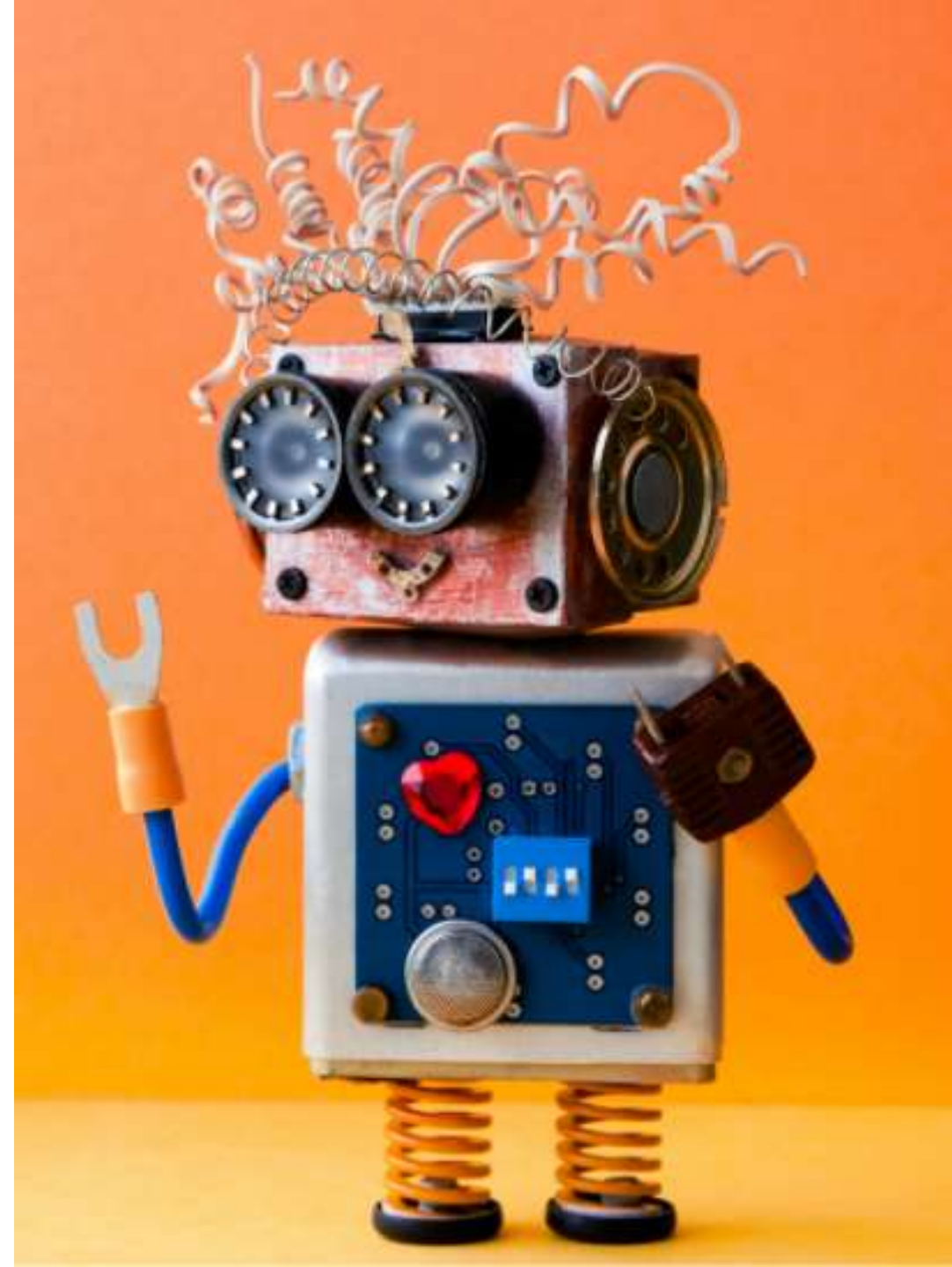
# CORPORATE INTELLIGENCE WITH AI

- Artificial intelligence is the science that empowers computers to mimic human intelligence such as decision making, text processing, and visual perception.
- In order to train Artificial intelligence and Machine Learning models, companies need a massive amount of data.



# DATA SOURCES AND TYPES

---



# DATA SOURCES

- Data can come from so many sources and forms such as images, audio, video, and text.
- Collecting, structuring and analysing this data is critical for companies to gain customers insights and set their marketing and product strategies.
- “Data preparation accounts for about 80% of the work of data scientists”, Forbes.

IMAGE/VIDEO



TEXT (CORPUS)



AUDIO/SOUND



TIMESERIES/SIGNALS



Source: <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=410bbd6a6f63>

Photo Credit: <https://pxhere.com/en/photo/1454351>

Photo Credit: <https://www.flickr.com/photos/29881930@N00/2086641598>

Photo Credit: [https://commons.wikimedia.org/wiki/File:Mobile\\_phone\\_text\\_messages.jpg](https://commons.wikimedia.org/wiki/File:Mobile_phone_text_messages.jpg)

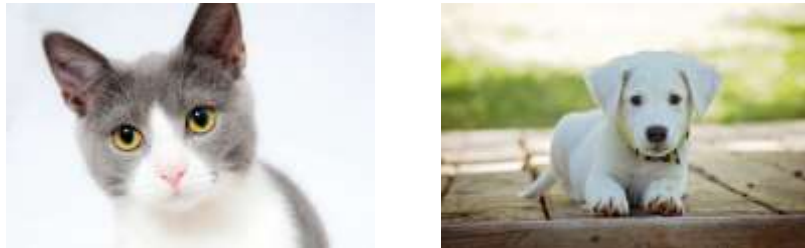
Photo Credit: [https://en.wikipedia.org/wiki/File:Messages\\_Yosemite.svg](https://en.wikipedia.org/wiki/File:Messages_Yosemite.svg)

Photo Credit: <https://www.pexels.com/photo/blue-and-yellow-graph-on-stock-market-monitor-159888/>

# DATA TYPES: LABELED VS. UNLABELED DATASET

- There are generally two types of data that we could use to train AI models.

## UNLABELED DATASET



*Unlabeled data consists of data that does not have explanation (class or tag) associated with it.*

## LABELED DATASET

**LABEL = "CAT"**



**LABEL = "DOG"**



*Labeled data consists of unlabeled data but with a "class" or "tag" associated with it.*

Photo Credit: <https://www.pexels.com/photo/grey-and-white-short-fur-cat-104827/>

Photo Credit: <https://www.pexels.com/photo/portrait-of-a-dog-257540/>

# GOOD Vs. BAD DATA

## GOOD DATA

Many samples (large number of data points)

Not Biased

Does not contain missing data points

Only contains (relevant) important features

Does not contain duplicate samples

## BAD DATA

Few samples (small number of data points)

Biased

Contains missing data points

Contains many irrelevant (useless) features

Contains duplicate samples



# WHERE DOES THIS DATA COME FROM?

- Data could also come from multiple sources such as Kaggle, UCI, and AWS Dataset.
- ImageNet is an open source repository of images consisting of 21,841 subcategories (classes) and over 14 million images.
- AWS dataset registry: <https://registry.opendata.aws/>

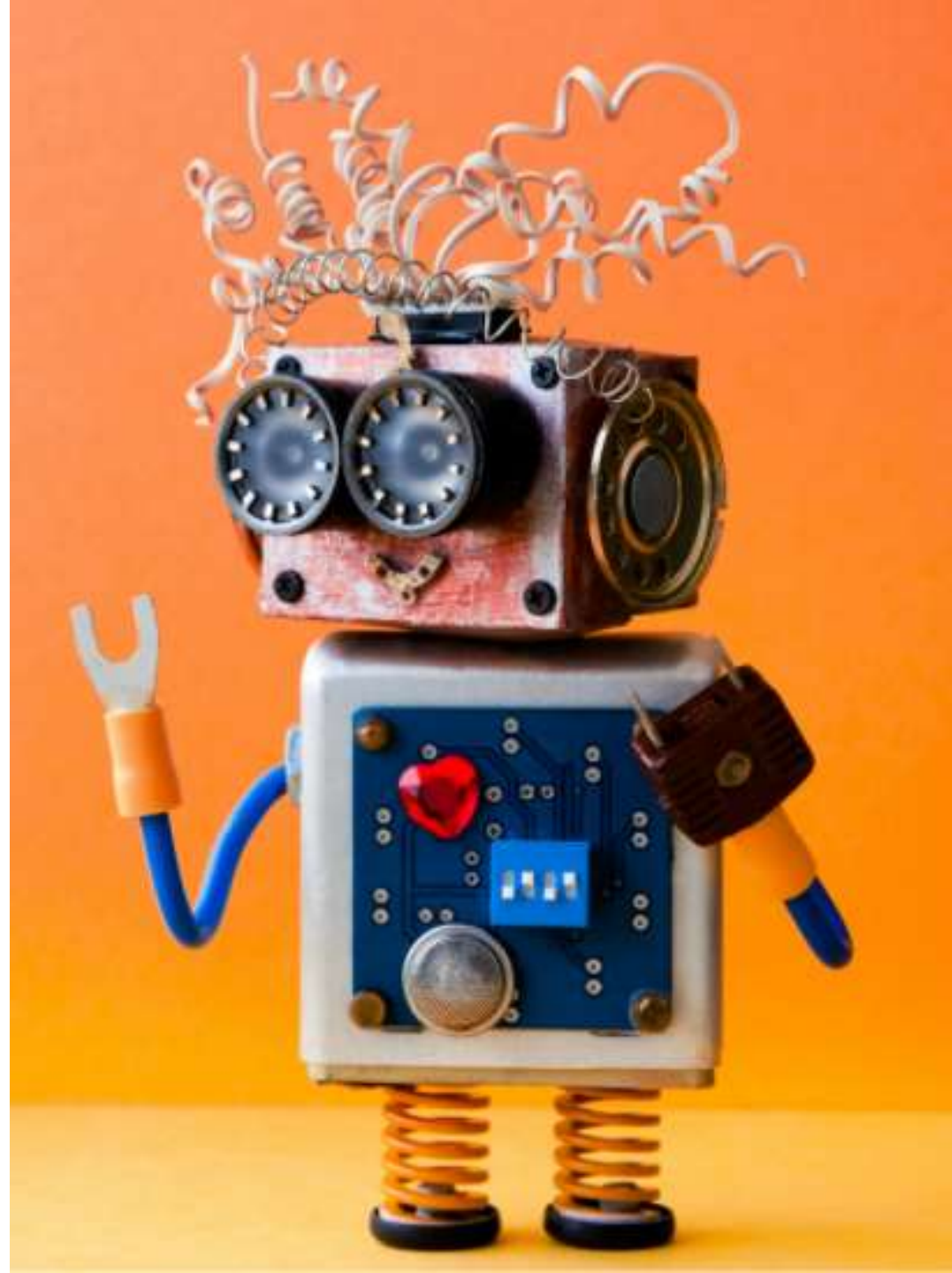
The image displays three screenshots of data repositories. The left screenshot shows the Kaggle Datasets page with a search bar and a list of datasets. The middle screenshot shows the UCI Machine Learning Repository page with a table of datasets. The right screenshot shows the AWS Registry of Open Data page with a search bar and a list of datasets.

Dataset Name	Data Type	Dataset Size	Number of Rows	Number of Columns	Number of Features	
Abalone	Tabular	Classification	Category: Integer	1101	8	1000
Adult	Tabular	Classification	Category: Integer	24512	14	1000
Amazon	Tabular	Classification	Category: Integer	76	35	1000
Amazon, Microsoft, Web Data	Tabular	Classification	Category: Integer	21111	240	1000
Barcode	Tabular	Classification	Category: Integer	813	175	1000
Artificial Character	Tabular	Classification	Category: Integer	1000	7	1001
Artificial Character	Tabular	Classification	Category: Integer	220	1	1001
Artificial Character	Tabular	Classification	Category: Integer	220	49	1001
Auto MPG	Tabular	Regression	Category: Integer	398	8	1001
Automobile	Tabular	Regression	Category: Integer	398	8	1001
California	Tabular	Regression	Category: Integer	204	5	1004

Check out website here: <https://archive.ics.uci.edu/ml/datasets.php>

Check out website here: <https://www.kaggle.com/datasets>

# WHY DO WE NEED LABELED DATA?



# ML TRAINING STRATEGIES

## Supervised Learning

- Used in cases where large dataset with known labels (outputs) are available.
- The learning algorithm evaluates output (i.e.: makes predictions), compares output against the label, and adjust model weights/parameters and repeat.

## Unsupervised Learning

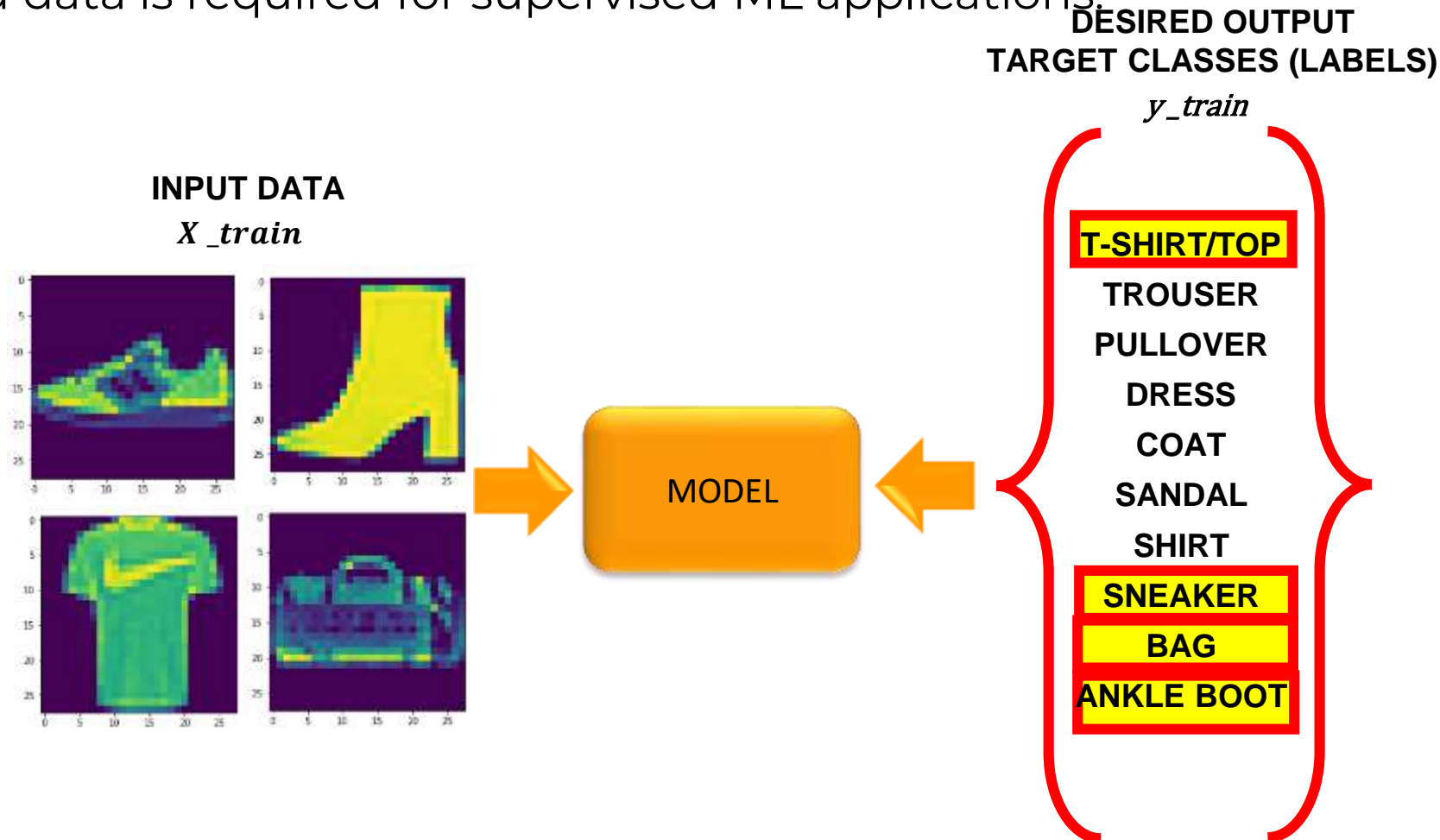
- Used with "unlabeled" data (not categorized) (Ex: k-means clustering).
- Since learning algorithm works with unlabeled data, there is no way to assess the accuracy of the structure suggested by the algorithm

## Reinforced Learning

- Learning algorithm takes actions that maximizes cumulative reward.
- Over time, ML models learn to prefer the right kind of action and avoid the wrong one.

