



AWS
re:Invent

AIM 33

How to design high quality data labeling pipelines

Fedor Zhdanov

Principal Applied Scientist
Amazon Web Services

Jonathan Buck

Software Dev Engineer
Amazon Web Services

Amazon SageMaker Ground Truth



Labeling

Set up and manage labeling jobs for your datasets.



Notebooks

Explore AWS data in your notebooks, and use algorithms to create models via training jobs.



Training

Track training jobs at your desk or remotely. Leverage high-performance AWS algorithms.



Inference

Create models for hosting from job outputs, or import externally trained models into SageMaker.

Integrated label-> build-> train-> deploy flow
Ground Truth makes data labeling easy, fast, and accurate

What you will be able to do after the talk

- Know how Amazon SageMaker Ground Truth works:
 - How it provides high quality annotations
 - How it does Active Learning to save on annotation costs
- Learn how to use confidences in annotations
- Learn how to chain jobs
- Learn how to verify and audit results of your labeling jobs
- Learn how to filter your data
- Learn how you can build hierarchical taxonomies in labeled classes

Human workforce options



MTurk

An on-demand 24x7 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

Provide details for a labeling job

Step 1
Specify job details

Step 2
Select workers and
configure tool

Amazon SageMaker > Labeling jobs > Create labeling job

Specify job details

Job overview

Job name

Input dataset location info
Provide a path to the S3 location where your manifest file is stored. [Create manifest file](#), if you don't have one.


Output dataset location info
Provide a path to the S3 location where you want your labeled dataset to be stored.


IAM role info
Provide the ID or ARN for your own IAM role for Amazon SageMaker to access your S3 bucket. Choose a role or let us create a role with the [SageMakerFullAccess](#) IAM policy attached.

► **Additional configuration - optional**
Dataset object selection, encryption key

Task type


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


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


☐ **Image classification**
Get workers to categorize images into specific classes. [Info](#)
☒ Basketball
☐ Soccer


☐ **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](#)


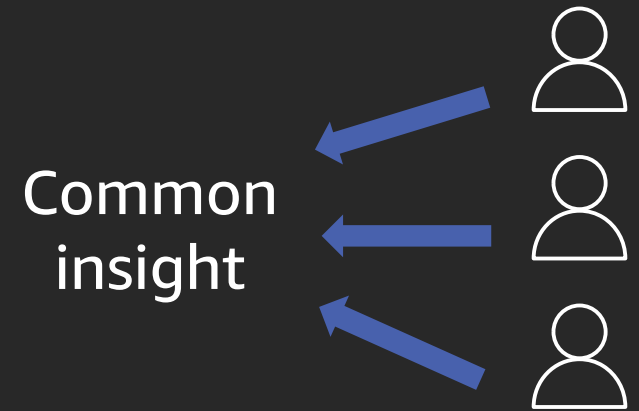
► **Tags - optional**

Cancel

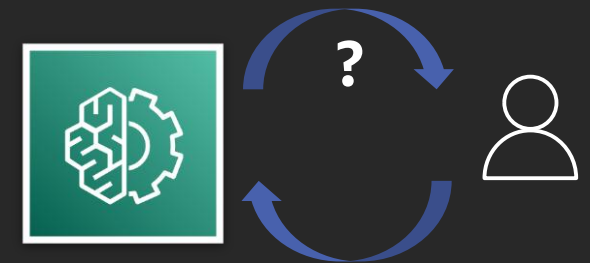
Next

Key ideas: Machine Learning and humans in the labeling loop

Consolidate annotations from multiple workers for greater accuracy



Only send to humans examples which are hard for the machines to label well:
Reduce annotation costs



AWS console: View labels for images

AWS Services Edit design@awsdesign Oregon Support

Amazon SageMaker > Labeling jobs > Identifying-dogs

Identifying-dogs

Actions

Labeling job summary

View labeling tool


Status Completed	Input dataset location s3://path-to-manifest/	Task type Bounding box
Creation time July 15, 2018 14:24 UTC	Dataset objects 20	Worker type Public
	Output dataset location s3://path-to-location/	

Output | Tags


Labeled dataset objects (20/20)

