

SEOUL

19.09.26

# DEV DAY



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

모두를 위한 컴퓨터 비전 딥러닝 툴킷, GluonCV 따라하기

# 1. Overall Lab Guide / Amazon SageMaker Ground Truth

김무현 데이터 사이언티스트  
Amazon Machine Learning Solutions Lab

# Our mission at AWS

---

Put machine learning in the  
hands of every developer

# The Amazon ML stack: 가장 넓고 깊은 기능

## AI SERVICES

App 개발자용



## ML SERVICES

ML 개발자 및  
데이터 과학자용



## ML FRAMEWORKS & INFRASTRUCTURE

ML 연구자 및  
학계용

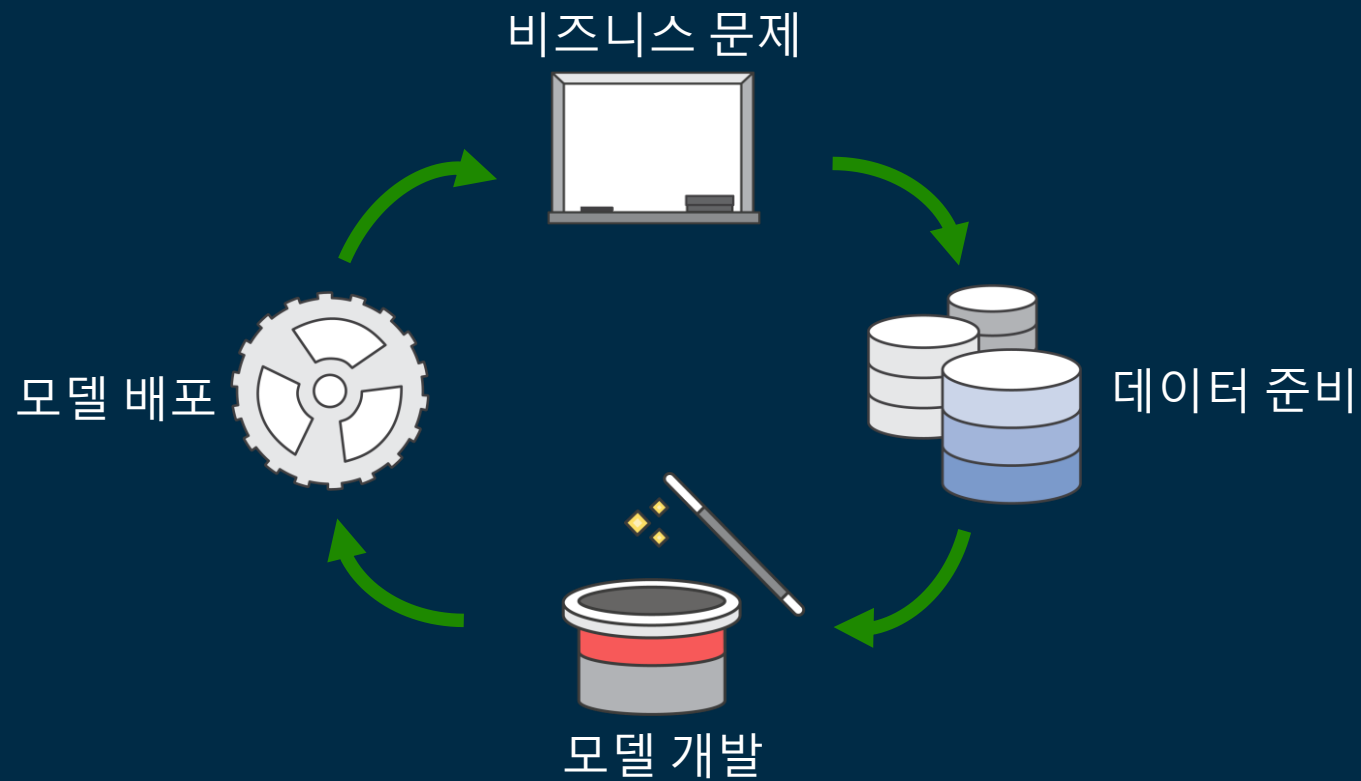


DEV DAY

# Amazon SageMaker



# 머신러닝 사이클



# Amazon SageMaker: Build, train, and deploy ML

Pre-built notebooks for common problems

Collect and prepare training data

intuit.

SIEMENS

Built-in, high performance algorithms

Choose and optimize your ML algorithm



THOMSON REUTERS

One-click training

Set up and manage environments for training

tinder.

GE Healthcare

Optimization

Train and tune model (trial and error)