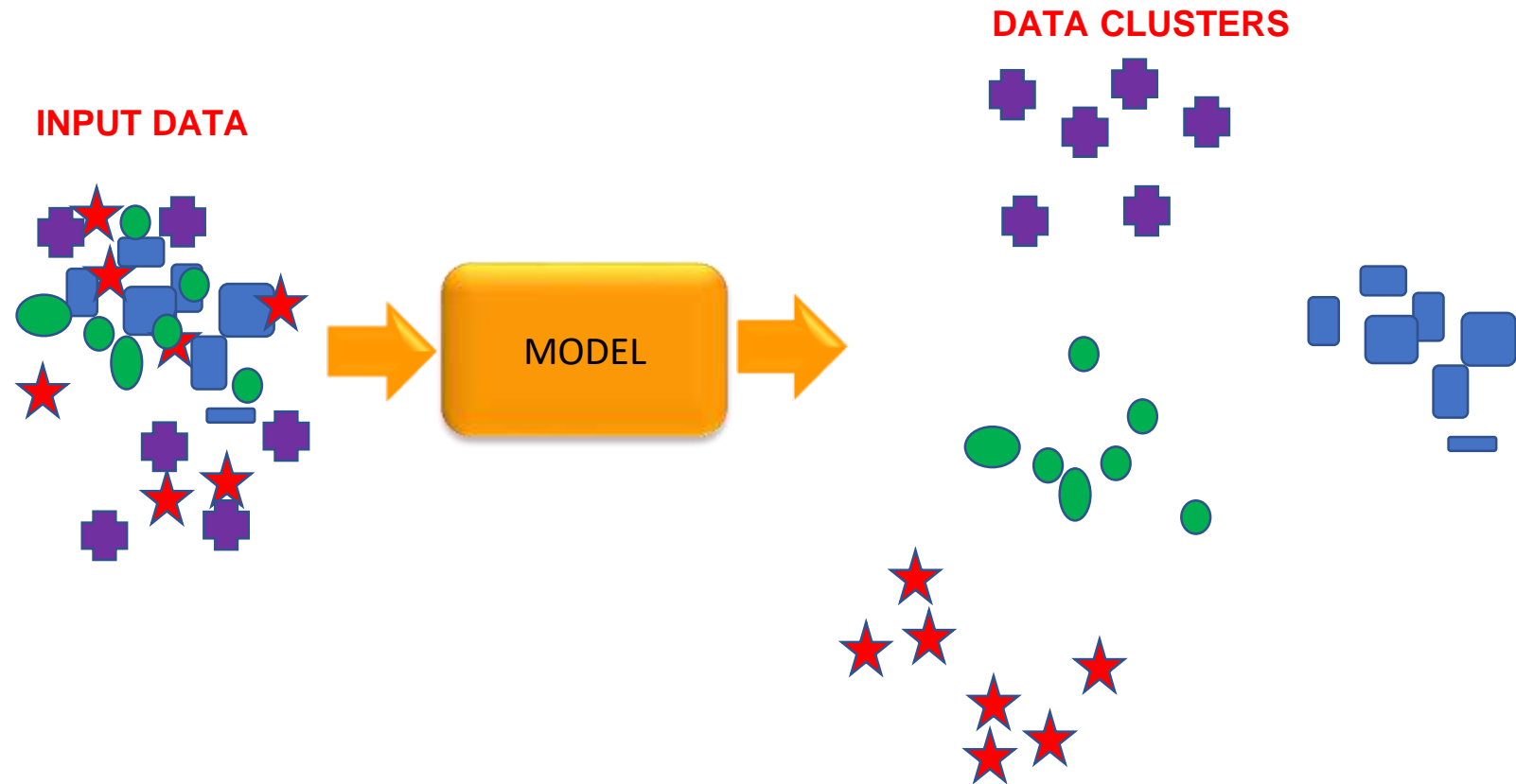
# MACHINE LEARNING: SUPERVISED LEARNING

- Supervised: used to train algorithms using labeled input and output data.
- Performance is assessed by comparing trained model prediction vs. real output.
- Labeled data is required for supervised ML applications.



# MACHINE LEARNING: UNSUPERVISED LEARNING

- Unsupervised learning: provides the algorithm with no labeled data.
- The algorithm attempts at discovering hidden patterns within the training data.
- Unsupervised learning methods can analyze complex data that humans might find difficult to interpret.
- No feedback!



# MACHINE LEARNING: REINFORCEMENT LEARNING

- Reinforcement learning allows machines take actions to maximize cumulative reward.
- Reinforcement algorithms learn by trial and error through reward/penalty.
- Two elements: environment and learning agent.
- The environment rewards the agent for correct actions.
- Based on the reward or penalty, agent improves its environment knowledge to make better decision.

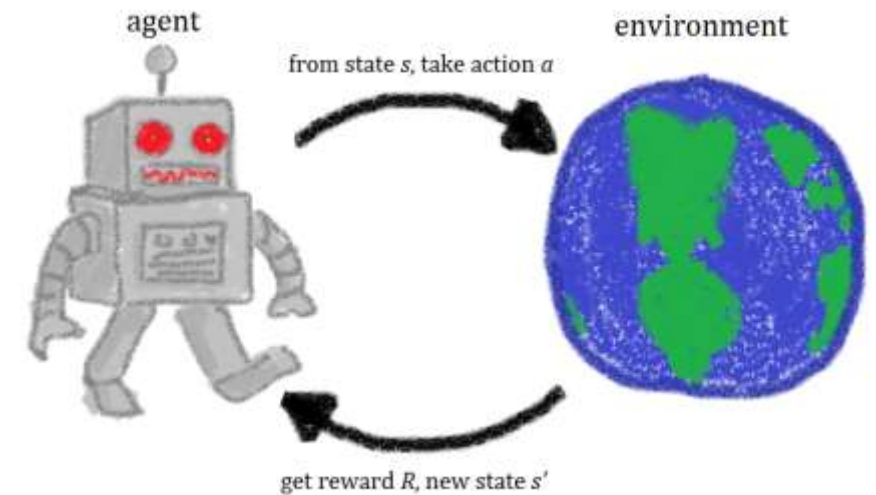
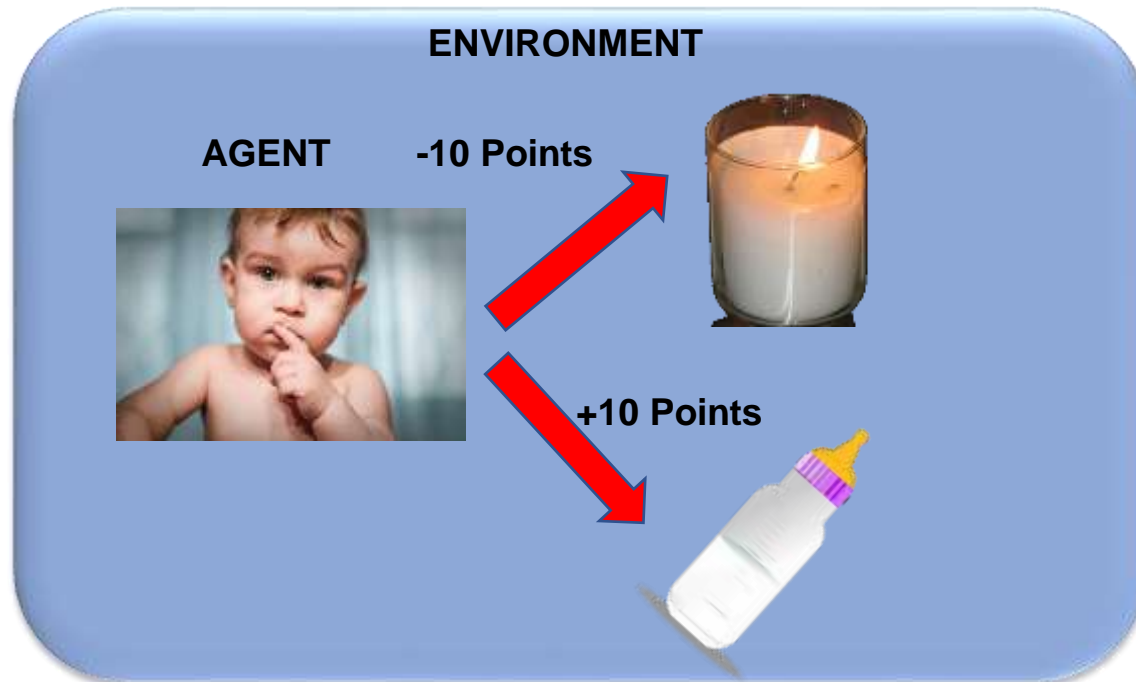
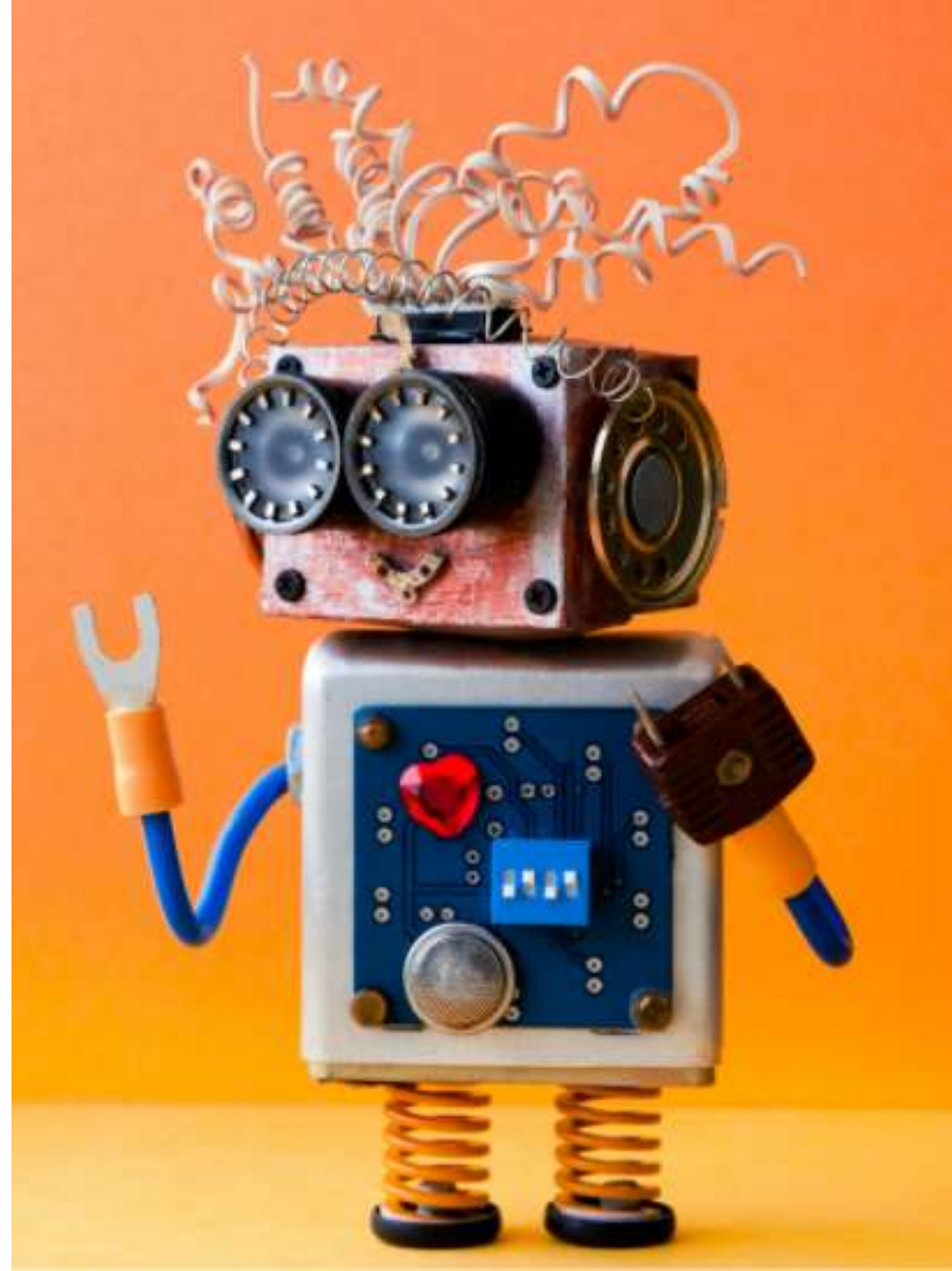


Photo Credit: [https://commons.wikimedia.org/wiki/File:RL\\_agent.png](https://commons.wikimedia.org/wiki/File:RL_agent.png)



# DATA LABELING CHALLENGES & APPLICATIONS



# DATA LABELING CHALLENGES

Modern state of the art deep learning models require a massive amount of labeled datasets

Large team of Human labelers are required to do the labeling

Inherent biases in human labelers require controls/standardization

Most data scientists' time is spent curating and labeling data

# USE CASES OF DATA LABELING

Text Data Analysis  
(News and Social Media  
Feeds)