Search by object name


< 1 2 3 4 5 ... 8 > ⚙




Name
Image 1234




Name
####




Name
####




Name
####




Name
####



Name
####



Name
####

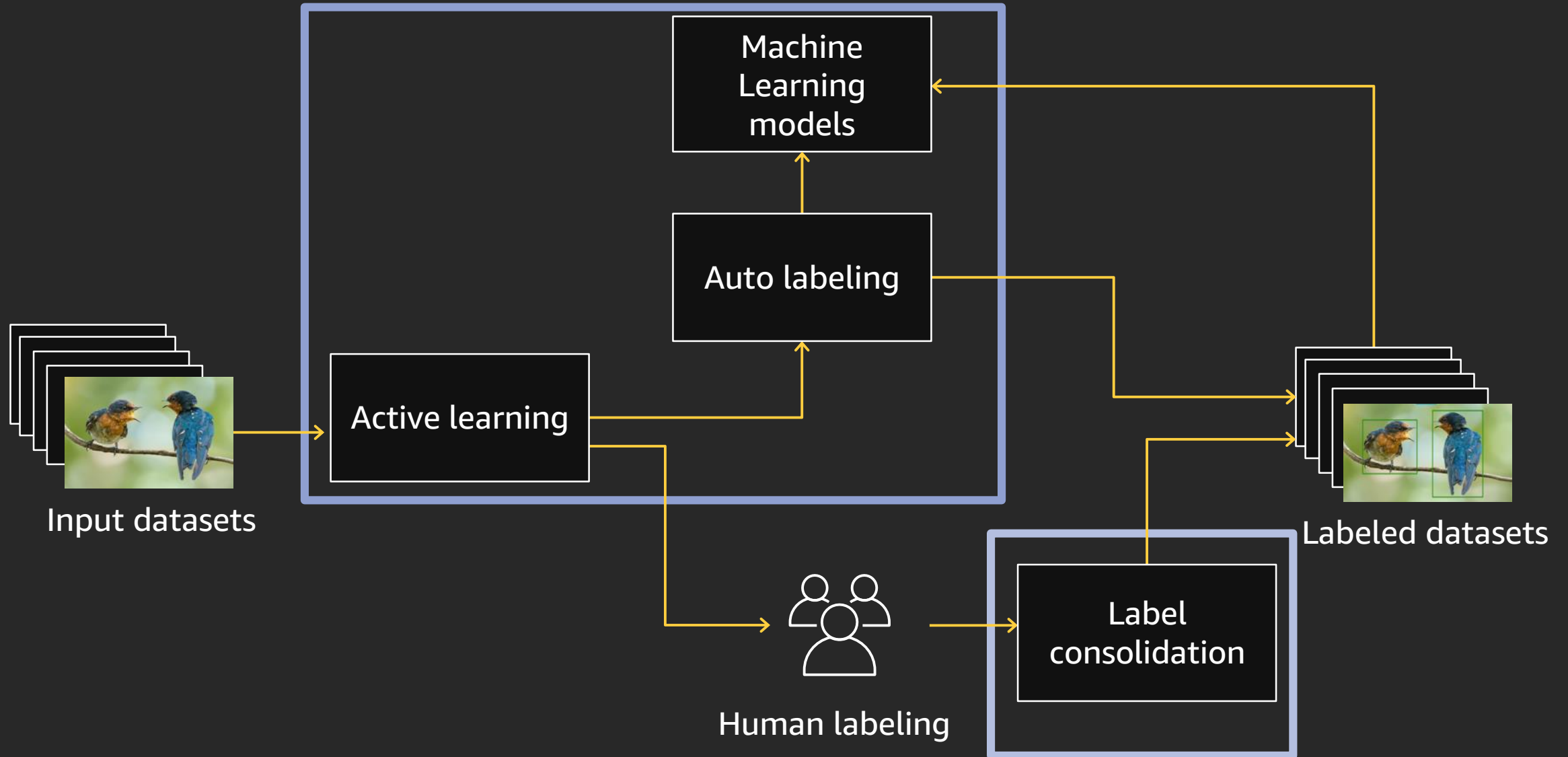


Name
####

Feedback English (US) © 2018 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