SIEMENS



One-click deployment

Deploy model in production



Fully managed with auto-scaling

Scale and manage the production environment

CONVOY



DOW JONES

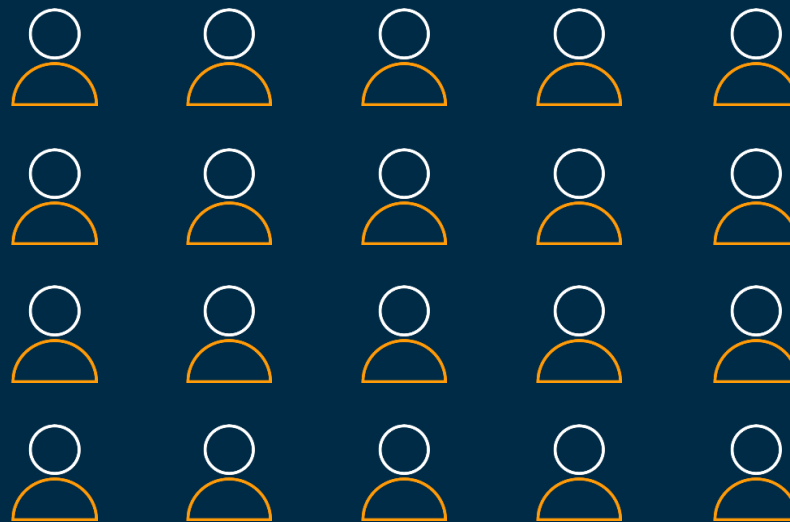
# Amazon Mechanical Turk

Introduction to Amazon Mechanical Turk

<https://www.youtube.com/watch?v=Pjm1uYbuyk4>



# What is crowdsourcing?



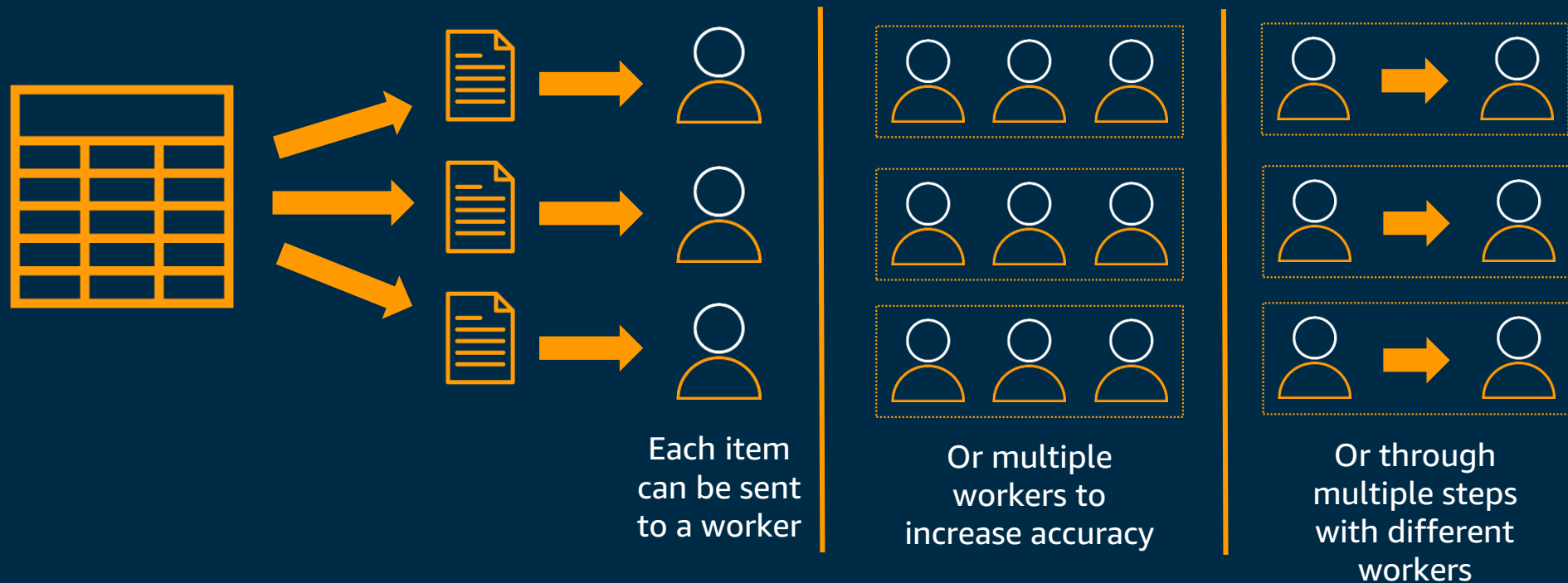
Harnessing the collective power of many individuals

# What is Amazon Mechanical Turk?



Amazon Mechanical Turk operates an online crowdsourcing marketplace for work that requires human intelligence

# What kind of work is on Mechanical Turk?



Microwork: Typically small, repeatable tasks that give workers variety and flexibility

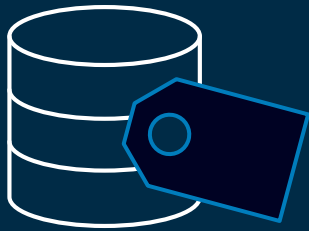
NEW

# Amazon SageMaker Ground Truth

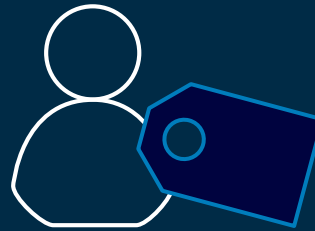
레이블링 작업에 ML을 적용하여 정밀한 훈련 데이터셋  
구축을 도와 드리고 작업 비용을 70%까지 줄입니다

# Amazon SageMaker Ground Truth

ML 훈련용 데이터를 쉽고 정확하게 레이블링



빠른 훈련 데이터  
레이블링



손쉽게 레이블 입력 지원



정확한 결과 획득

---

## KEY FEATURES

머신 러닝을 통한  
자동 레이블링

Ready-made and  
custom workflows for  
image bounding box,  
segmentation, and  
text

Private and public  
human workforce

레이블 관리

# Key Features

## 1. Worker Selection

## 2. Annotation Consolidation

combines the results of multiple worker's annotations into one high-fidelity label.

<https://docs.aws.amazon.com/sagemaker/latest/dg/sms-annotation-consolidation.html>

## 3. Automated Data Labeling

uses machine learning to label portions of your data automatically without having to send them to human workers.