Manufacturing Services:  
identify defects from  
images

Self-Driving Cars: label  
pedestrians, vehicles  
and traffic devices

AI-powered precision  
Agriculture: Identify  
precise locations of  
crops vs. weeds from  
images

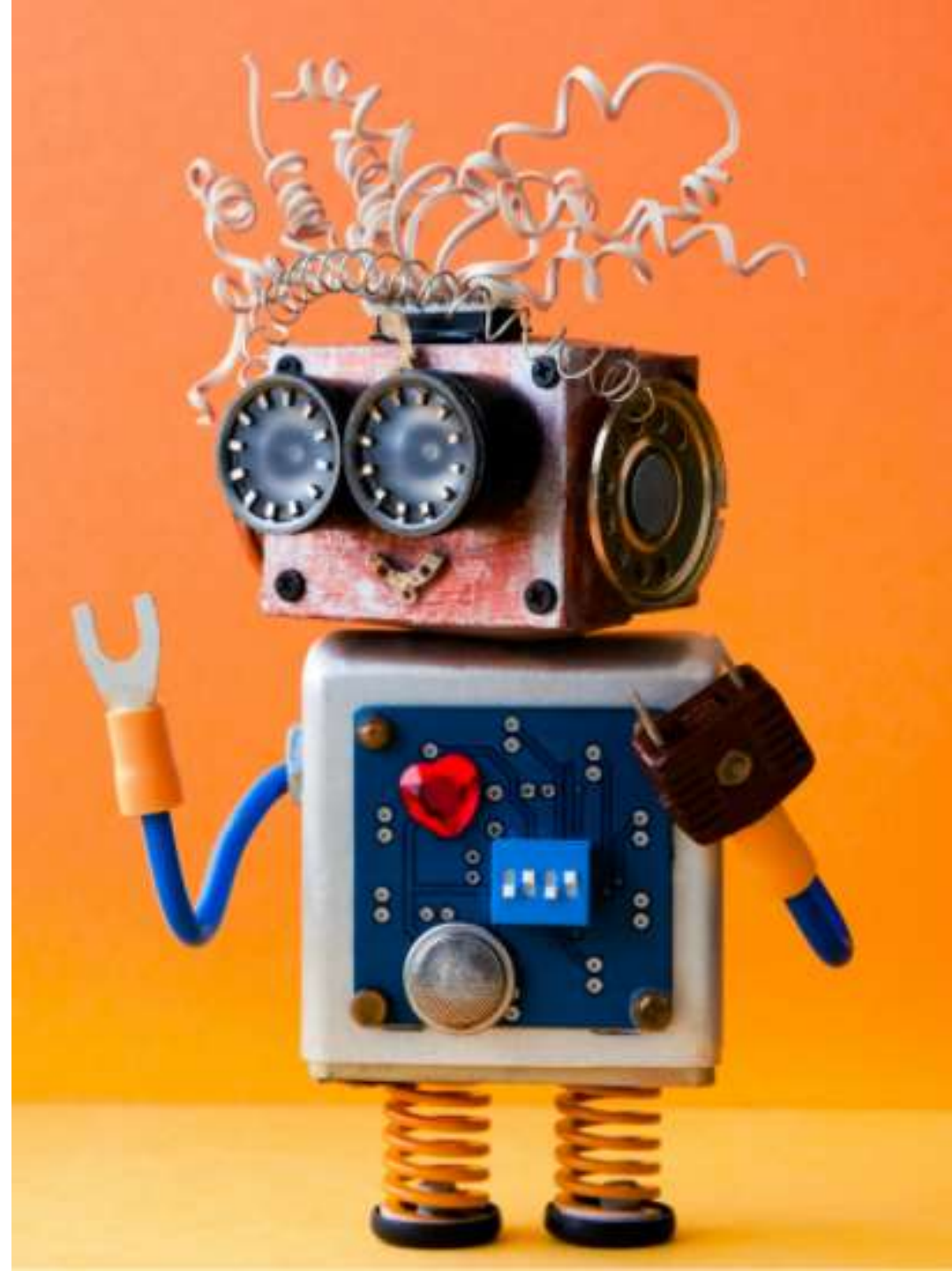
# JSON LINES FORMAT & MANIFEST FILES

---



EASY

ADVANCED



# JSON LINES (JSONL)

- JSON Lines text format is a format used for storing structured data that could be processed one record at a time.
- JSON Line format is generally used to store data labels.

```
{"Image 1": "Cat"}  
{"Image 2": "Dog"}  
{"Image 3": "Lion"}
```

# MANIFEST FILES 101

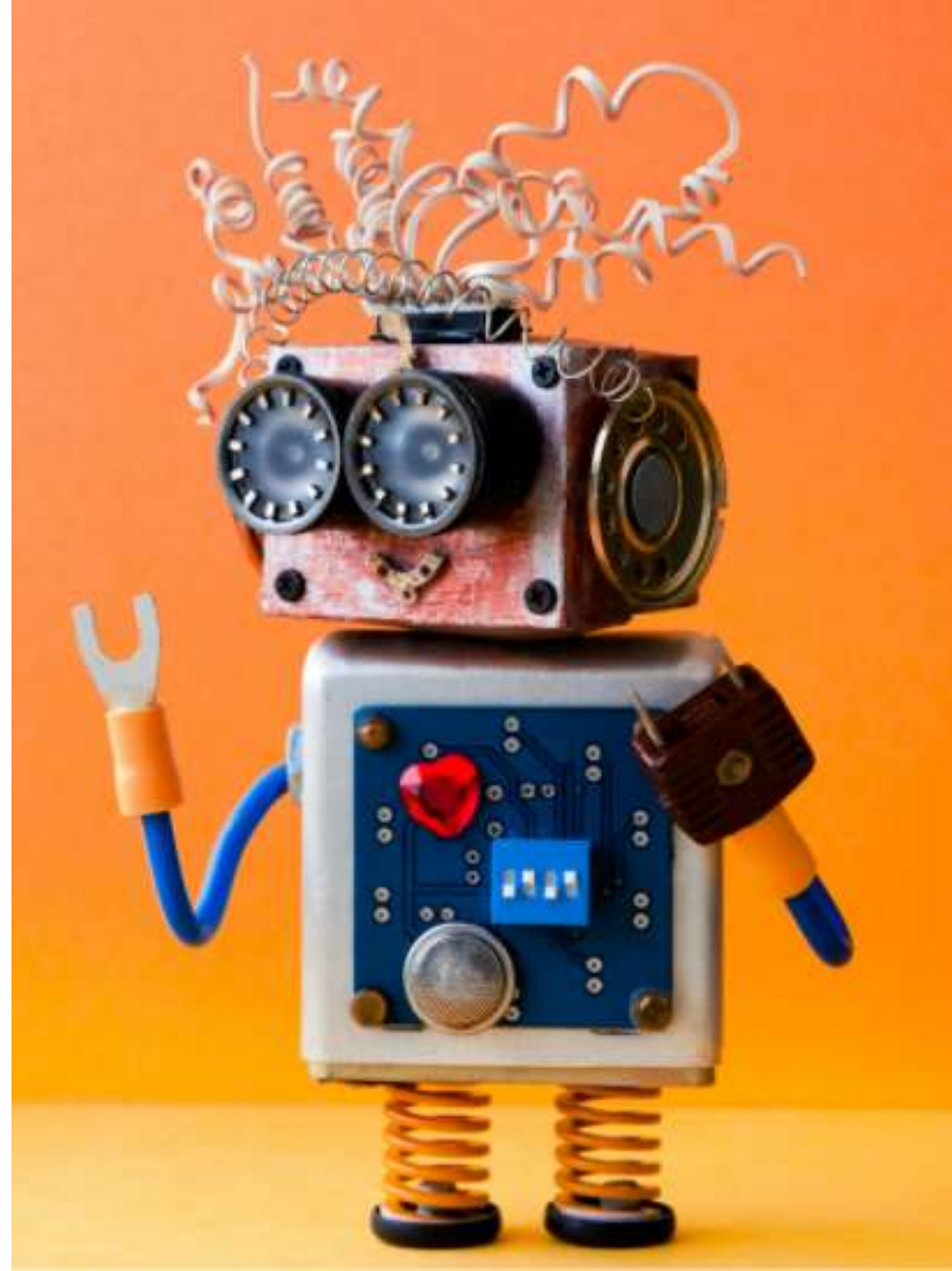
- In computer vision classification-type problems (Supervised ML applications), a manifest file can be useful since it contains data about the inputs (images) and outputs (labels).
- When you run an Amazon SageMaker GroundTruth Job, the output from this process includes a manifest file.
- Manifest files are in JSON lines format where each line is a complete JSON object representing the labeling information for an image.
- Each manifest file itself contains N JSON objects, where N is the number of images that are used in this dataset.

```
{ "source-ref": "s3://aws-ml-engineer/diabetic-retinopathy/train/severe/c3cd0200df79.png", "auto-label": 1, "auto-label-metadata": { "confidence": 1, "job-name": "labeling-job/auto-label", "class-name": "severe", "human-annotated": "yes", "creation-date": "2017-03-01", "type": "groundtruth/image-classification" } }  
{ "source-ref": "s3://aws-ml-engineer/diabetic-retinopathy/train/severe/913490237ad4.png", "auto-label": 1, "auto-label-metadata": { "confidence": 1, "job-name": "labeling-job/auto-label", "class-name": "severe", "human-annotated": "yes", "creation-date": "2017-03-01", "type": "groundtruth/image-classification" } }  
{ "source-ref": "s3://aws-ml-engineer/diabetic-retinopathy/train/severe/a80dab8eddf4.png", "auto-label": 1, "auto-label-metadata": { "confidence": 1, "job-name": "labeling-job/auto-label", "class-name": "severe", "human-annotated": "yes", "creation-date": "2017-03-01", "type": "groundtruth/image-classification" } }  
{ "source-ref": "s3://aws-ml-engineer/diabetic-retinopathy/train/severe/910bfd38e2f5.png", "auto-label": 1, "auto-label-metadata": { "confidence": 1, "job-name": "labeling-job/auto-label", "class-name": "severe", "human-annotated": "yes", "creation-date": "2017-03-01", "type": "groundtruth/image-classification" } }  
{ "source-ref": "s3://aws-ml-engineer/diabetic-retinopathy/train/severe/f6f433f3306f.png", "auto-label": 1, "auto-label-metadata": { "confidence": 1, "job-name": "labeling-job/auto-label", "class-name": "severe", "human-annotated": "yes", "creation-date": "2017-03-01", "type": "groundtruth/image-classification" } }  
{ "source-ref": "s3://aws-ml-engineer/diabetic-retinopathy/train/healthy/3b2b91590590.png", "auto-label": 1, "auto-label-metadata": { "confidence": 1, "job-name": "labeling-job/auto-label", "class-name": "healthy", "human-annotated": "yes", "creation-date": "2017-03-01", "type": "groundtruth/image-classification" } }
```