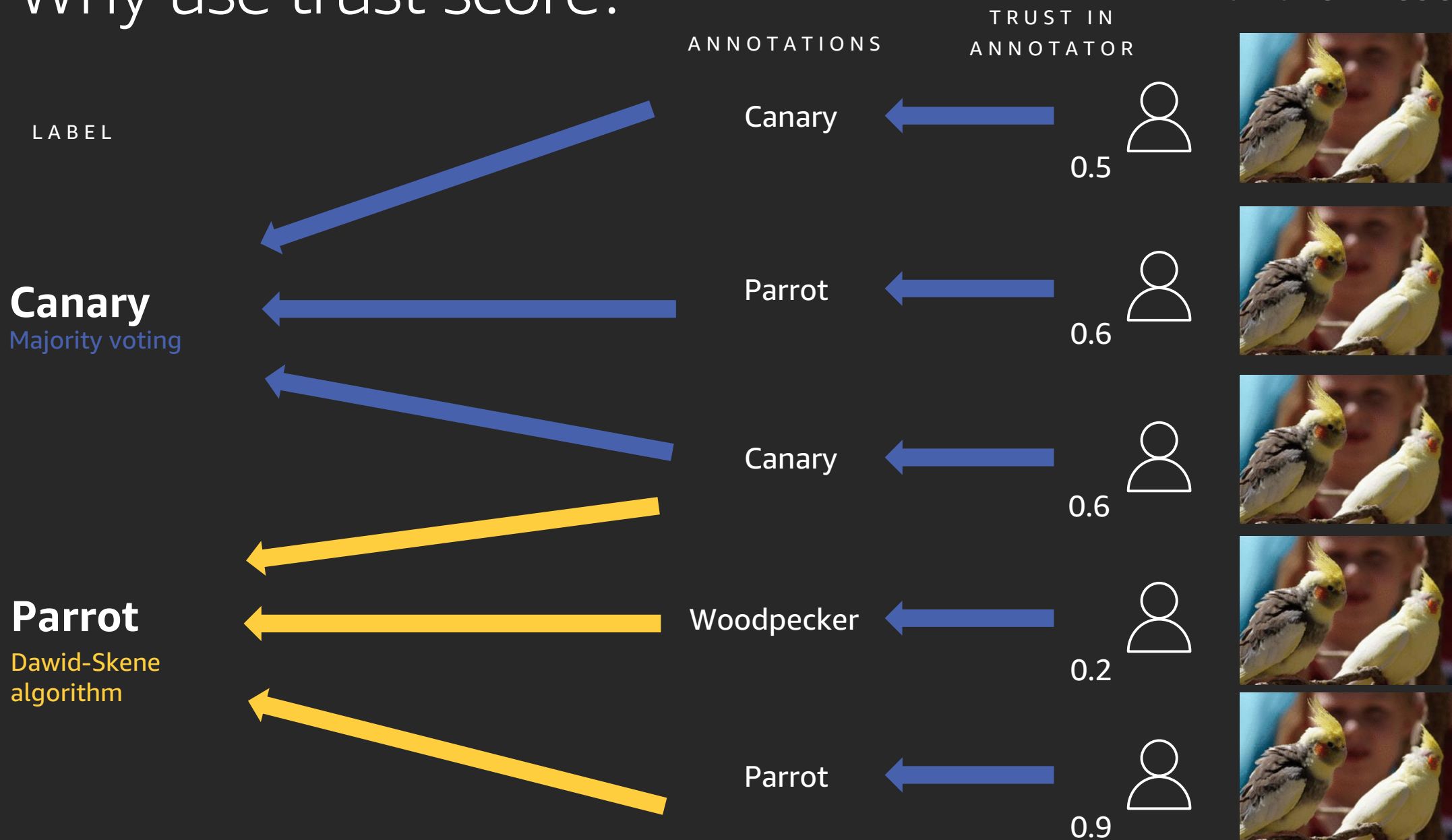
Deep dive into “how”

Reminder: Machine Learning with humans in the loop

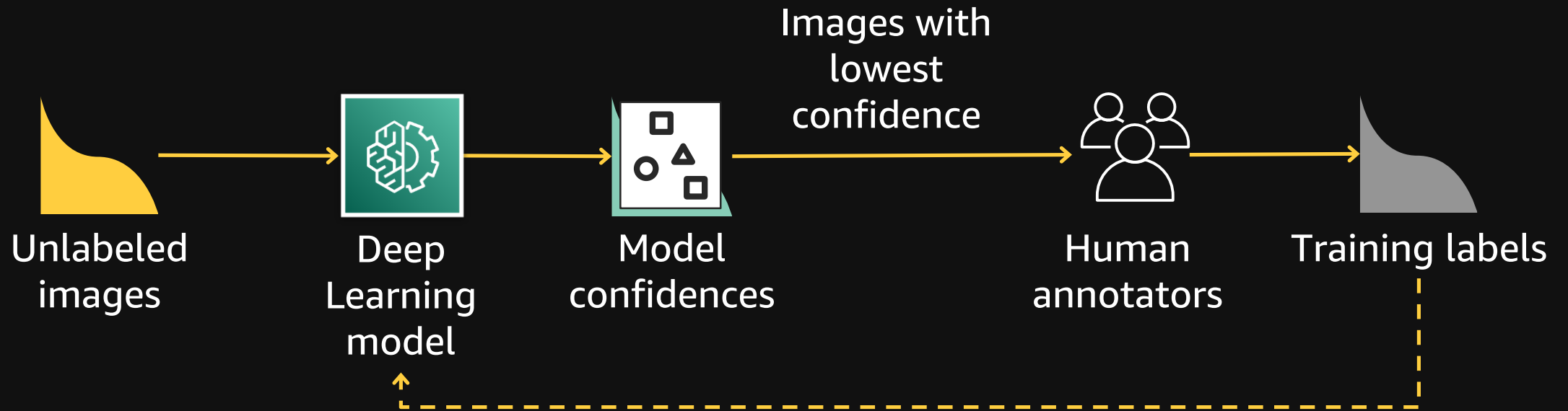


Why use trust score?