<https://docs.aws.amazon.com/sagemaker/latest/dg/sms-automated-labeling.html>

# 1. Workforce Selection



## Public

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



## Private

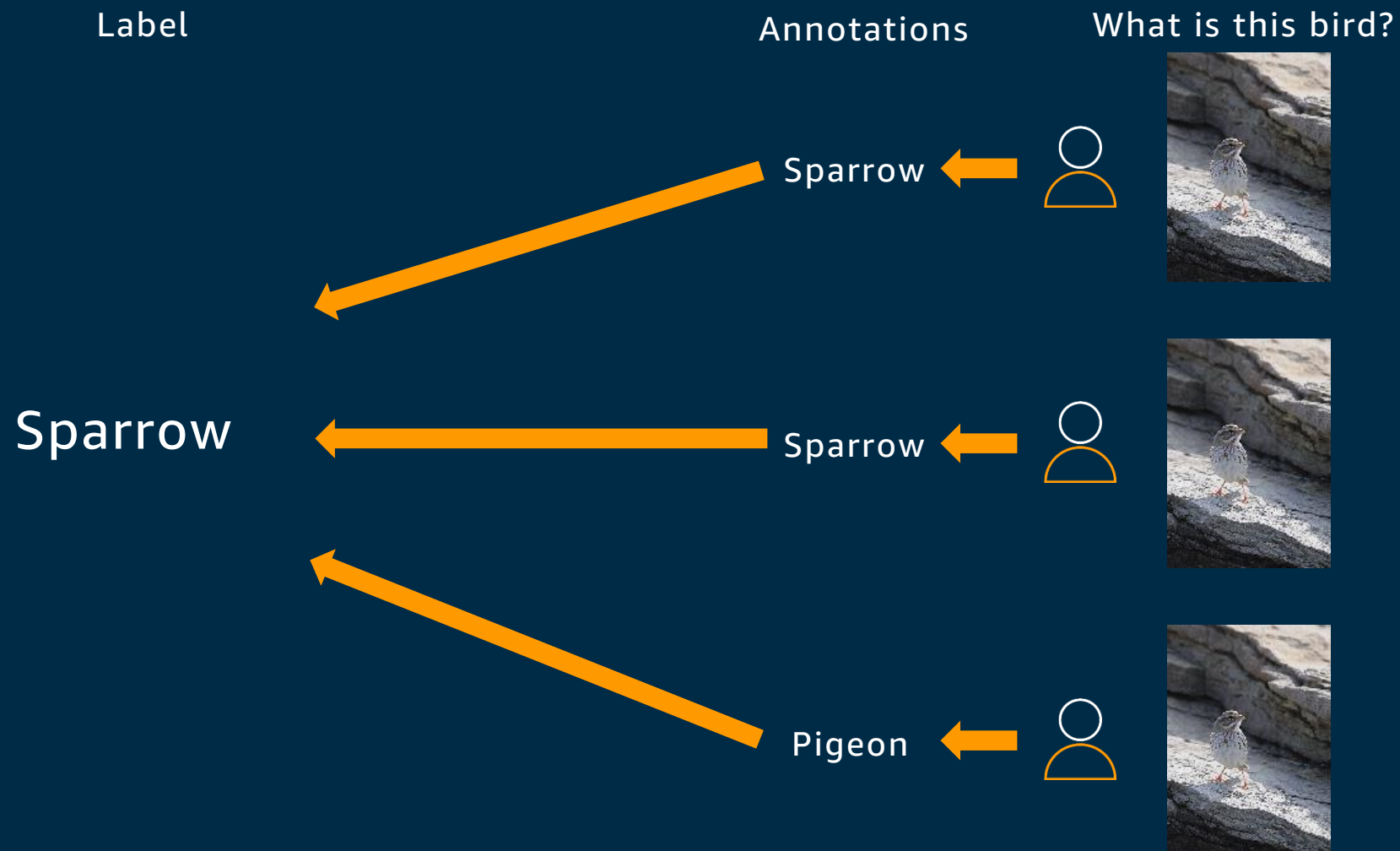
A team of workers that you have sourced yourself, including your own employees or contractors for handling data that needs to stay within your organization



## Vendors

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

## 2. Annotation Consolidation - Why consolidate?





## 2. Annotation Consolidation - Why consolidate?

- The annotators are imperfect:
  - Hard / ambiguous examples
  - Cultural differences (e.g. golf course vs park)
  - Varying worker capabilities
- Solution: requesting redundant labels for each example, and combining them to improve quality
- Annotation Consolidation (AC): a suite of algorithms that enable this combination

# Voting-based Label Consolidation



bulldog



sharpei



bulldog



bulldog



bulldog

- Common solution: **Majority Voting (MV) - baseline**
- Improvement: some workers are better at certain tasks than the others => give their votes a higher preference

# Probability-based Label Consolidation



bulldog



sharpei



bulldog



bulldog

Probabilities of  
correct labels



bulldog 0.1  
sharpei 0.9

$$P(x_1, x_2, x_3, x_4|B) = \prod_i P(x_i|B) = 0.7 * 0.1 * 0.5 * 0.3 \approx 0.01$$

$$P(x_1, x_2, x_3, x_4|S) = \prod_i P(x_i|S) = 0.3 * 0.9 * 0.5 * 0.7 \approx 0.1$$

$$\frac{P(S|x_1, \dots, x_4)}{P(B|x_1, \dots, x_4)} = \frac{P(x_1, \dots, x_4|S)}{P(x_1, \dots, x_4|B)} \approx 10$$

# Annotation Consolidation

- Technical challenge: how do we infer worker quality?
- **Our approach:** an improvement of the well-known **Dawid Skene algorithm**\*, doesn't require a test dataset



\* Dawid AP, Skene AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. Applied statistics. 1979 Jan 1:20-8.