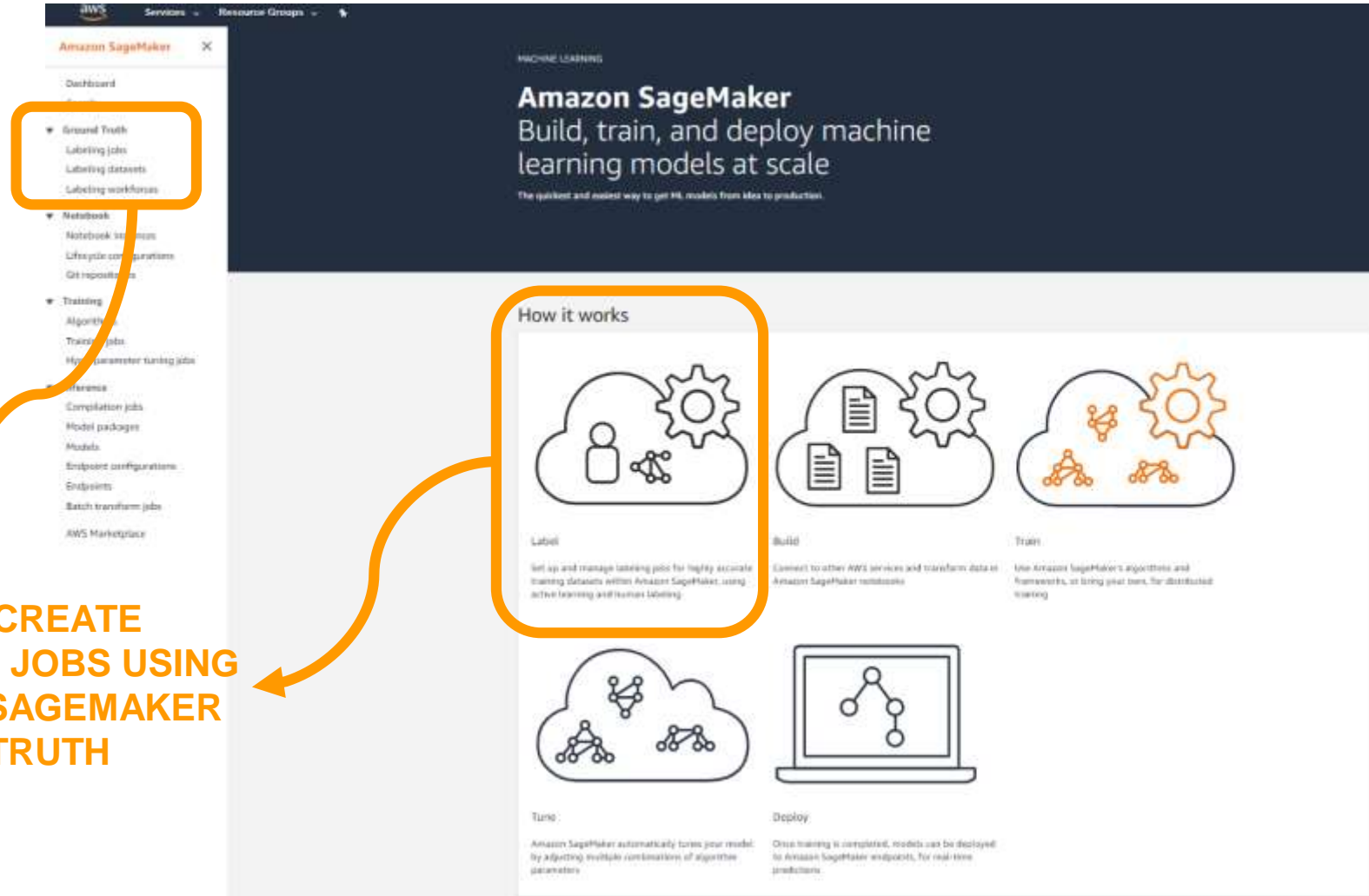


# DATA LABELING IN SAGEMAKER GROUNDTRUTH DEMO – PART 1

---



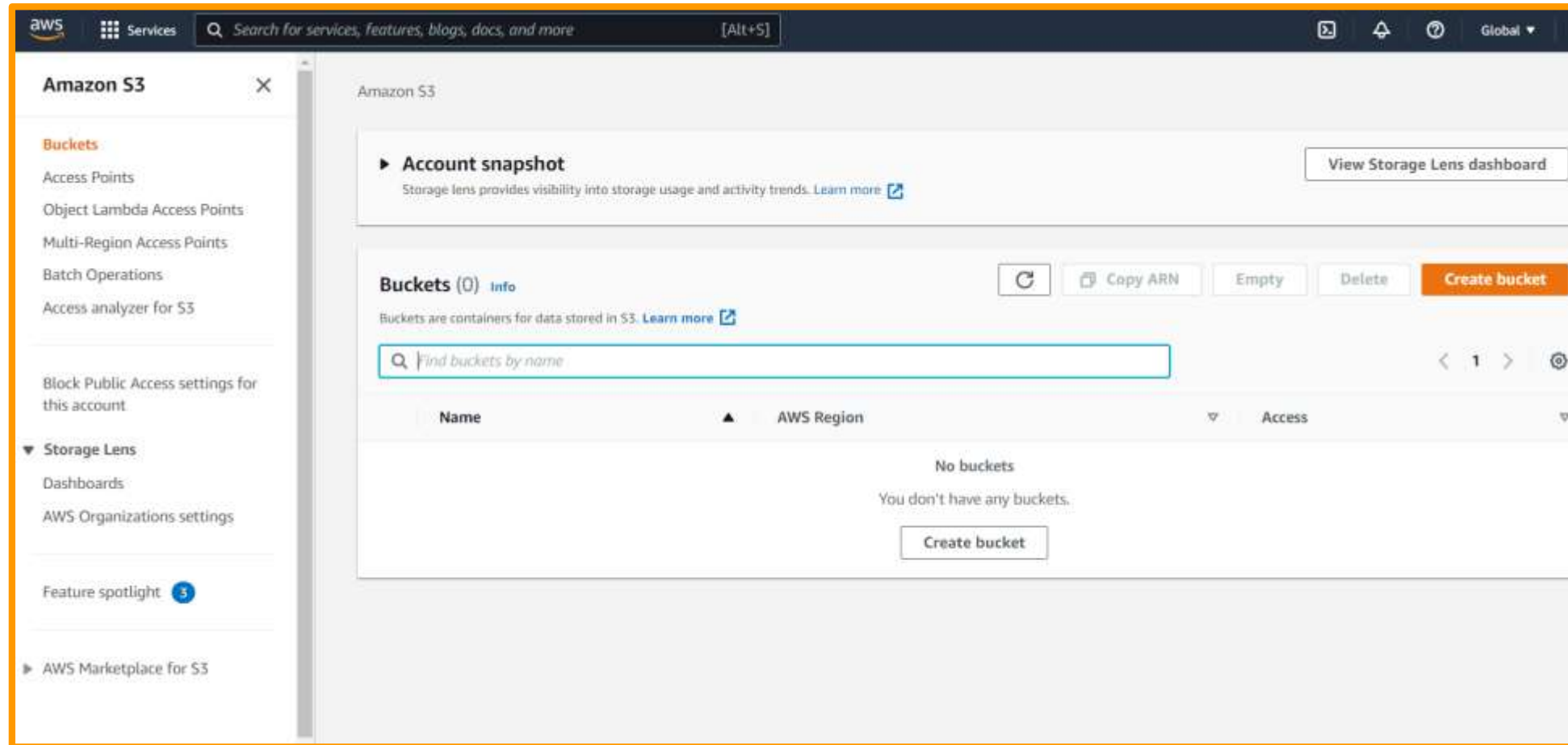
# HOW TO OBTAIN LABELED DATA USING AWS? SAGEMAKER GROUNDTRUTH



<https://aws.amazon.com/sagemaker/groundtruth/>

# SAGEMAKER GOUNDTRUTH DEMO

UPLOAD IMAGES TO S3, GO TO S3 AND  
CLICK ON CREATE BUCKET



# SAGEMAKER GOUNDTRUTH DEMO

GIVE IT NAME AND CLICK CREATE BUCKET  
THEN UPLOAD IMAGES TO S3

**Amazon S3** X

**Buckets**

- Access Points
- Object Lambda Access Points
- Multi-Region Access Points
- Batch Operations
- Access analyzer for S3

Block Public Access settings for this account

▼ **Storage Lens**

- Dashboards
- AWS Organizations settings

Feature spotlight 3

► AWS Marketplace for S3

Amazon S3 > Create bucket

## Create bucket [Info](#)

Buckets are containers for data stored in S3. [Learn more](#)

### General configuration

Bucket name

fashion-labeling

Bucket name must be unique and must not contain spaces or uppercase letters. [See rules for bucket naming](#)

AWS Region

US East (N. Virginia) us-east-1

Copy settings from existing bucket - optional

Only the bucket settings in the following configuration are copied.

Choose bucket

### Object Ownership [Info](#)

Control ownership of objects written to this bucket from other AWS accounts and the use of access control lists (ACLs). Object ownership determines who can specify access to objects.

☒ **ACLs disabled (recommended)**

All objects in this bucket are owned by this account. Access to this bucket and its objects is specified using only policies.

☐ **ACLs enabled**

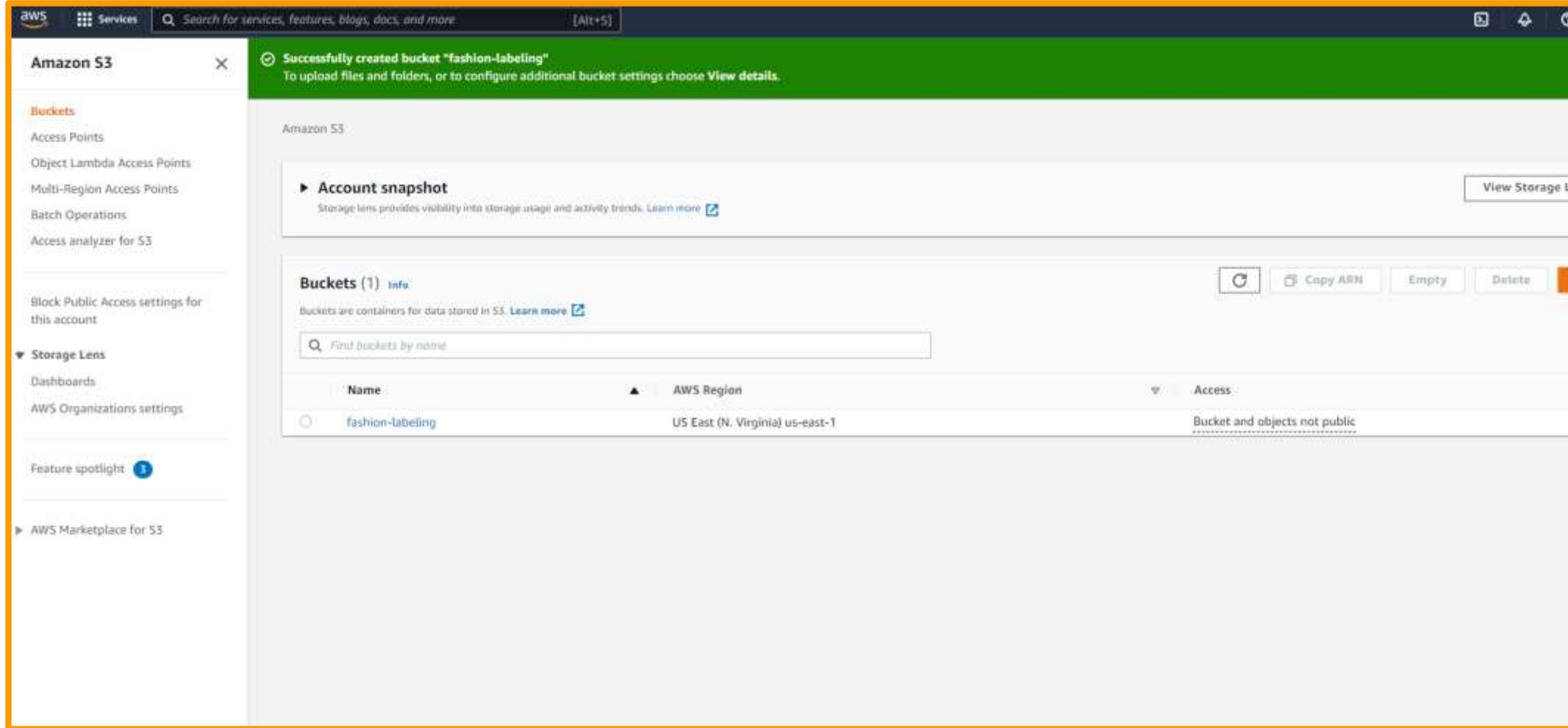
Objects in this bucket can be owned by other AWS accounts. Access to this bucket and its objects can be specified using ACLs.

Object Ownership

Bucket owner enforced

# SAGEMAKER GOUNDTRUTH DEMO

CLICK ON THE NEWLY CREATED BUCKET  
AND UPLOAD THE IMAGES



# SAGEMAKER GOUNDTRUTH DEMO

## CLICK UPLOAD

The screenshot shows the Amazon S3 console interface for the 'fashion-labeling' bucket. The page is titled 'Upload' and includes an 'Info' link. A message instructs users to add files and folders for upload, with a note that files larger than 160GB require the AWS CLI, SDK, or REST API. A dashed box contains the instruction: 'Drag and drop files and folders you want to upload here, or choose **Add files**, or **Add folders**.' Below this, a section titled 'Files and folders (20 Total, 348.2 KB)' shows a list of 10 files. Each file has a checkbox, name, folder path, type, and size. At the bottom, the 'Destination' section shows the path 's3://fashion-labeling'.

Amazon S3 > fashion-labeling > Upload

### Upload [Info](#)

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. [Learn more](#)

Drag and drop files and folders you want to upload here, or choose **Add files**, or **Add folders**.

**Files and folders (20 Total, 348.2 KB)** [Remove](#) [Add files](#) [Add folder](#)

All files and folders in this table will be uploaded.

<input type="checkbox"/>	Name	Folder	Type	Size
<input type="checkbox"/>	001.jpg	-	image/jpeg	20.4 KB
<input type="checkbox"/>	002.jpg	-	image/jpeg	20.4 KB
<input type="checkbox"/>	003.jpg	-	image/jpeg	23.6 KB
<input type="checkbox"/>	004.jpg	-	image/jpeg	23.5 KB
<input type="checkbox"/>	005.jpg	-	image/jpeg	15.3 KB
<input type="checkbox"/>	006.jpg	-	image/jpeg	18.5 KB
<input type="checkbox"/>	007.jpg	-	image/jpeg	812.0 B
<input type="checkbox"/>	008.jpg	-	image/jpeg	11.5 KB
<input type="checkbox"/>	009.jpg	-	image/jpeg	17.8 KB
<input type="checkbox"/>	010.jpg	-	image/jpeg	19.8 KB

**Destination**

Destination  
s3://fashion-labeling



# SAGEMAKER GOUNDTRUTH DEMO

IMAGES ARE NOW UPLOADED TO S3

The screenshot displays the AWS S3 console interface. At the top, a green banner indicates "Upload succeeded" with a link to "View details below." Below this, a summary section shows the destination as "s3://fashion-labeling" and the upload status as "Succeeded" with a green checkmark icon, reporting "20 files, 348.2 KB (100.00%)". A "Failed" section shows "0 files, 0 B (0%)". The "Files and folders" tab is selected, showing a list of 20 files. The table includes columns for Name, Folder, Type, Size, Status, and Error. The files are named 001.jpg through 010.jpg, all of type image/jpeg, and all have a status of "Succeeded".

**Summary**

Destination: [s3://fashion-labeling](#)

Succeeded: 20 files, 348.2 KB (100.00%)

Failed: 0 files, 0 B (0%)

**Files and folders (20 Total, 348.2 KB)**

Find by name

Name	Folder	Type	Size	Status	Error
<a href="#">001.jpg</a>	-	image/jpeg	20.4 KB	Succeeded	-
<a href="#">002.jpg</a>	-	image/jpeg	20.4 KB	Succeeded	-
<a href="#">003.jpg</a>	-	image/jpeg	23.6 KB	Succeeded	-
<a href="#">004.jpg</a>	-	image/jpeg	23.5 KB	Succeeded	-
<a href="#">005.jpg</a>	-	image/jpeg	15.3 KB	Succeeded	-
<a href="#">006.jpg</a>	-	image/jpeg	18.5 KB	Succeeded	-
<a href="#">007.jpg</a>	-	image/jpeg	812.0 B	Succeeded	-
<a href="#">008.jpg</a>	-	image/jpeg	11.5 KB	Succeeded	-
<a href="#">009.jpg</a>	-	image/jpeg	17.8 KB	Succeeded	-
<a href="#">010.jpg</a>	-	image/jpeg	19.8 KB	Succeeded	-