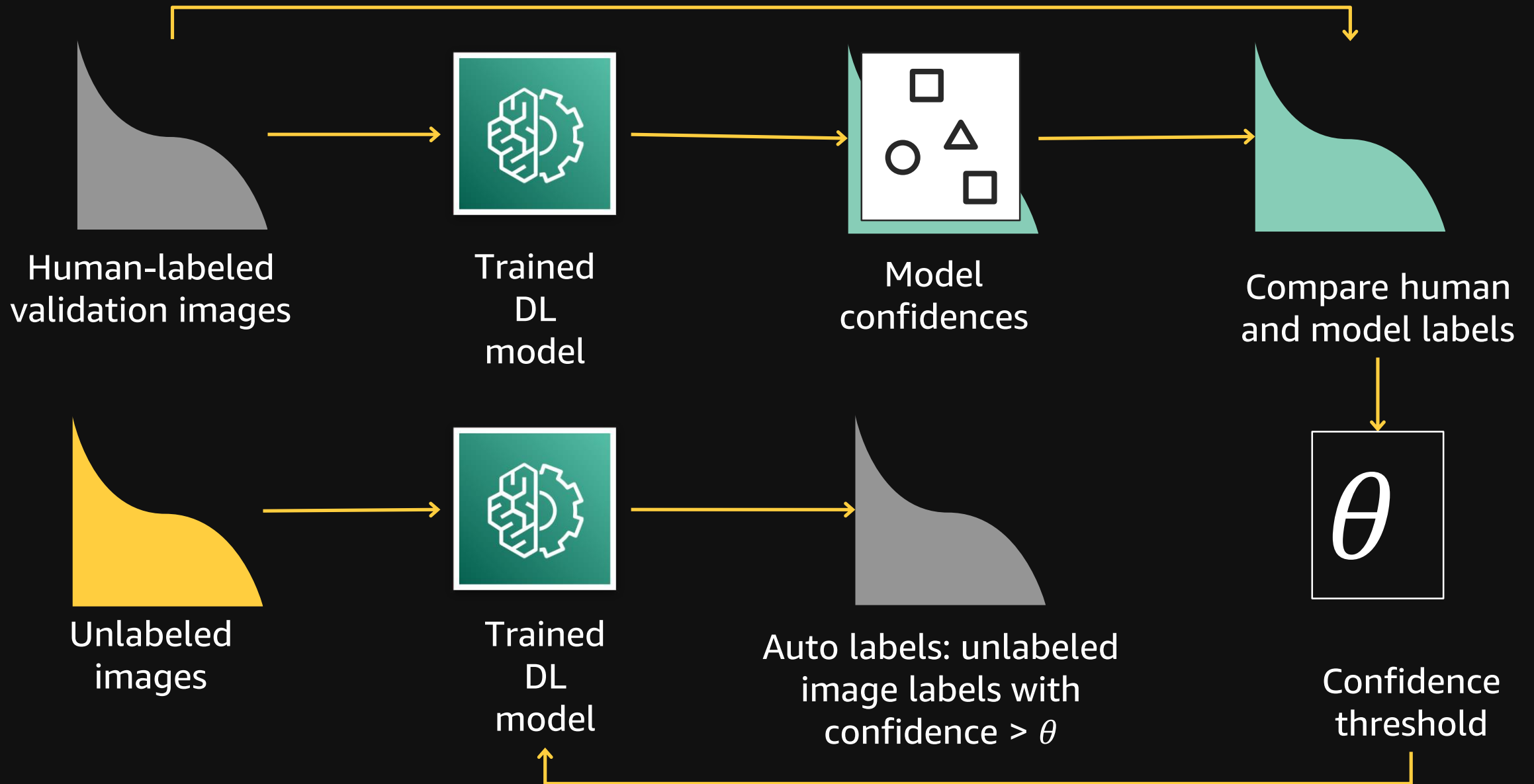
What are these birds?