# Annotation Consolidation

## Algorithm: modified Dawid-Skene (MDS) model

Estimate individual worker accuracies AND the final label at the same time

The exact algorithms are slight different per each task

## Resources

**Paper:** Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics), 28(1), 20-28 ([pdf](#))

**Blog:** <https://aws.amazon.com/ko/blogs/machine-learning/use-the-wisdom-of-crowds-with-amazon-sagemaker-ground-truth-to-annotate-data-more-accurately/>

**Implementation Code (Jupyter notebook):** [https://github.com/aws-labs/amazon-sagemaker-examples/blob/master/ground\\_truth\\_labeling\\_jobs/annotation\\_consolidation/ACSBlogPost.ipynb](https://github.com/aws-labs/amazon-sagemaker-examples/blob/master/ground_truth_labeling_jobs/annotation_consolidation/ACSBlogPost.ipynb)

# 3. Automated Data Labeling

**Optional** - you don't have to

Supported task (now)

Image classification, Object detection, Text classification

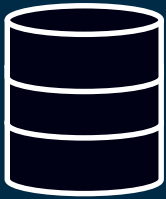
Guides & Tips

Strongly suggest a minimum with **5,000 objects**

Higher accuracy requirements = less number of auto-labeled data

<https://docs.aws.amazon.com/sagemaker/latest/dg/sms-automated-labeling.html>

# Active Learning Loop



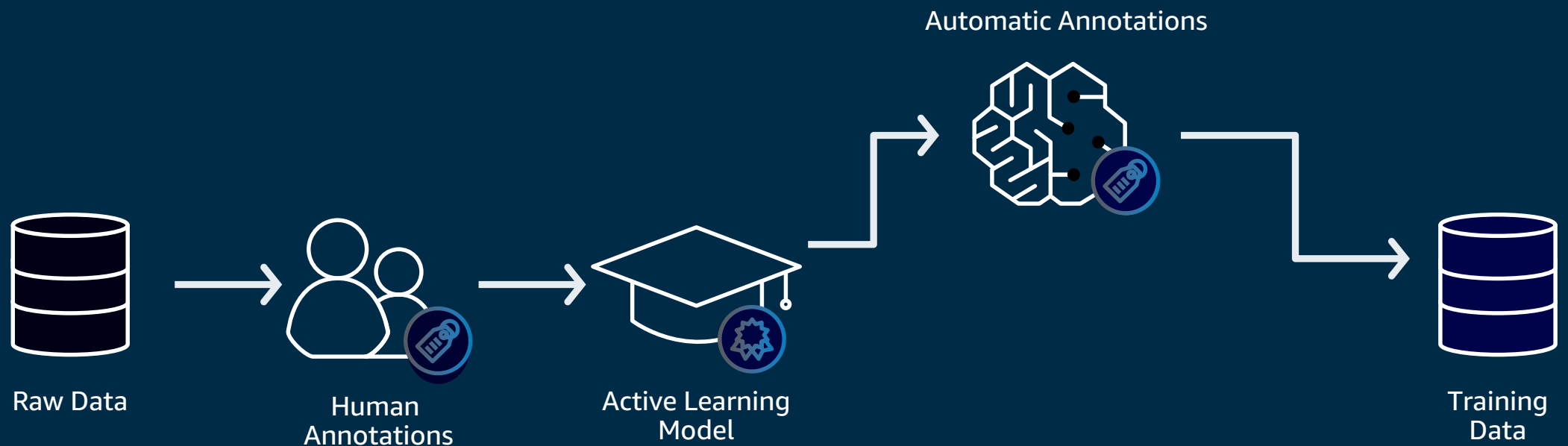
Raw Data

# Active Learning Loop

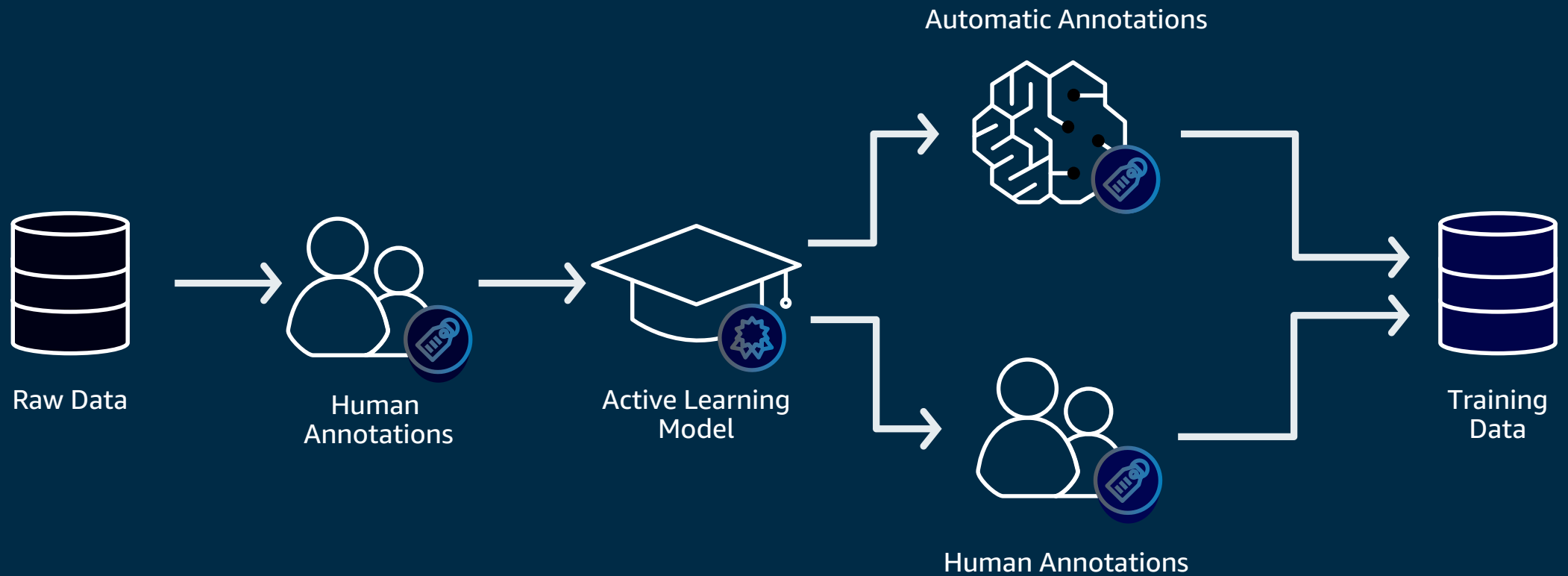




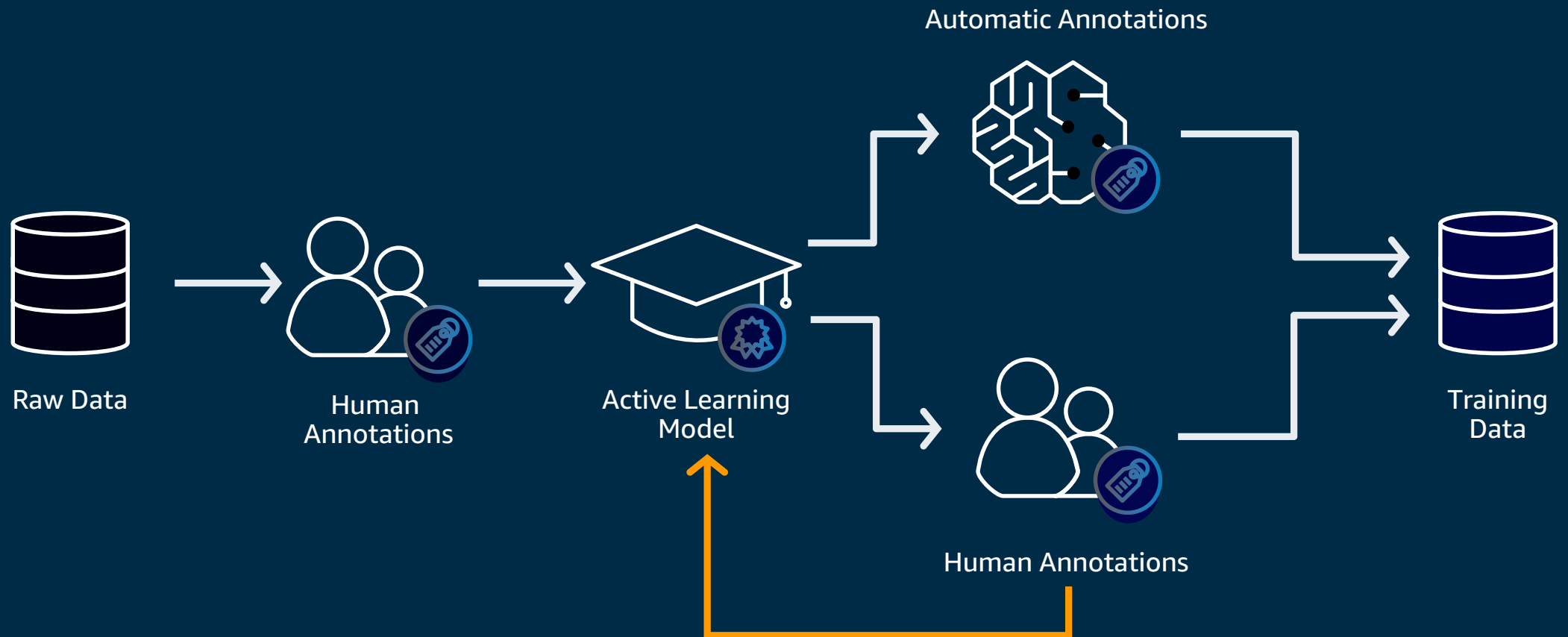
# Active Learning Loop



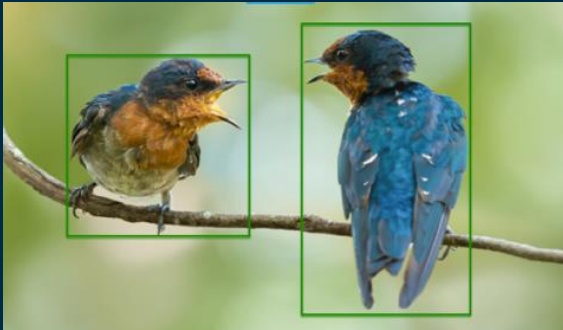
# Active Learning Loop



# Active Learning Loop



# Use Pre-built Labeling Workflows or Set Up Your Own



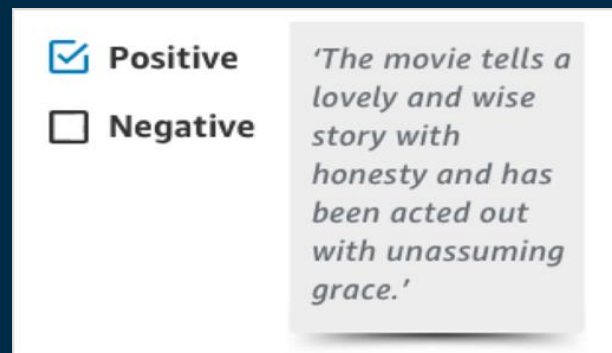
Bounding boxes



Image classification



Semantic segmentation

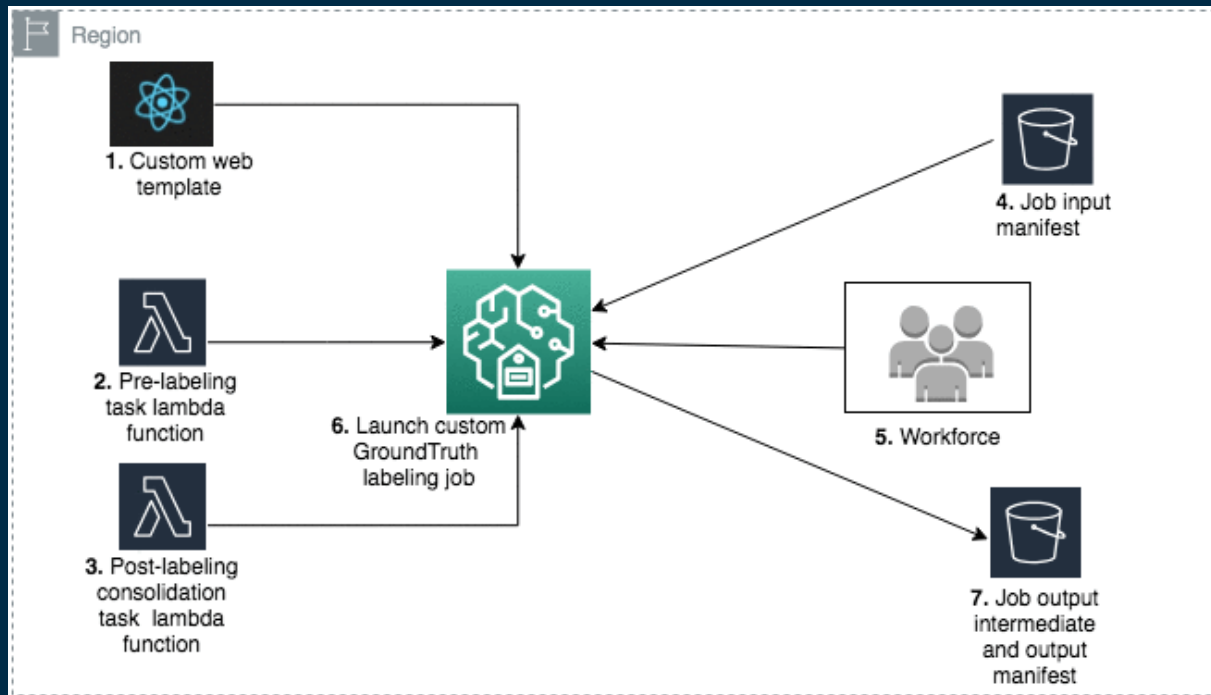


Text classification



Custom

# How to Set Up Your Own Custom Labeling Workflow



**Instructions**

The task can be completed with blank, or saved and returned to when time is available to make more progress. If there is evidence in the record to support or deny abstract quality, **highlight** it with the cursor and select **Yes** or **No**. Add any notes you have for each task in the **Notes** free text area.

CONCLUSIONS

Astronomy Swarthmore College, Swarthmore, PA 19081, USA Center for Astrophysical Sciences, Johns Hopkins University, Baltimore MD 21218

**ABSTRACT** We leverage new high-quality data from Hubble Space Telescope program GO-14164 to explore the variation in horizontal branch morphology among globular clusters in the Large Magellanic Cloud (LMC). **BACKGROUND** Our new observations lead to photometry with a precision commensurate with that available for the Galactic globular cluster population. **CONCLUSIONS** Our analysis indicates that, once metallicity is accounted for, clusters in the LMC largely share similar horizontal branch morphologies regardless of their location within the system. Furthermore, the LMC clusters possess, on average, slightly redder morphologies than most of the inner halo Galactic population; we find, instead, that their characteristics tend to be more similar to those exhibited by clusters in the outer Galactic halo. Our results are consistent with previous studies showing a correlation between horizontal branch morphology and age.

**Key words:** (galaxies:) Magellanic Clouds, galaxies: star clusters: general, (galaxies:) globular clusters: general, stars: horizontal branch

Notes:

Missing Limitations

Is this a good Abstract?