# SAGEMAKER GOUNDTRUTH DEMO

20 IMAGES BELONGS TO 4 CLASSES  
(BALANCED DATASET)

BAGS



001



002



003



004



005

EYEWEAR



006



007



008



009



010

FLIPFLOPS



011



012



013



014



015

WATCHES



016



017



018



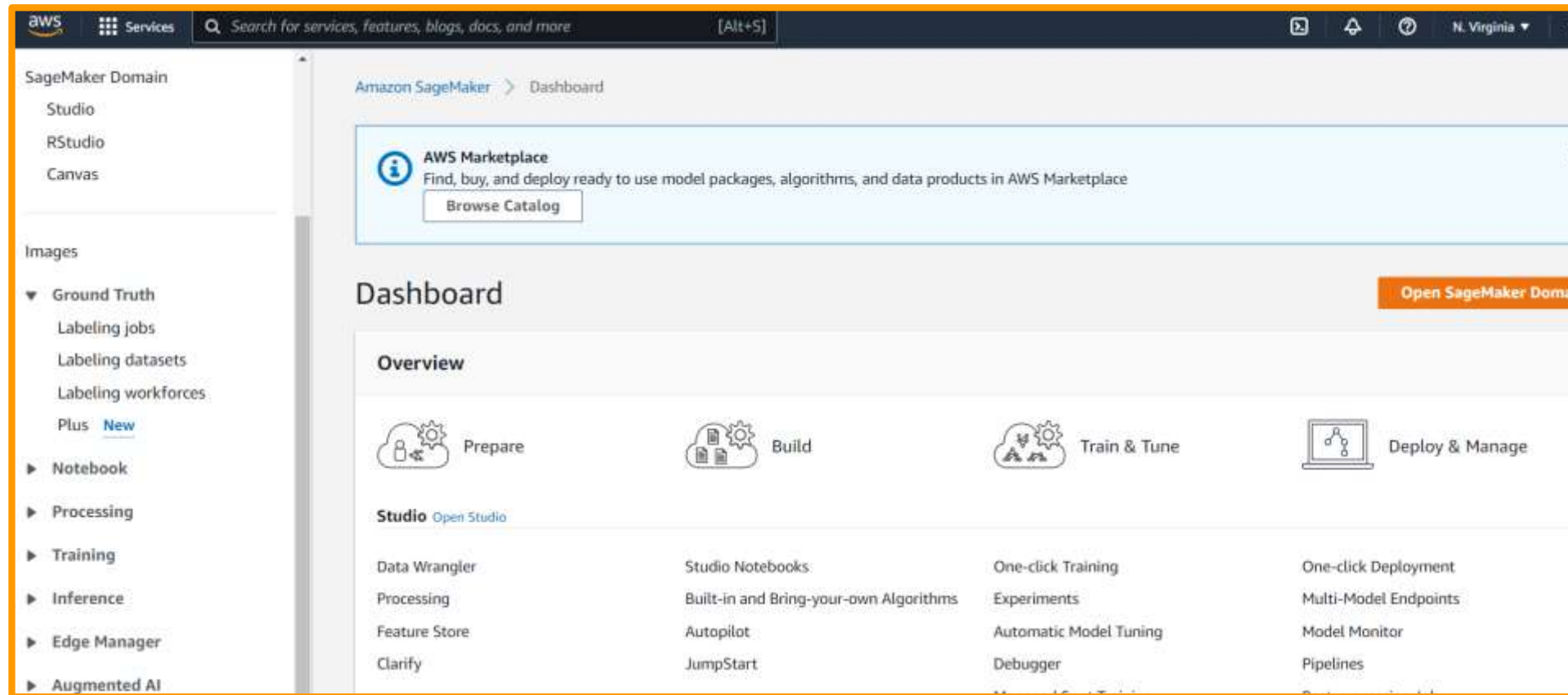
019



020

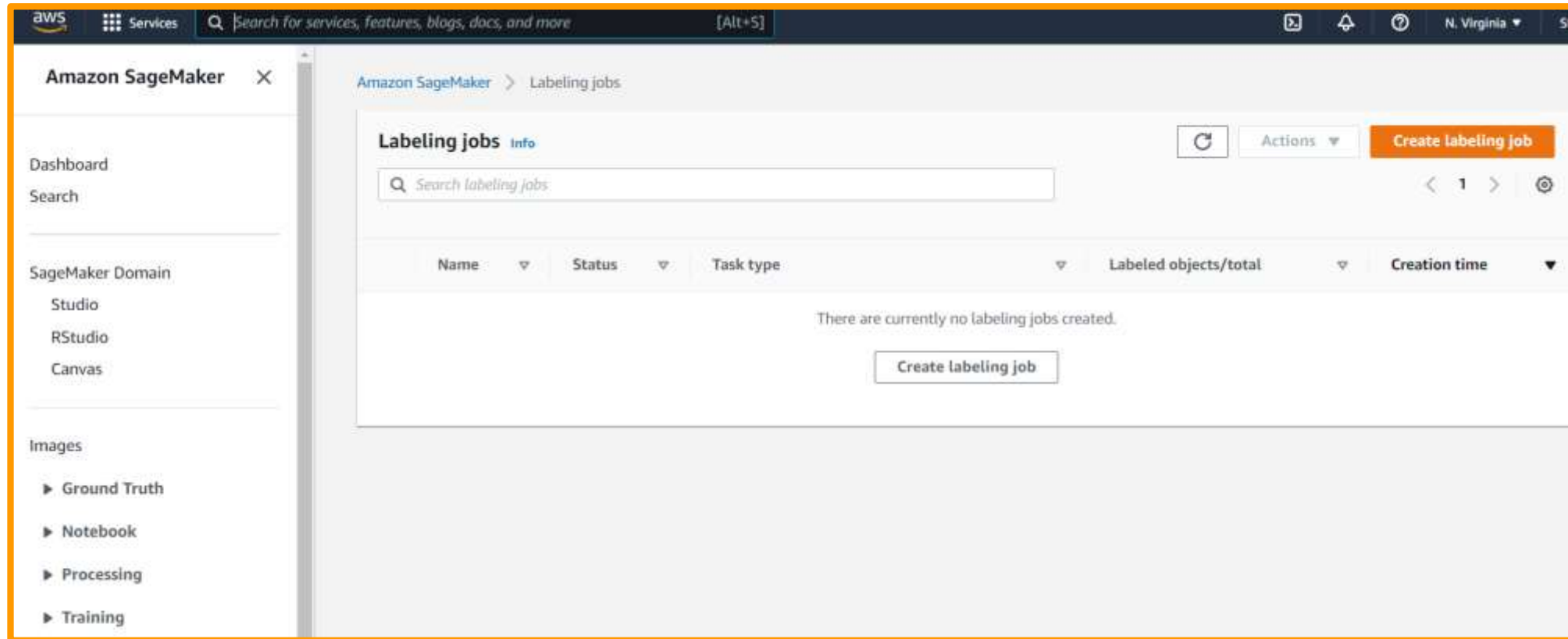
# SAGEMAKER GOUNDTRUTH DEMO

OPEN SAGMAKER AND CLICK ON LABELING JOBS



# SAGEMAKER GOUNDTRUTH DEMO

CLICK ON CREATE LABELING JOB



# SAGEMAKER GOUNDTRUTH DEMO

GIVE A NAME TO THE JOB AND  
CLICK ON BROWSE S3

aws Services Search for services, features, blogs, docs, and more [Alt+S] N. Virginia

Amazon SageMaker X

Dashboard  
Search

SageMaker Domain

- Studio
- RStudio
- Canvas

Images

- Ground Truth
- Notebook
- Processing
- Training
- Inference
- Edge Manager

Step 1  
Specify job details

Step 2  
Select workers and configure tool

## Specify job details

### Job overview

Job name

my-first-labeling-job

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.

Data setup

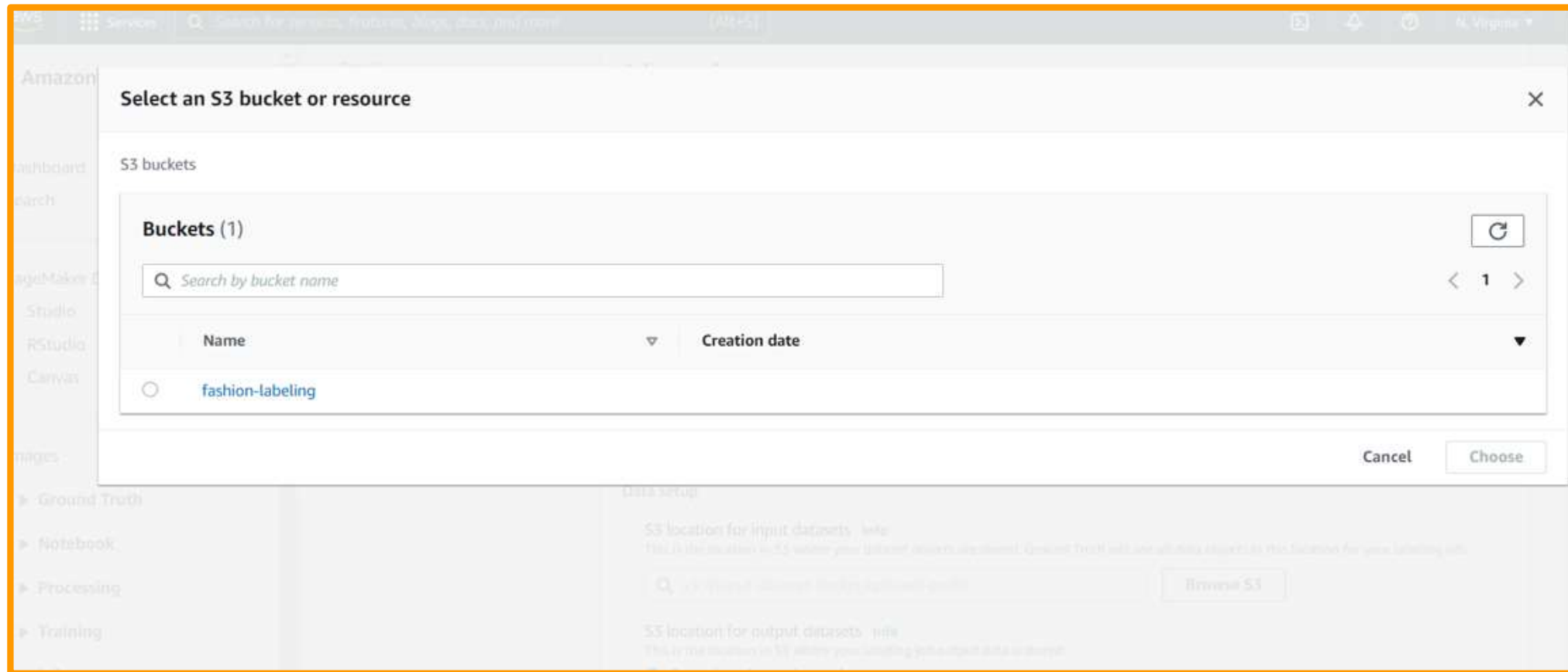
S3 location for input datasets [Info](#)

This is the location in S3 where your dataset objects are stored. Ground Truth will use all data objects in this location for your labeling job.

[Browse S3](#)

# SAGEMAKER GROUNDTRUTH DEMO

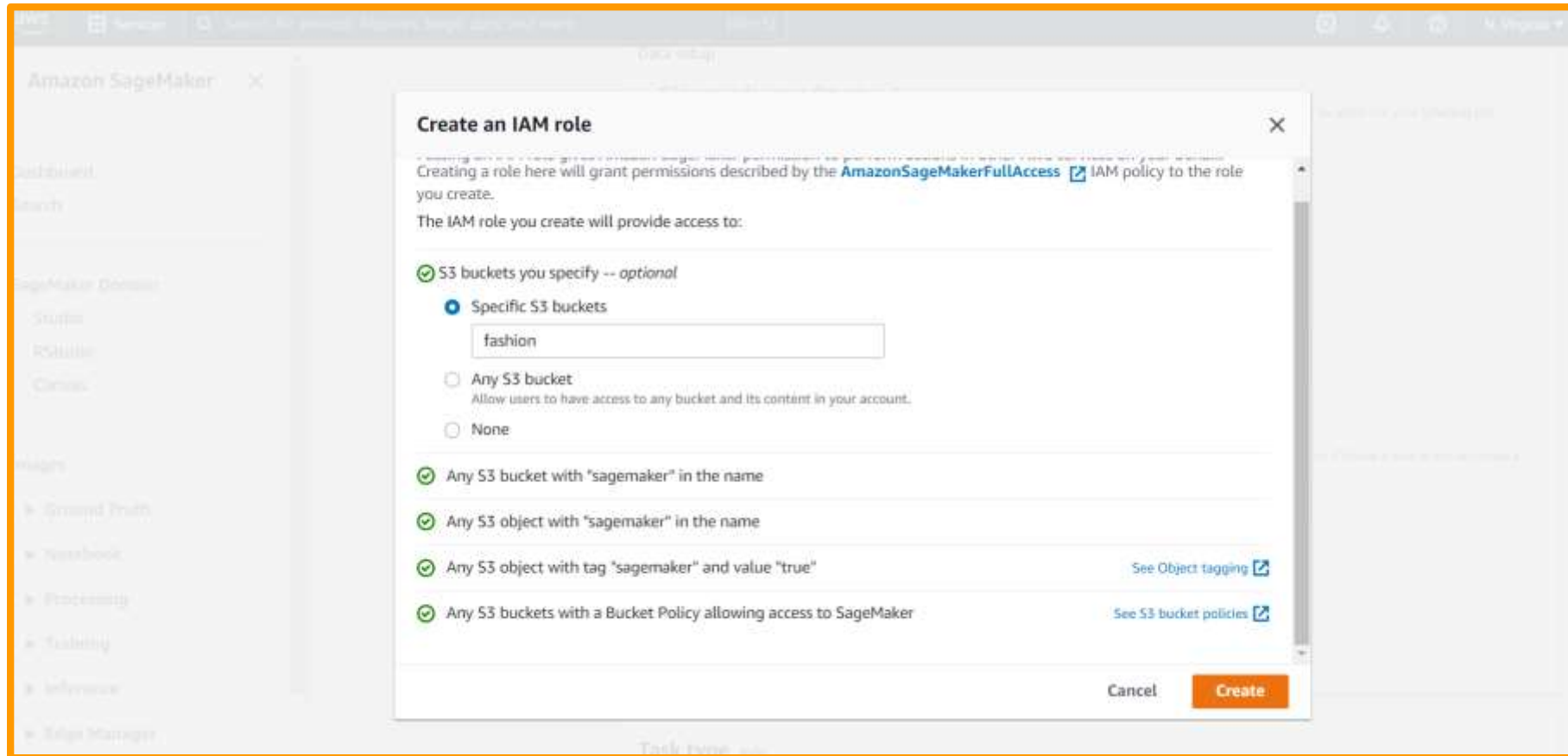
CLICK ON BROWSE S3 AND  
POINT TO THE DATASET





# SAGEMAKER GOUNDTRUTH DEMO

CREATE A NEW IAM ROLE AND GIVE  
ACCESS TO A SPECIFIC BUCKET





# SAGEMAKER GOUNDTRUTH DEMO

DON'T FORGET TO CLICK ON  
“COMPLETE DATA SETUP”

### Job overview

Job name

my-first-labeling-job

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

Data setup

S3 location for input datasets [Info](#)

This is the location in S3 where your dataset objects are stored. Ground Truth will use all data objects in this location for your labeling job.

Q s3://fashion-labeling/fashion/ X

Browse S3

S3 location for output datasets [Info](#)

This is the location in S3 where your labeling job output data is stored.

☒ Same location as input dataset

☐ Specify a new location

Data type

Image

Supported formats are .jpg, .jpeg, and .png.

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.

AmazonSageMaker-ExecutionRole-20220203T055170

Use this button to process and complete your input data setup.