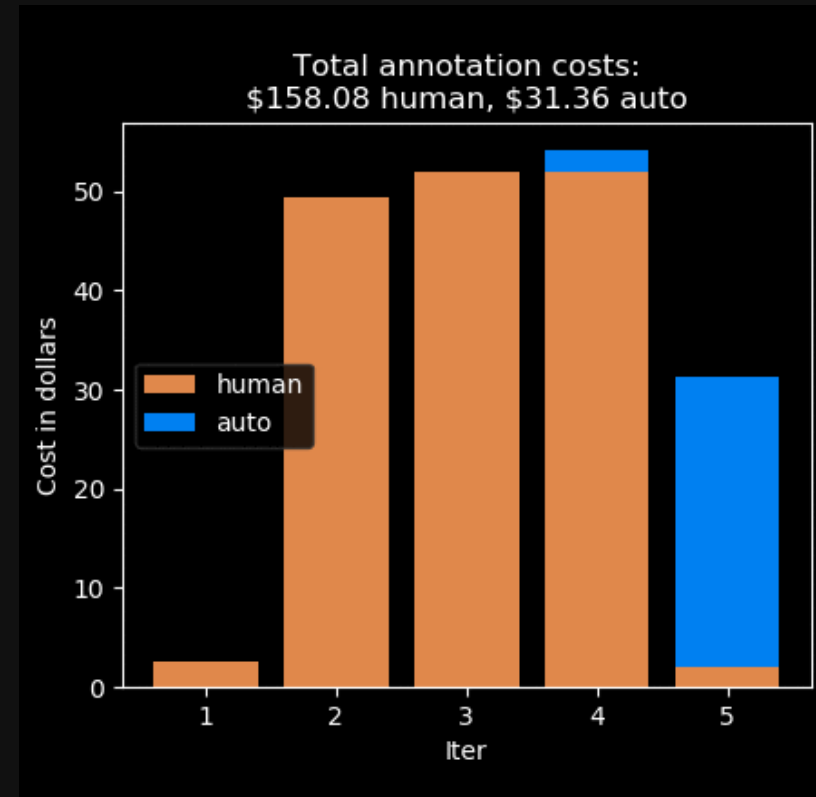
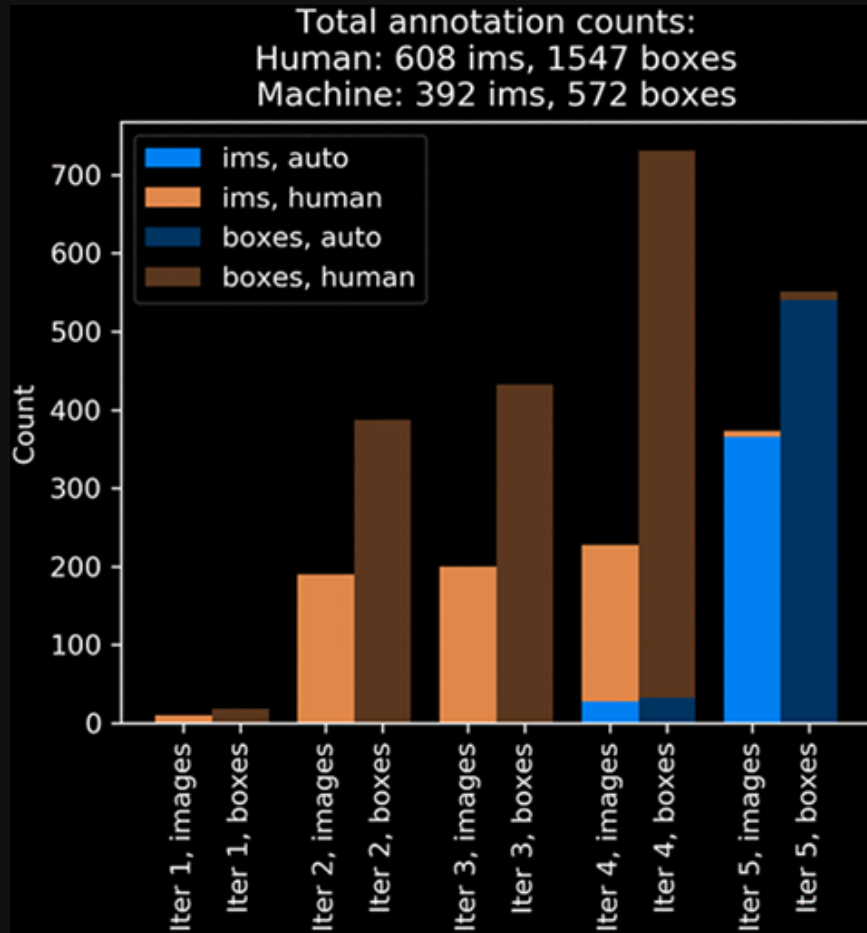
Active learning



Auto labeling



Auto-annotation for the bounding boxes



Total cost \$189.44 instead of \$260

Cost saving 27%

Larger datasets often bring more savings

Read more on auto-annotation:

<https://aws.amazon.com/blogs/machine-learning/annotate-data-for-less-with-amazon-sagemaker-ground-truth-and-automated-data-labeling/>

Indicator of annotation quality

```
{
  "source-ref": "S3 bucket location",
  "bounding-box":
  {
    "image_size": [{ "width": 500, "height": 400, "depth": 3}],
    "annotations":
    [
      {"class_id": 0, "left": 111, "top": 134,
        "width": 61, "height": 128},
      {"class_id": 5, "left": 161, "top": 250,
        "width": 30, "height": 30},
      {"class_id": 5, "left": 20, "top": 20,
        "width": 30, "height": 30}
    ]
  },
  "bounding-box-metadata":
  {
    "objects":
    [
      {"confidence": 0.8},
      {"confidence": 0.9},
      {"confidence": 0.9}
    ],
    "class-map":
    {
      "0": "dog",
      "5": "bone"
    },
    "type": "groundtruth/object_detection",
    "human-annotated": "yes",
    "creation-date": "2018-10-18T22:18:13.527256",
    "job-name": "identify-dogs-and-toys"
  }
}
```

- Can we have any indicators for annotation quality?