Yes

No

**Submit**

<https://aws.amazon.com/blogs/machine-learning/build-a-custom-data-labeling-workflow-with-amazon-sagemaker-ground-truth/>

<https://github.com/nitinaws/gt-custom-workflow>

DEV DAY

# Demo



## Task type [Info](#)

### Task selection

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

☒ **Image classification**

Get workers to categorize images into specific classes. [Info](#)

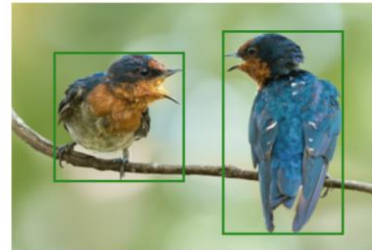
☒ **Basketball**

☐ **Soccer**



☐ **Bounding box**

Get workers to draw bounding boxes around specified objects in your images. [Info](#)



☐ **Text classification**

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

☒ **Positive**

☐ **Negative**

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

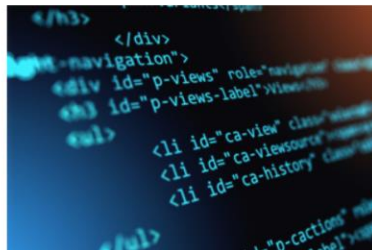
☐ **Semantic segmentation**

Get workers to draw pixel level labels around specific objects and segments in your images. [Info](#)



☐ **Custom**

Customize tasks for your workers to label your dataset. [Info](#)





Step 1

Specify job details

Step 2

Select workers and  
configure tool

## Select workers and configure tool

### Workers [Info](#)

#### Worker types

☒ **Public**

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 you 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.

#### Price per task

We recommend you choose a price consistent with the approximate time it takes to complete a task. We have provided time estimates for each price as guideline to help you decide how you want to price your task.

\$0.840

Time estimate: 3 mins - 3.5 mins

☐ The dataset does not contain adult content. [Info](#)

☐ I understand that my dataset will be viewed by the Amazon Mechanical Turk public workforce and I acknowledge that my dataset does not contain personally identifiable information (PII). [Info](#)

#### ▼ Additional configuration - optional

Automated data labeling, workers per dataset object

#### Number of workers per dataset object

The number of distinct workers you want to perform the same task on a dataset object. This can help increase the accuracy of the data labels.

3

workers

### Semantic segmentation

Use semantic segmentation to identify the contents of an image to the pixel level. Workers choose a class and then use the tool to surround portions of the image with a polygon. Workers can apply a different class to each polygon that they draw in the image. You define one or more classes when you create the labeling job. Use the instructions to guide your users to make the correct choices.

#### Learn more

- [Amazon SageMaker Ground Truth](#) [↗](#)



Step 1

Specify job details

Step 2

Select workers and  
configure tool

## Select workers and configure tool

Workers [Info](#)

Worker selection

### Worker types

☒ Public

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.

### Price per task

We recommend you choose a price consistent with the approximate time it takes to complete a task. We have provided time estimates for each price as guideline to help you decide how you want to price your task.

\$0.840

Time estimate: 3 mins - 3.5 mins

☐ The dataset does not contain adult content. [Info](#)

☐ I understand that my dataset will be viewed by the Amazon Mechanical Turk public workforce and I acknowledge that my dataset does not contain personally identifiable information (PII). [Info](#)

### Additional configuration - optional

Automated data labeling, workers per dataset object

number of workers on the SAME data

### Number of workers per dataset object

The number of distinct workers you want to perform the same task on a dataset object. This can help increase the accuracy of the data labels.

3

workers

## Semantic segmentation

Use semantic segmentation to identify the contents of an image to the pixel level. Workers choose a class and then use the tool to surround portions of the image with a polygon. Workers can apply a different class to each polygon that they draw in the image. You define one or more classes when you create the labeling job. Use the instructions to guide your users to make the correct choices.

### Learn more

- [Amazon SageMaker Ground Truth](#)


Instructions

View full instructions

View tool guide


Good example

Enter description to explain the correct label to the workers




Bad example

Enter description of an incorrect label



Please find a real photo that is not a drawing or synthesized image.



Select an option

Real Photo1

Non-real Photo2

Zoom in


Zoom out

Move

Fit image

Submit

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.



How to Check x Amazon Sage x https://s3.am x sagemaker gr x Use the wisdc x annotation\_cc x output.manife x ACSBlogPost x S3 Managem x

https://s3.amazonaws.com/sagemaker-ground-truth-labeling-task-preview/preview.html


### Instructions

[View full instructions](#)

[View tool guide](#)


#### Good example

Enter description to explain the correct label to the workers



#### Bad example

Enter description of an incorrect label



Please find a real photo that is not a drawing or synthesized image.

Text description for the workers

Select an option

Real Photo	1
Non-real Photo	2

Predefined TAGs for the workers

image data in my s3 bucket

good/bad examples for the workers

out Move Fit image

Submit



Hello Julia

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

Log out

## Instructions



[View full instructions](#)

[View tool guide](#)

### GOOD EXAMPLE

Only draw bounding boxes on dogs. The sides of the box should touch the top, bottom, left, and right boundaries of the dog.



### BAD EXAMPLE

The box drawn below does not cover all sides of the dog.



Draw a bounding box on any dog that you see in the image.



## Label

☒ Dog

1

Example:  
Object  
Detection



Box



Undo



Redo



Zoom



Fit image



Move

☐ Nothing to label

Submit

Hello Julia

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

Log out

## Instructions



[View full instructions](#)

[View tool guide](#)

In this task, you will need to select the category that best describes the image.

### Basketball



### Swimming



### Football

Which sport is played in the image?



Select an option

Basketball	1
Swimming	2
Football	3
Soccer	4

Example:  
Image  
Classification



Zoom



Fit image



Move

Submit



Hello Julia

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

Log out

## Instructions



[View full instructions](#)

[View tool guide](#)

Inspect the image and select a label first before each annotation.