Complete data setup

# SAGEMAKER GOUNDTRUTH DEMO

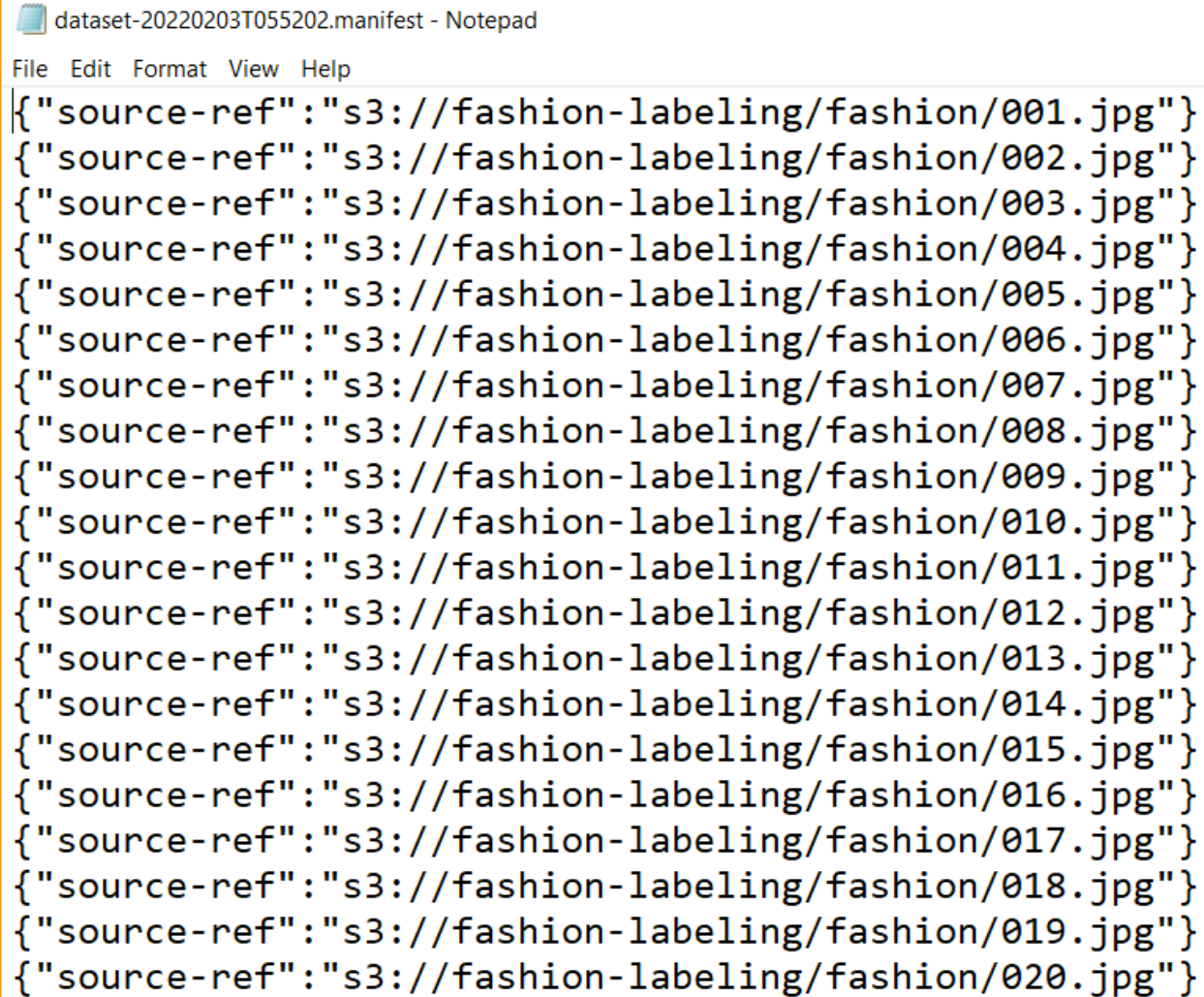
WHEN YOU CLICK ON “COMPLETE DATA  
SETUP”, A MANIFEST FILE WILL BE  
GENERATED IN S3

The screenshot displays the Amazon S3 console interface. On the left, the 'Amazon S3' sidebar is visible with various navigation options. The main area shows a bucket containing 22 objects. The objects are listed in a table with columns for Name, Type, Last modified, Size, and Storage class. The objects include 20 JPEG files (002.jpg to 020.jpg) and two manifest files (dataset-20220203T055202.manifest and dataset-20220203T055202.manifest.json).

Name	Type	Last modified	Size	Storage class
002.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	20.4 KB	Standard
003.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	23.6 KB	Standard
004.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	23.5 KB	Standard
005.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	15.3 KB	Standard
006.jpg	jpg	February 3, 2022, 05:41:03 (UTC-05:00)	18.5 KB	Standard
007.jpg	jpg	February 3, 2022, 05:41:03 (UTC-05:00)	812.0 B	Standard
008.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	11.5 KB	Standard
009.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	17.8 KB	Standard
010.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	19.8 KB	Standard
011.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	12.9 KB	Standard
012.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	20.2 KB	Standard
013.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	18.9 KB	Standard
014.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	13.1 KB	Standard
015.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	38.1 KB	Standard
016.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	17.0 KB	Standard
017.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	1.8 KB	Standard
018.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	15.8 KB	Standard
019.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	19.6 KB	Standard
020.jpg	jpg	February 3, 2022, 05:41:04 (UTC-05:00)	19.0 KB	Standard
dataset-20220203T055202.manifest	manifest	February 3, 2022, 05:52:35 (UTC-05:00)	1.1 KB	Standard
dataset-20220203T055202.manifest.json	json	February 3, 2022, 05:52:35 (UTC-05:00)	83.0 B	Standard

# SAGEMAKER GOUNDTRUTH DEMO

INPUT MANIFEST FILE SHOULD LOOK LIKE THIS

A screenshot of a Notepad window titled "dataset-20220203T055202.manifest - Notepad". The window has a menu bar with "File", "Edit", "Format", "View", and "Help". The main text area contains a list of 20 JSON objects, each with a "source-ref" key and a value representing an S3 path to a fashion image. The paths are sequential, starting from "s3://fashion-labeling/fashion/001.jpg" and ending at "s3://fashion-labeling/fashion/020.jpg".

```
dataset-20220203T055202.manifest - Notepad
File Edit Format View Help
{"source-ref":"s3://fashion-labeling/fashion/001.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/002.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/003.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/004.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/005.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/006.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/007.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/008.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/009.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/010.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/011.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/012.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/013.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/014.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/015.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/016.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/017.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/018.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/019.jpg"}
{"source-ref":"s3://fashion-labeling/fashion/020.jpg"}
```

# SAGEMAKER GOUNDTRUTH DEMO

NOTE THAT YOU CAN EITHER LABEL THE ENTIRE DATASET OR CHOOSE A RANDOM SAMPLE

Use this button to process and complete your input data setup.

**Complete data setup**

✔ Input data connection successful. [View more details](#)

▼ **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

**Encryption key - optional**

If you want Amazon SageMaker to encrypt the output of your labeling job using your own AWS KMS encryption key instead of the default S3 service key, provide its ID or ARN.

*Choose an option* ▼



# SAGEMAKER GOUNDTRUTH DEMO

NOTE THAT ONCE YOU ESTABLISH THE DATA CONNECTION, THE MANIFEST FILE PATH IS UPDATED AS SHOWN BELOW

☒ 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

Data setup

S3 location for input datasets [Info](#)  
This is the location in S3 where your dataset objects are stored. Ground Truth will use all data objects in this location for your labeling job.

Q s3://fashion-labeling/fashion/dataset-20220203T060220.m X

Browse S3

S3 location for output datasets [Info](#)  
This is the location in S3 where your labeling job output data is stored.

☒ Same location as input dataset

☐ Specify a new location

Data type

Image

Supported formats are .jpg, .jpeg, and .png.

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.

AmazonSageMaker-ExecutionRole-20220203T060145

Use this button to process and complete your input data setup.

Complete data setup

✔ Input data connection successful. [View more details](#)

► Additional configuration - optional

Dataset object selection, encryption

# SAGEMAKER GOUNDTRUTH DEMO

## LET'S KICK START A PRIVATE DATA LABELING JOB

The screenshot shows the Amazon SageMaker Ground Truth console. The left sidebar contains navigation links: Dashboard, Search, SageMaker Domain (Studio, RStudio, Canvas), Images (Ground Truth, Notebook, Processing, Training, Inference, Edge Manager, Augmented AI, AWS Marketplace), and a search bar. The main content area is titled 'Create labeling job' and shows 'Step 2: Select workers and configure tool'. The 'Workers' tab is active, displaying three worker types: Amazon Mechanical Turk, Private (selected), and Vendor managed. The 'Private' option is highlighted with a blue border. Below the worker types, there are fields for 'Team name' (set to 'super-labelers'), 'Invite private annotators' (with email 'ryanahmedaly@gmail.com'), 'Task timeout' (0 hours, 5 mins, 0 secs), 'Task expiration time' (10 days, 0 hours, 0 mins, 0 secs), 'Organization' (set to 'Fashion Images Classification'), and 'Contact email' (set to 'ryanahmedaly@gmail.com').

Amazon SageMaker

Search for services, features, blogs, docs, and more

Labeling jobs > Create labeling job

Step 1  
Select job domain

Step 2  
Select workers and configure tool

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  
super-labelers  
Maximum of 32 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.  
ryanahmedaly@gmail.com  
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

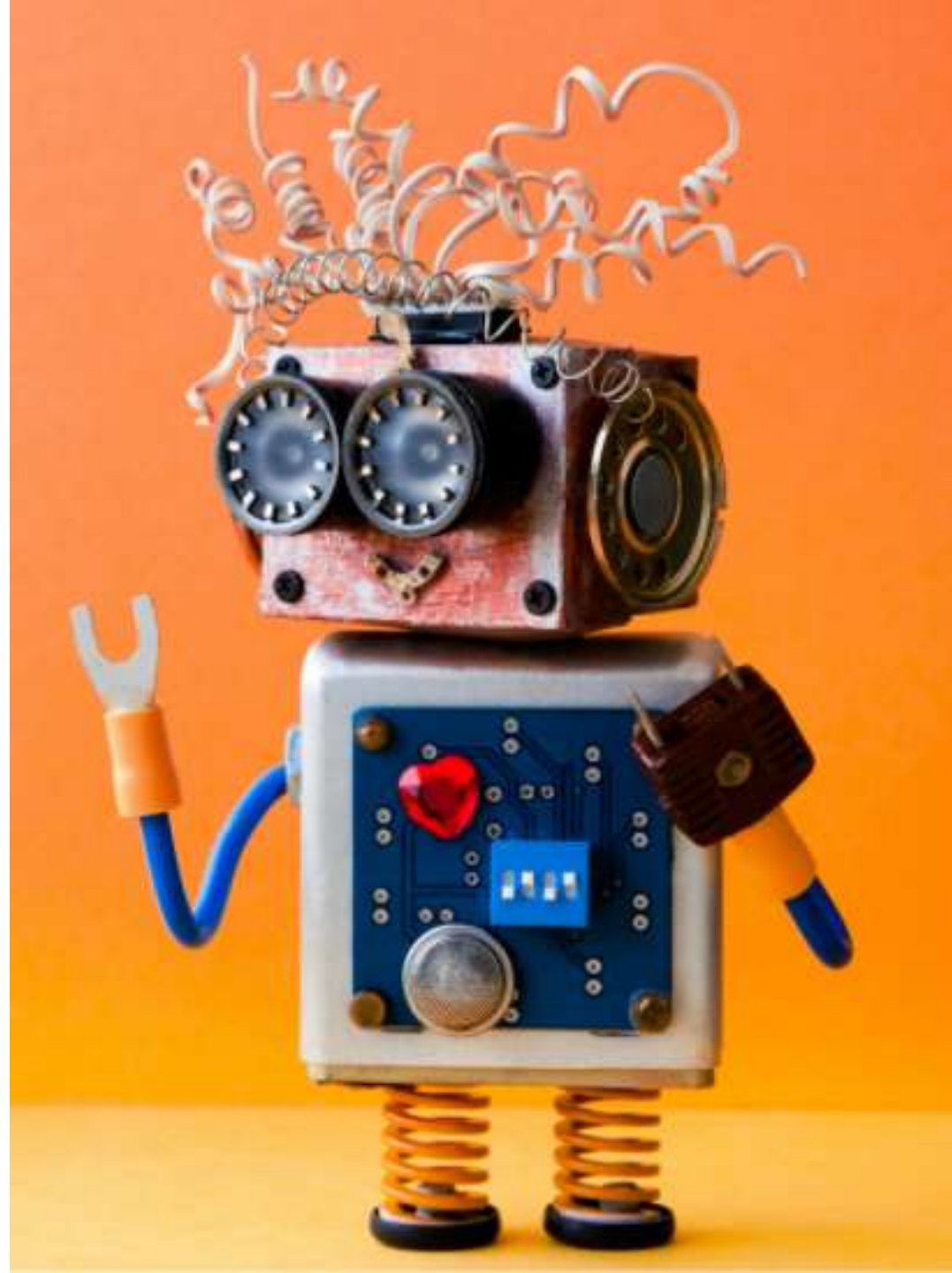
Task expiration time  
The amount of time that a task remains available to workers before expiring. If you want to use values beyond 10 days, contact AWS Support.  
10 days 0 hours 0 mins 0 secs

Organization  
We use this information to customize the worker invitation.  
Fashion Images Classification

Contact email  
Workers can use this to report issues related to the job.  
ryanahmedaly@gmail.com  
Enter one email address only.

# DATA LABELING IN SAGEMAKER GROUNDTRUTH DEMO – PART 2

---



# SAGEMAKER GOUNDTRUTH DEMO

## SPECIFY THE LABELING JOB REQUIREMENTS


☐ 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

Image classification (Single Label) labeling tool [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](#) [🔗](#)

H1 H2 B I A  
🔗 📎

**Good example**  
Here is an example of an eyewear  


**Bad example**  
Enter description of an incorrect label  
  
Add image here

Classify Fashion Images to one of 4 categories



Select an option  
Add up to 30 labels

Bag ×

Eyewear ×

Flipflop ×

Watch ×

Add new label

You can add 26 more labels.

▶ Additional instructions - optional

Cancel

Previous

Create

# SAGEMAKER GOUNDTRUTH DEMO

NOTE THAT THE LABELING JOB HAS  
BEEN SUCCESSFULLY CREATED

The screenshot displays the Amazon SageMaker console interface. At the top, there's a navigation bar with the AWS logo, a 'Services' menu, and a search bar. Below this, a green banner indicates that the labeling job 'my-first-labeling-job' was successfully created. The left sidebar contains navigation links for 'Dashboard', 'Search', 'SageMaker Domain' (with sub-links for Studio, RStudio, and Canvas), and 'Images' (with sub-links for Ground Truth, Notebook, Processing, Training, and Inference). The main content area shows the 'Labeling jobs' page, which includes a search bar and a table of jobs. The table has columns for Name, Status, Task type, and Labeled objects/total. A single job, 'my-first-labeling-job', is listed with a status of 'In progress' and a task type of 'Image Classification (Single Label)'.

aws Services Search for services, features, blogs, docs, and more [Alt+S]

Amazon SageMaker ×

Dashboard  
Search

SageMaker Domain  
Studio  
RStudio  
Canvas

Images  
▶ Ground Truth  
▶ Notebook  
▶ Processing  
▶ Training  
▶ Inference

Labeling job my-first-labeling-job was successfully created.