Confidence scores for the annotations

- Auto-annotated and Human-annotated confidences are different as one is coming from the Deep Neural Network and the other is calculated from worker agreement.
- Only compare the confidence scores with each other for the objects within the same labeling job, and not rely on the high or low confidence scores as absolute values

Confidence Scores for Human Annotation

- Confidence score is the posterior probability of the output class from the Dawid-Skene model.
- Object Detection: Model predicts the Intersection over Union (IoU) using features like worker disagreement on box corners and number of annotators.
- Semantic Segmentation: Model predicts the Intersection over Union (IoU) using features like image complexity, worker disagreement across all pixels and number of annotators

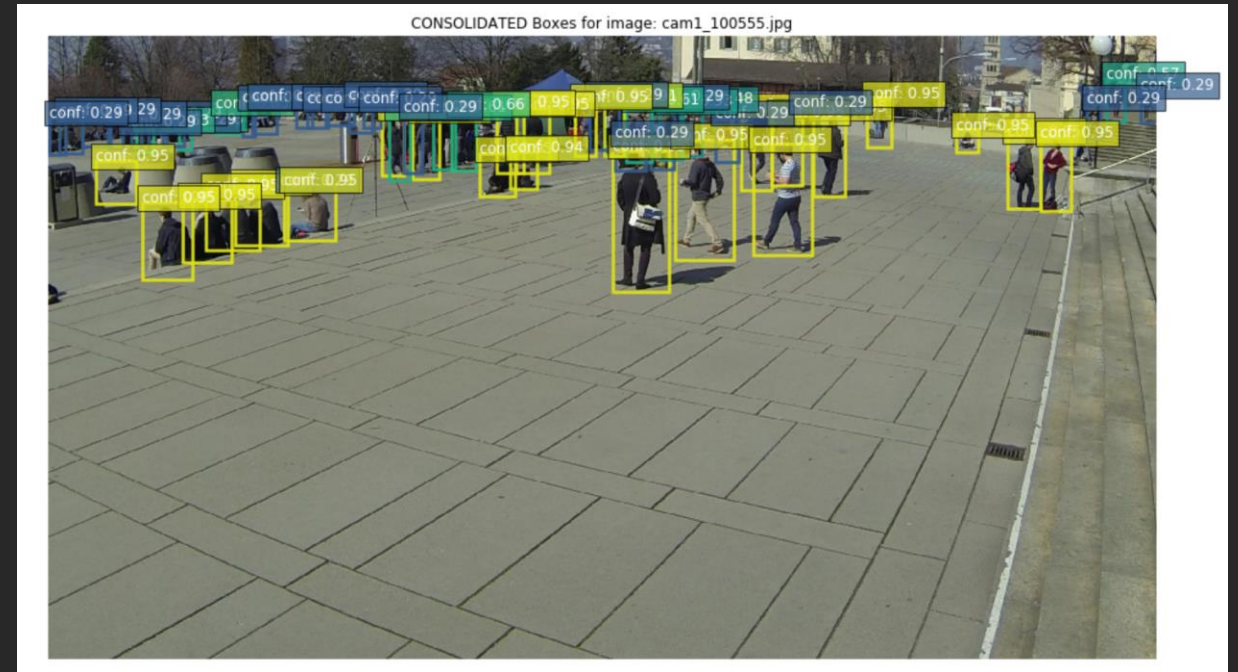
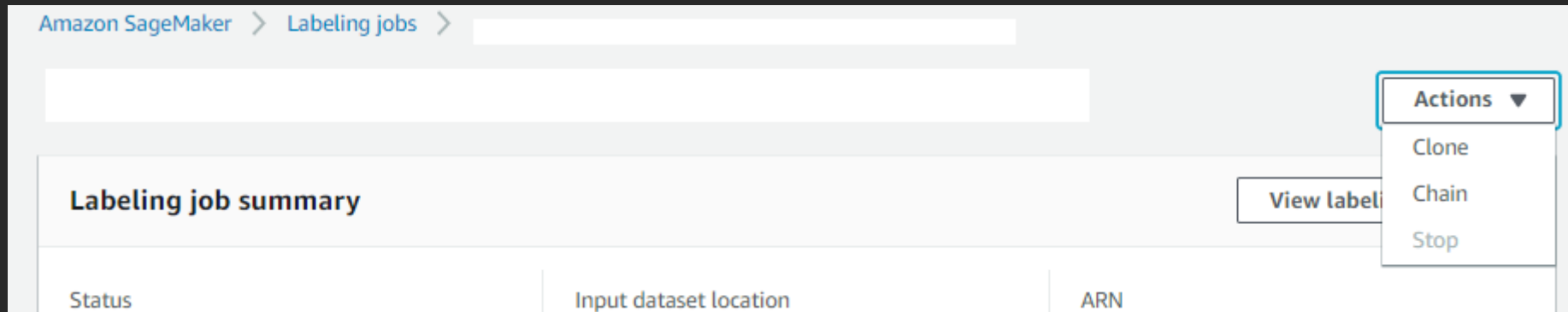


Image from
T. Chavdarova, P. Baqu, S. Bouquet, A. Maksai, C. Jose, T. Bagautdinov, L. Lettry, P. Fua, L. Van Gool, and F. Fleuret, "WILDTRACK: A multicamera HD dataset for dense unscripted pedestrian detection," 2018.