### GOOD EXAMPLE

Label is correctly assigned. The object is highlighted nicely.

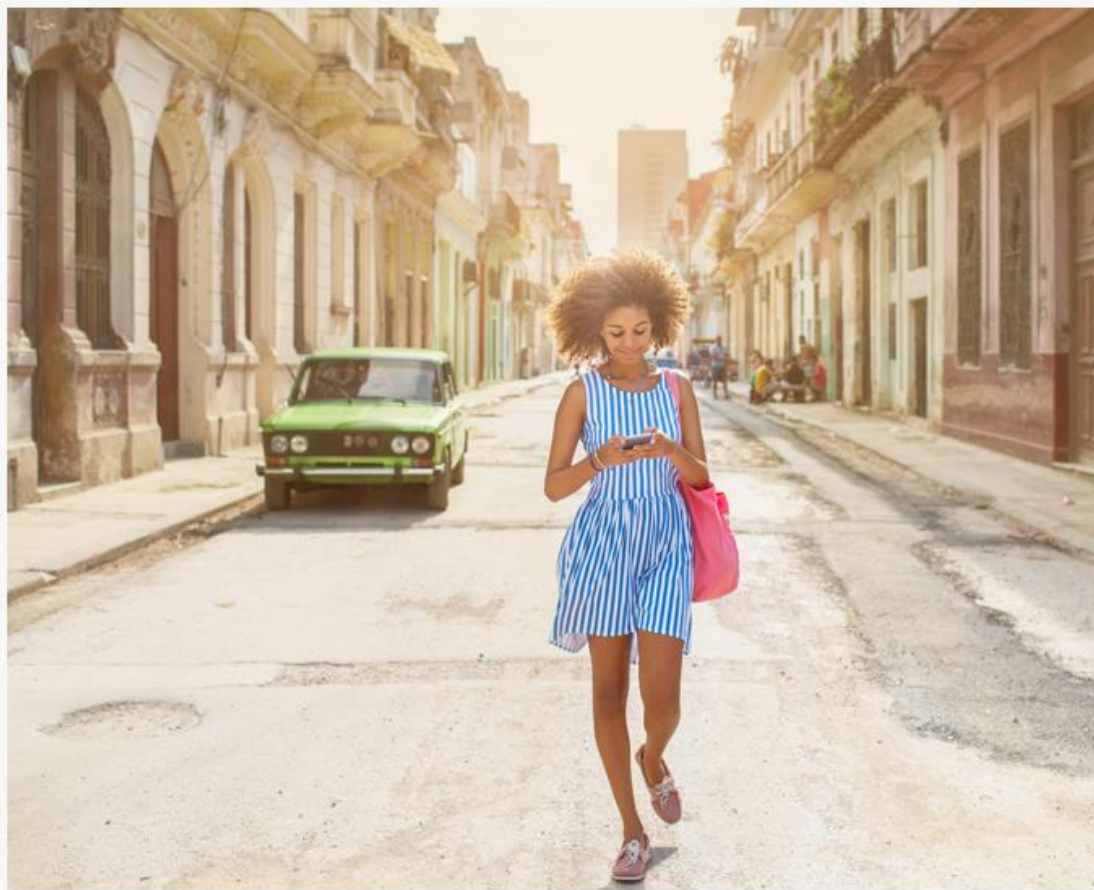


### BAD EXAMPLE

Object car is not labeled.




## Highlight the image for each label listed.



## Labels



	Car	 1
	People	 2

Example:  
Semantic  
Segmentation



Polygon



Brush



Eraser



Dimmer



Undo



Redo



Zoom



Fit image



Move

☐ Nothing to label

Submit

## Instructions

[View full instructions](#)[View tool guide](#)**Positive**

Positive means some aspects of the text uncover a positive mood, such as praise, pleasure, recommendation or a favorable comparison.

Example: This view is amazing.

**Negative**

Negative means some aspects of the text uncover a negative mood, such as criticism, insults or a negative comparison.

Example: I dislike old cabin cruisers.

**Neutral**

Neutral means no emotions are implied.

Example: It's going to rain tomorrow.

**Unsure**

Select this option when you are not sure what sentiment the content is

## What is the overall sentiment in the text?

My boyfriend and I went to watch The Guardian. At first I didn't want to watch it, but I loved the movie- It was definitely the best movie I have seen in sometime. They portrayed the USCG very well, it really showed me what they do and I think they should really be appreciated more. Not only did it teach but it was a really good movie. The movie shows what the really do and how hard the job is. I think being a USCG would be challenging and very scary. It was a great movie all around. I would suggest this movie for anyone to see. The ending broke my heart but I know why

## Select an option

☒ Positive 1☐ Negative 2☐ Neutral 3☐ Unsure 4

Example:  
Text  
Classification

Submit

DEV DAY

# Thank you!





# 여러분의 피드백을 기다립니다!



## 강연 평가 및 설문 조사

QR 코드를 통해 AWS DEV DAY SEOUL에  
대한 여러분의 의견을 공유해주세요.  
강연 평가 및 설문 조사에 참여해 주신 분께는  
등록데스크에서 특별한 기념품을 드립니다.



## 강연 영상

AWS DEV DAY SEOUL 강연 영상은  
행사 종료 후 메일로 공유드릴 예정입니다.



## #AWSDEVDAYSEOUL

소셜미디어에 행사 참여 소감을 공유해주세요!