Amazon SageMaker > Labeling jobs

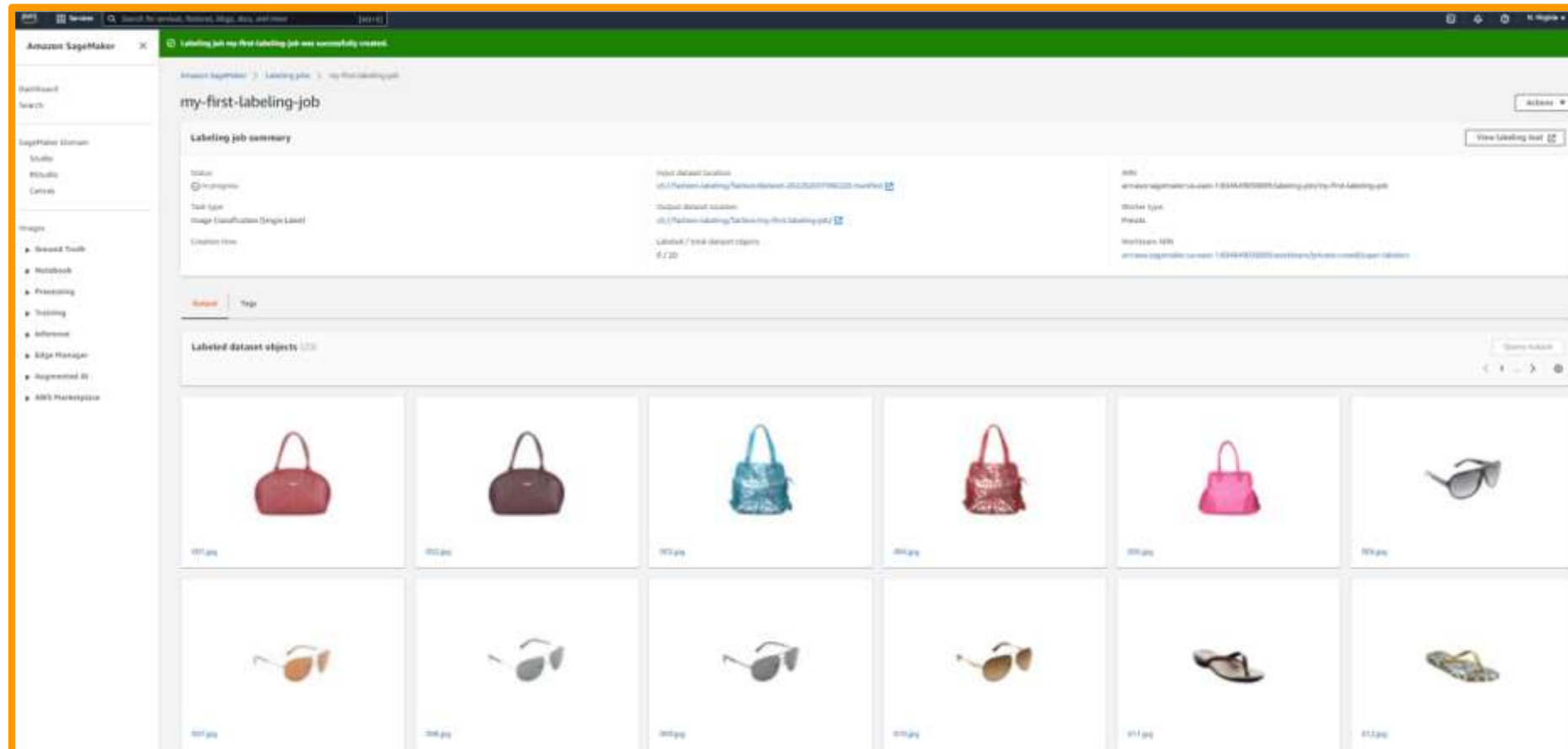
Labeling jobs Info

Search labeling jobs

Name	Status	Task type	Labeled objects/total
my-first-labeling-job	In progress	Image Classification (Single Label)	-

# SAGEMAKER GOUNDTRUTH DEMO

CLICK ON THE LABELING JOB TO CHECK THE STATUS AND SEE IF IT MAKES SENSE!

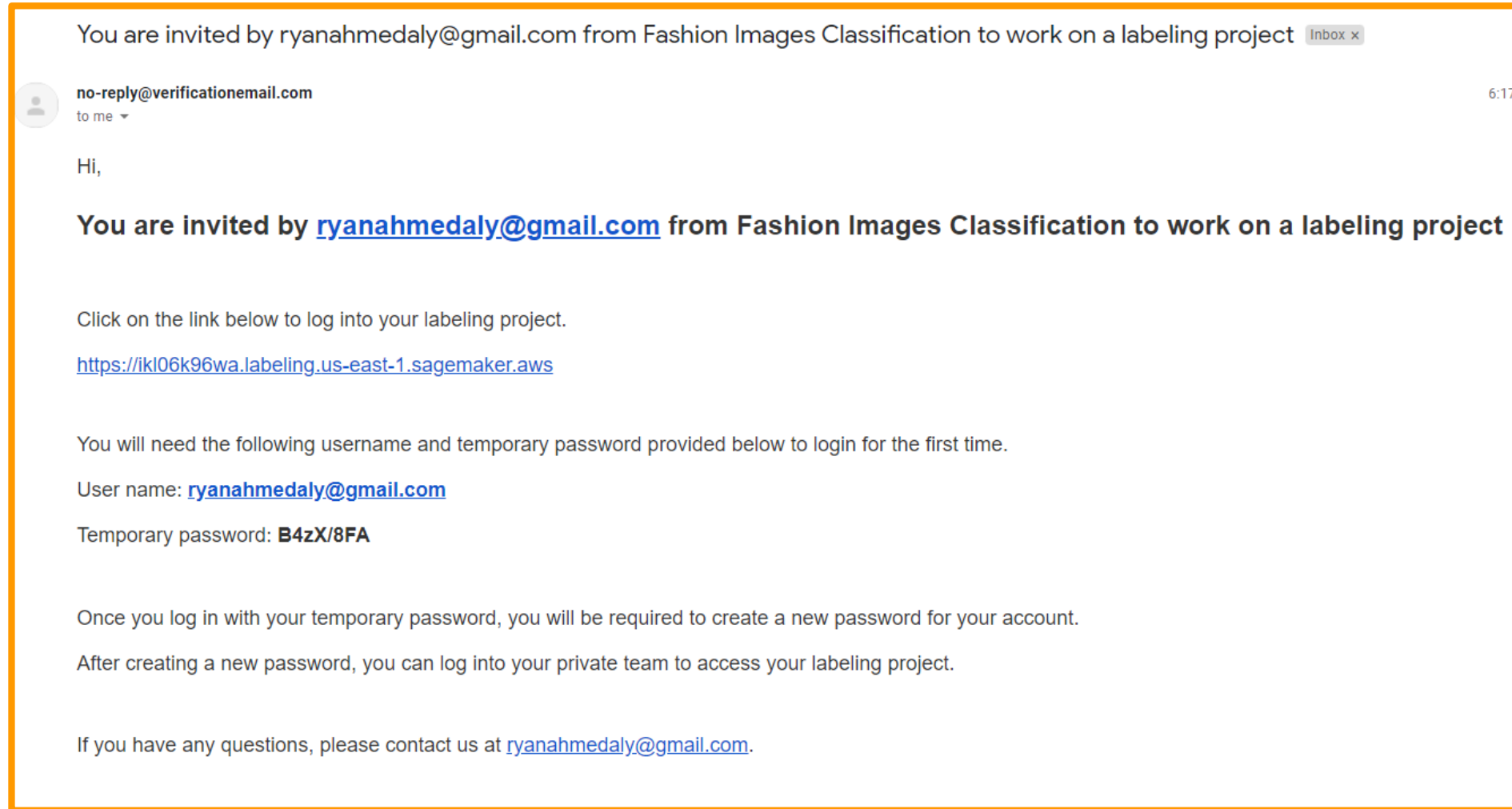


The screenshot displays the Amazon SageMaker console interface. On the left, a navigation sidebar includes links to Dashboard, Search, SageMaker Studio, Studio, StudioLab, and StudioLab. The main content area is titled 'my-first-labeling-job' and features a 'Labeling job summary' section. This summary includes the job name, status (In progress), task type (Image Classification (Single Label)), and creation time (6/2/20). It also lists the input dataset location, output dataset location, and the SageMaker Labeling tool version (1.0.0). A 'View labeling tool' button is present. Below the summary, a 'Labeled dataset objects' section shows a grid of 12 image thumbnails, each with a filename below it. The images include various handbags and sunglasses. The console interface is framed by an orange border.



# SAGEMAKER GOUNDTRUTH DEMO

## THIS IS THE INVITE THAT LABELERS WILL RECEIVE IN THEIR EMAILS



# SAGEMAKER GOUNDTRUTH DEMO

THIS IS WHAT THE TASK LOOK LIKE!

Hello, ryanahmedaly@gmail.com

Customer ID: 6946...

Task description: Categorize images i...

Task time: 0:07 of 5 Min

Decline task


Release task

Stop and resume later

Instructions

Shortcuts

Classify Fashion Images to one of 4 categories



Select an option

Bag	1
Eyewear	2
Flipflop	3
Watch	4



Submit

# SAGEMAKER GOUNDTRUTH DEMO

THE LABELING JOB IS NOW  
COMPLETE, YOU CAN PROCEED WITH  
EXPLORING THE LABELS

Amazon SageMaker

my-first-labeling-job

Labeling job summary

Status: Complete

Task type: Image Classification (Single Label)

Creation time:

Input dataset location: [s3://fashion-labeling/fashion/dataset-2022/01/000220.marshfirst](#)

Output dataset location: [s3://fashion-labeling/fashion/my-first-labeling-job/](#)

Labeled / total dataset objects: 20 / 20

ARN: [arn:aws:sagemaker:us-east-1:5045490158895:Labeling-job/my-first-labeling-job](#)

Worker type: Private

Work team ARN: [arn:aws:sagemaker:us-east-1:5045490158895:workteam/private-crowd/super-labelers](#)

View labeling tool

Output | Tags

Labeled dataset objects (20)

Query output

001.jpg  
Label: Bag

002.jpg  
Label: Bag

003.jpg  
Label: Bag

004.jpg  
Label: Bag

005.jpg  
Label: Bag

006.jpg  
Label: Eyewear

007.jpg  
Label: Eyewear

008.jpg  
Label: Eyewear

009.jpg  
Label: Eyewear

010.jpg  
Label: Eyewear

011.jpg  
Label: Flipflop

012.jpg  
Label: Flipflop

# SAGEMAKER GOUNDTRUTH DEMO

GO TO S3 AND CHECK OUT THE  
MANIFEST OUTPUT FILE

Amazon S3

Buckets

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

Access analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

AWS Organizations settings

Feature spotlight

AWS Marketplace for S3

We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

Amazon S3 > fashion-labeling > fashion/ > my-first-labeling-job/ > manifests/ > output/

output/

Objects

Properties

Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Copy S3 URI

Copy URL

Download

Open


Delete

Actions

Create folder

Upload

Find objects by prefix

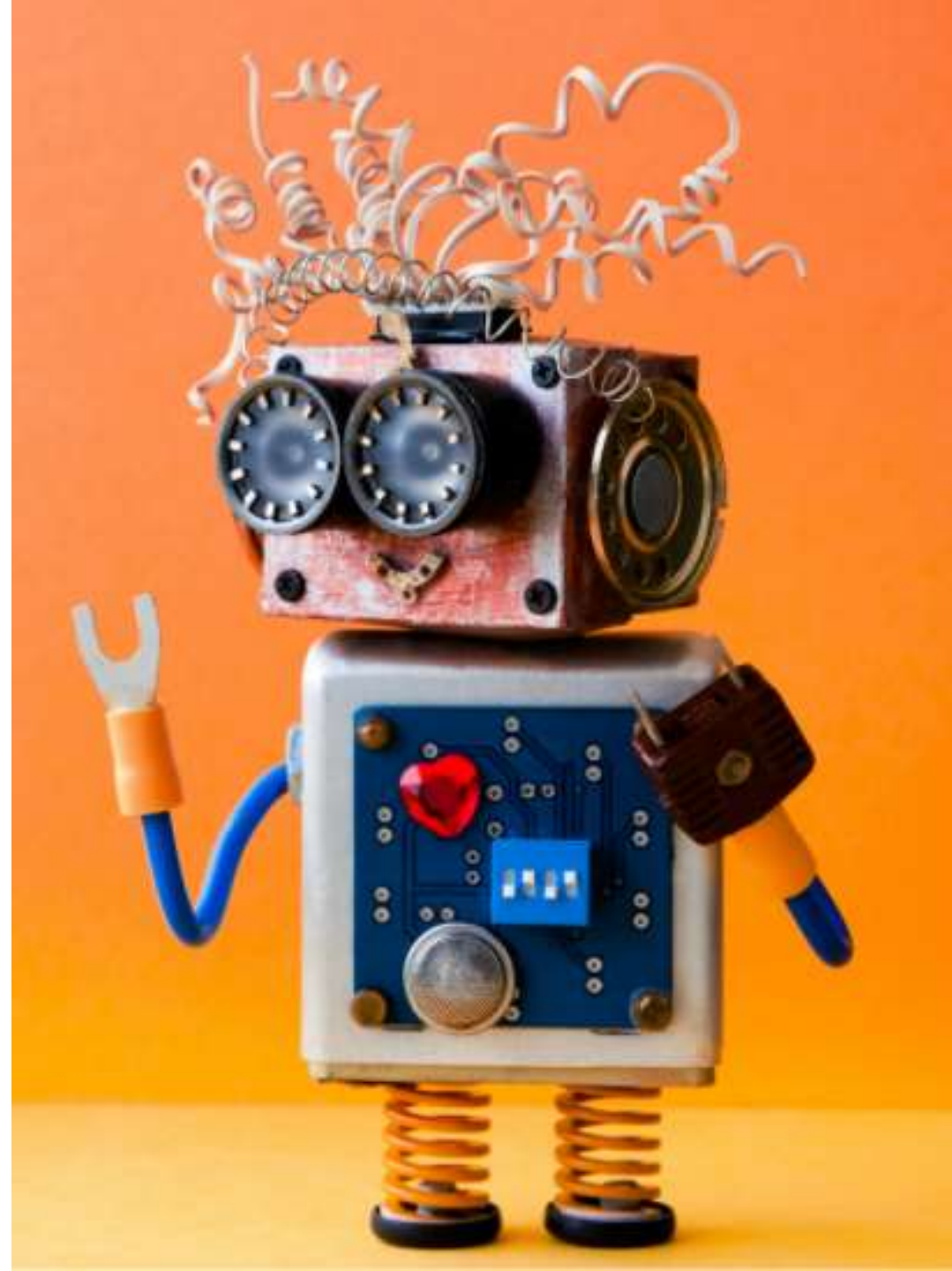
<input type="checkbox"/>	Name	Type	Size	Storage class
<input type="checkbox"/>	 output.manifest	manifest	6.1 KB	Standard