Advanced features

Job Chaining

- Select a labeling job in the console, and select “chain”
- Subsequent job can be same or different modality
- All annotations and metadata from first job will be carried over to the output of the subsequent job



Audit Workflows

- Start with the “chain” feature
- Use with advanced functionality
 - Filter low-confidence annotations from previous labeling job
 - Filter specific classes
- The original task type is *maintained*
 - Existing annotations are updated/adjusted

▼ Existing-labels display options - optional

Enable this option to display existing labels for label adjustments.

☒ I want to display existing labels from the dataset for this job. [Info](#)

Label attribute name [Info](#)



Choose the label attribute name that you want to adjust.

jonabuck-catsvdogs-1018 ▼


Bounding box labeling tool [Preview](#)

Provide labeling instructions with examples below for workers. Workers will be viewing these instructions when they perform your task. You can add up to 10 labels for workers to choose from. See guidelines for [creating high-quality instructions](#)

H1 H2 B I A

Draw a box around each cat and dog in the image.



Labels

Add up to 10 labels

Cat ×

Dog ×

[Add label](#)

► Additional instructions - optional

Verification Workflows

- Start with the “chain” feature
- The original labeling task is transformed into a *classification* task, e.g.,
 - Correct / Incorrect
 - Occluded / Not Occluded
- Consider downstream filters to build hierarchies

Task type [Info](#)


Task category
Select the type of data being labeled to view available task templates for it or select 'Custom' to create your own.

Image ▼

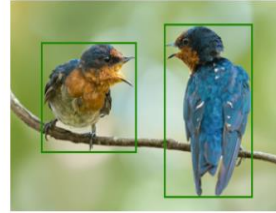
Task selection
Select the task that a human worker will perform to label objects in your dataset.

☐ **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](#)




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



☒ **Label verification**
Get workers to verify existing labels in your dataset. [Info](#)

☒ Correct label
☐ Incorrect label

Car



Existing-labels display options [Info](#)
Existing labels from the dataset will be displayed for verifying.

Label attribute name [Info](#)
Choose the label attribute name that you want to verify.

jonabuck-catsvdogs-1018 ▼

Label verification tool [Preview](#)

Provide instructions to help workers identify correct and incorrect labels. Workers will refer to these instructions for each task to verify existing labels. You can add up to 30 labels for workers to choose from. See guidelines for [creating high-quality instructions](#).

► Existing labels

Review the existing labels on the objects and choose the appropriate option.

About existing labels
Draw a box around each cat and dog in the image.
Good example
Provide instructions to help workers understand how the task was supposed to be done.

Bad example
Provide examples of mislabeled items that should be rejected.

Select an option
Add up to 30 labels

Correct
Incorrect
Add label

► Additional instructions - optional

Decision Points for Filtering Data

- Class/Label
- Machine labeled vs. human labeled
- Confidence score
- Modality
- Verification status (yes/no)
- Audit status (adjusted/not adjusted)

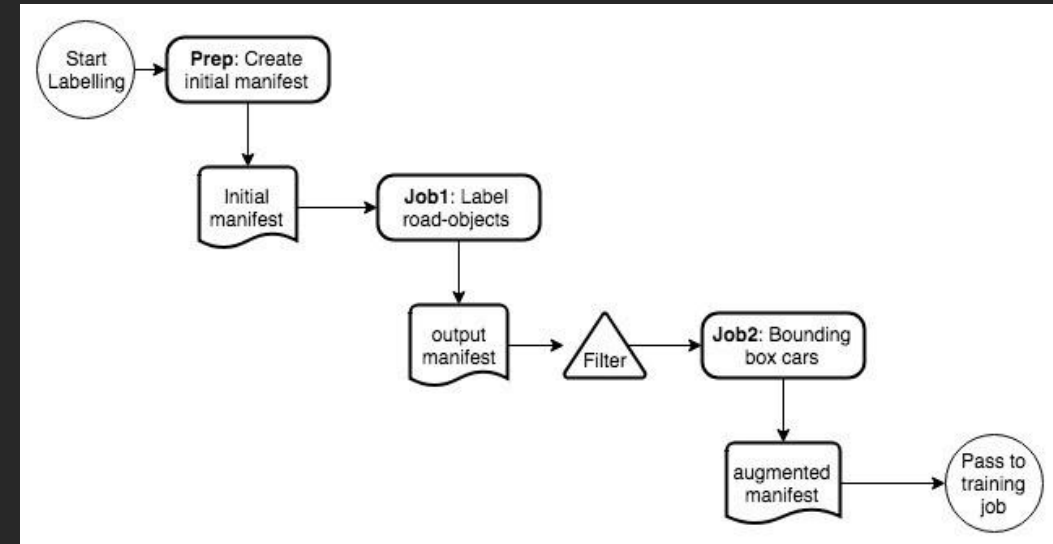
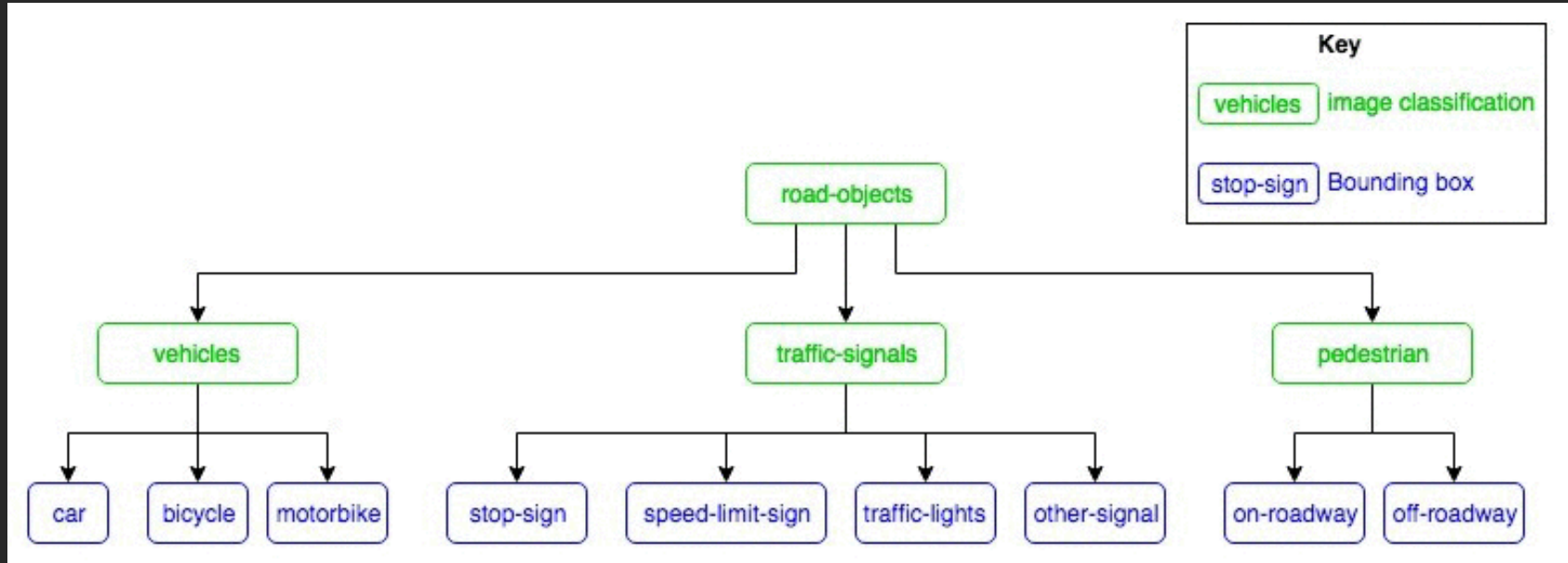
Using SQL to Query Output Data

```
1  {
2    "source-ref": "s3://jsimon-groundtruth-demo/SSDB00001.JPG",
3    "GroundTruthDemo": {
4      "annotations": [
5        {"class_id": 0, "width": 54, "top": 482, "height": 39, "left": 337},
6        {"class_id": 0, "width": 69, "top": 495, "height": 53, "left": 461},
7        {"class_id": 0, "width": 52, "top": 482, "height": 41, "left": 523},
8        {"class_id": 0, "width": 71, "top": 481, "height": 62, "left": 589},
9        {"class_id": 0, "width": 347, "top": 479, "height": 120, "left": 573}
10     ],
11     "image_size": [{"width": 1280, "depth": 3, "height": 960}
12   ],
13 },
14 "GroundTruthDemo-metadata": {
15   "job-name": "labeling-job/groundtruthdemo",
16   "class-map": {"0": "Car"},
17   "human-annotated": "yes",
18   "objects": [
19     {"confidence": 0.94},
20     {"confidence": 0.94},
21     {"confidence": 0.94},
22     {"confidence": 0.94},
23     {"confidence": 0.94}
24   ],
25   "creation-date": "2018-11-26T04:01:09.038134",
26   "type": "groundtruth/object-detection"
27 }
28 }
```

Objective: All images with at least **5** cars present
with confidence score of at least **80**

```
select * from s3object s
where s."demo-full-dataset-2" is not null
and 'Car' in s."demo-full-dataset-2-metadata"."class-map".*
and size(s."demo-full-dataset-2"."annotations") >= 5
and min(s."demo-full-dataset-2 metadata".objects[*].confidence) >= 0.8
```

Hierarchical Taxonomies of Data



Thank you!



Please complete the session
survey in the mobile app.