CHECK OUT THE OUTPUT MANIFEST FILE. NOTE THAT CONFIDENCE IS SHOWN ZERO SINCE WE DON'T HAVE MULTIPLE LABELERS DOING THE SAME JOB AND WE DIDN'T LEVERAGE THE AUTOLABEL

[illegible]

# FINAL END-OF-DAY CAPSTONE PROJECT

---





# FINAL CAPSTONE PROJECT

1. Using the Traffic signs datasets included in the course package.
2. Create a labeling job using Amazon SageMaker GroundTruth
3. Review the input manifest file
4. Perform the labeling job
5. Review the output manifest file and ensure that the labeling job was successful



1



2



3



4



5



6



7



8

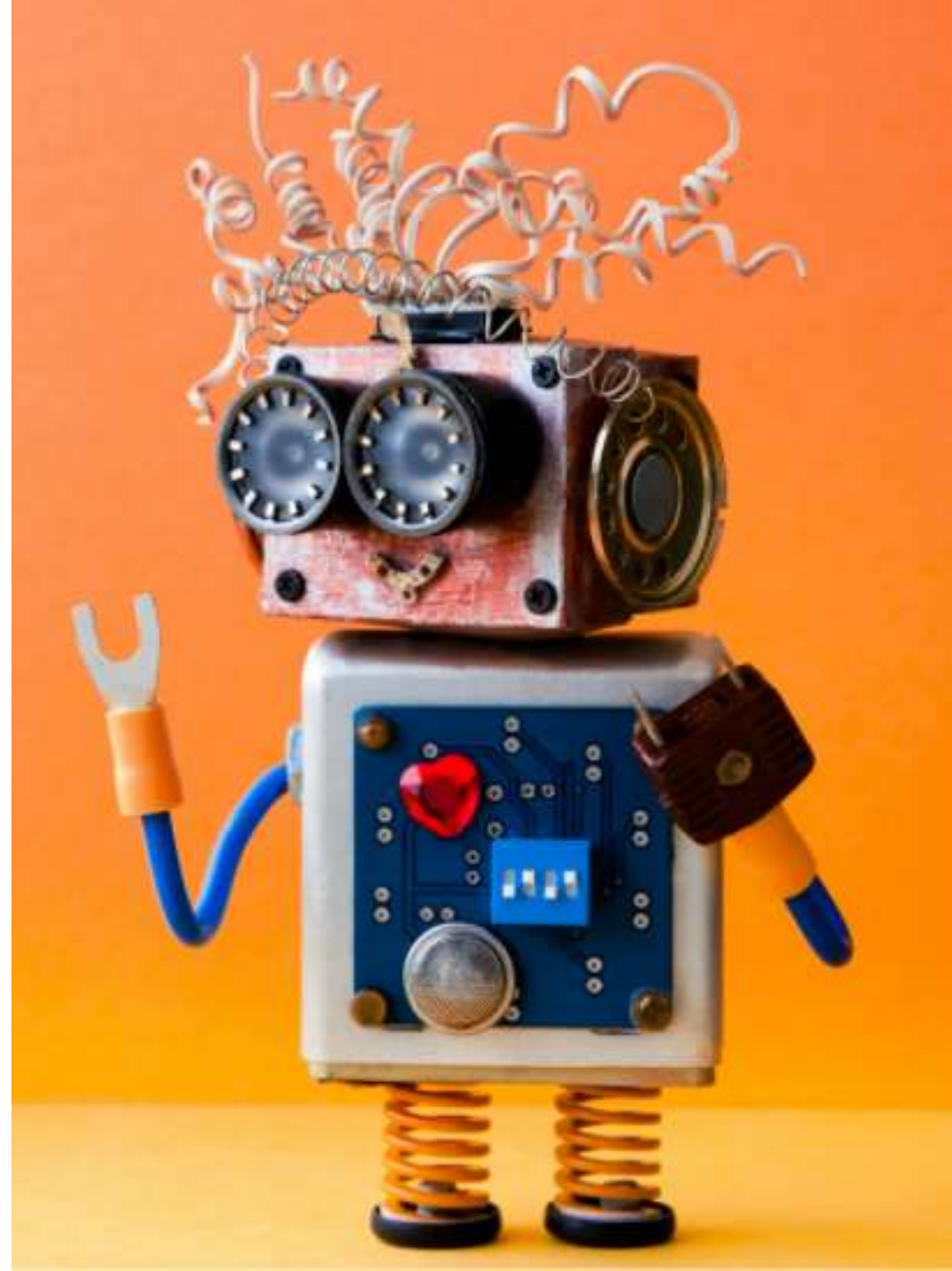


9

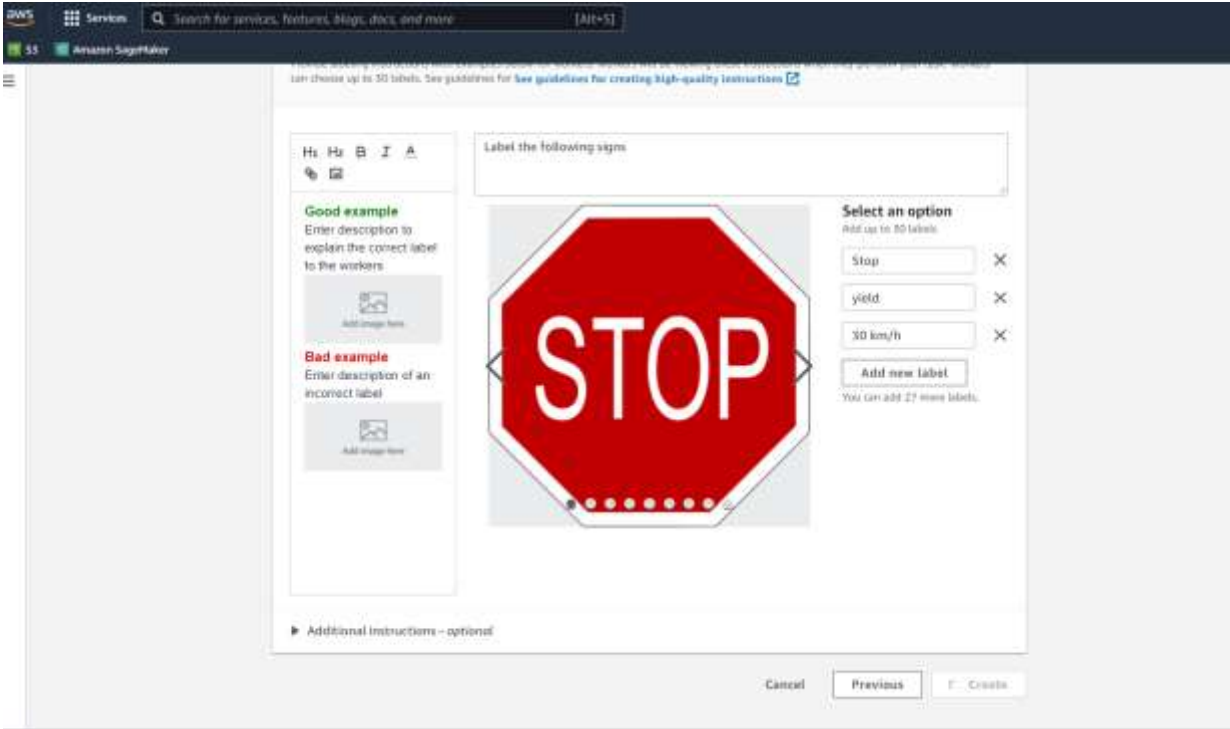
AWS SageMaker Pricing Examples: [https://aws.amazon.com/sagemaker/data-labeling/pricing/?nc=sn&loc=3&refid=ps\\_a134p0000078pqxaae&trkcampaign=acq\\_paid\\_search\\_brand](https://aws.amazon.com/sagemaker/data-labeling/pricing/?nc=sn&loc=3&refid=ps_a134p0000078pqxaae&trkcampaign=acq_paid_search_brand)

# FINAL END-OF-DAY CAPSTONE PROJECT SOLUTION

---



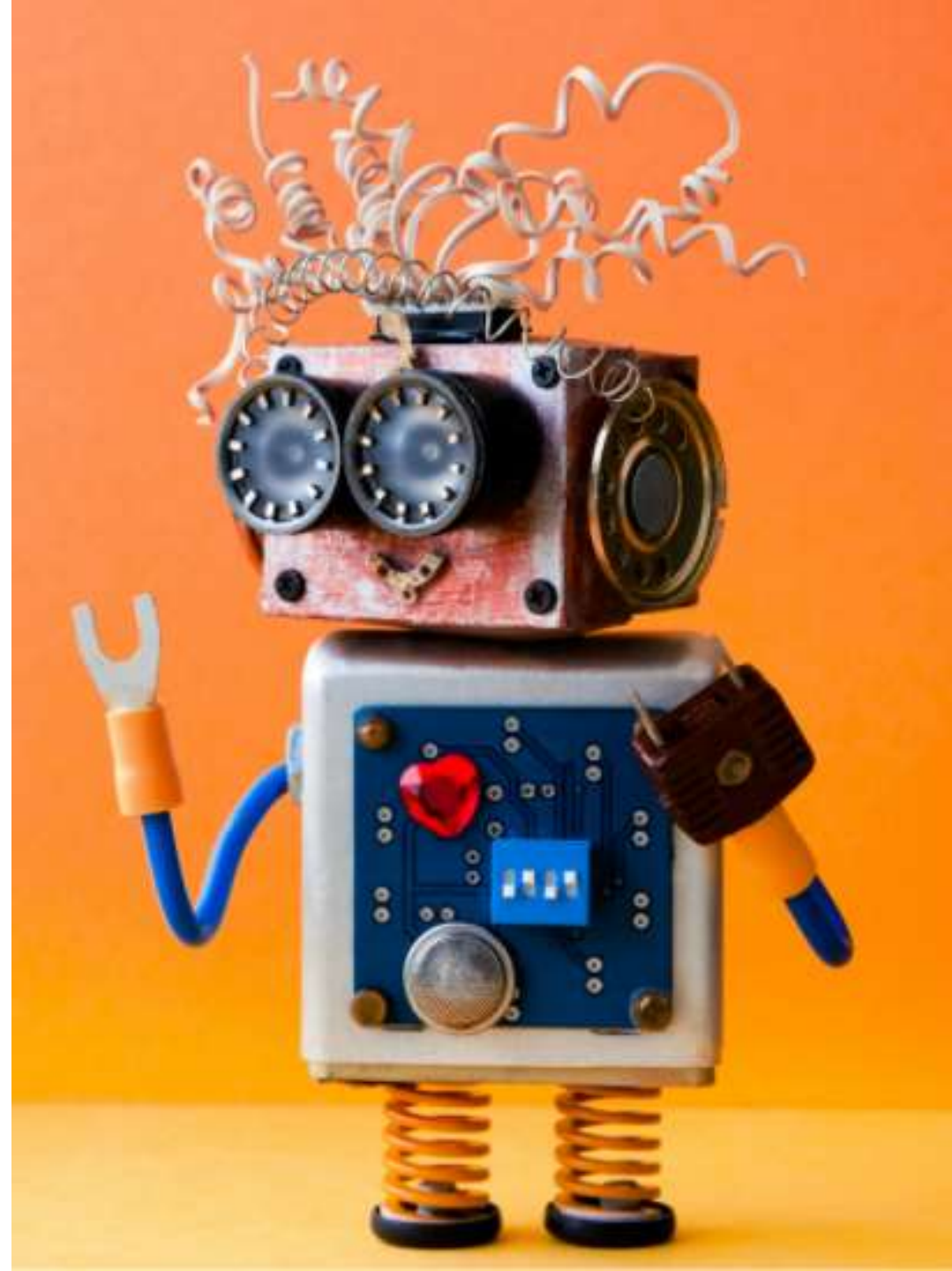
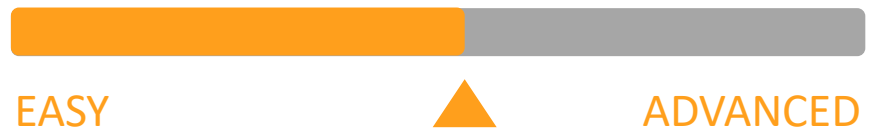
# PROJECT SOLUTION



```
File Edit View
[{"datasetObjectId": "4", "consolidatedAnnotation": {"content": {"labeling-job-project": 1, "labeling-job-project-metadata": {"class-name": "yield", "job-name": "labeling-job/labeling-job-project", "confidence": 0.0, "type": "groundtruth/image-classification", "human-annotated": "yes", "creation-date": "2022-03-25T20:05:00.000Z"}, "labeling-job-project": 0, "labeling-job-project-metadata": {"class-name": "stop", "job-name": "labeling-job/labeling-job-project", "confidence": 0.0, "type": "groundtruth/image-classification", "human-annotated": "yes", "creation-date": "2022-03-25T20:05:00.000Z"}, "labeling-job-project": 0, "labeling-job-project-metadata": {"class-name": "30 km/h", "job-name": "labeling-job/labeling-job-project", "confidence": 0.0, "type": "groundtruth/image-classification", "human-annotated": "yes", "creation-date": "2022-03-25T20:05:00.000Z"}}}
```

# EXTRAS!

---



# GROUNDTRUTH VS. GROUNDTRUTH PLUS

## AMAZON SAGEMAKER GROUND TRUTH

- Amazon SageMaker ground truth empowers companies and individuals to build and manage data labeling workflows.
- You need to manage human annotators such as mechanical turks, third-party vendors, or your own human labelers.

## SAGEMAKER GROUND TRUTH PLUS

- Amazon SageMaker Ground truth plus creates and manages the workflow on your behalf.
- There is no need to manage workforces. Instead, a trained ML team will handle all data security, labeling, privacy and compliance.
- It reduces data labeling costs by 40%
- All what you need to do is to upload your datasets.



# AUTOMATED DATA LABELING

- Auto labeling uses active learning to determine if input data is well understood or not (Easy or hard).
- Input unlabeled images are used with labeled output images to continuously train machine learning model which in turn reduces the number of human labelers required overtime (~70% reduction).
- If there is no enough dataset, auto labeling is not recommended (threshold 1000 images)

