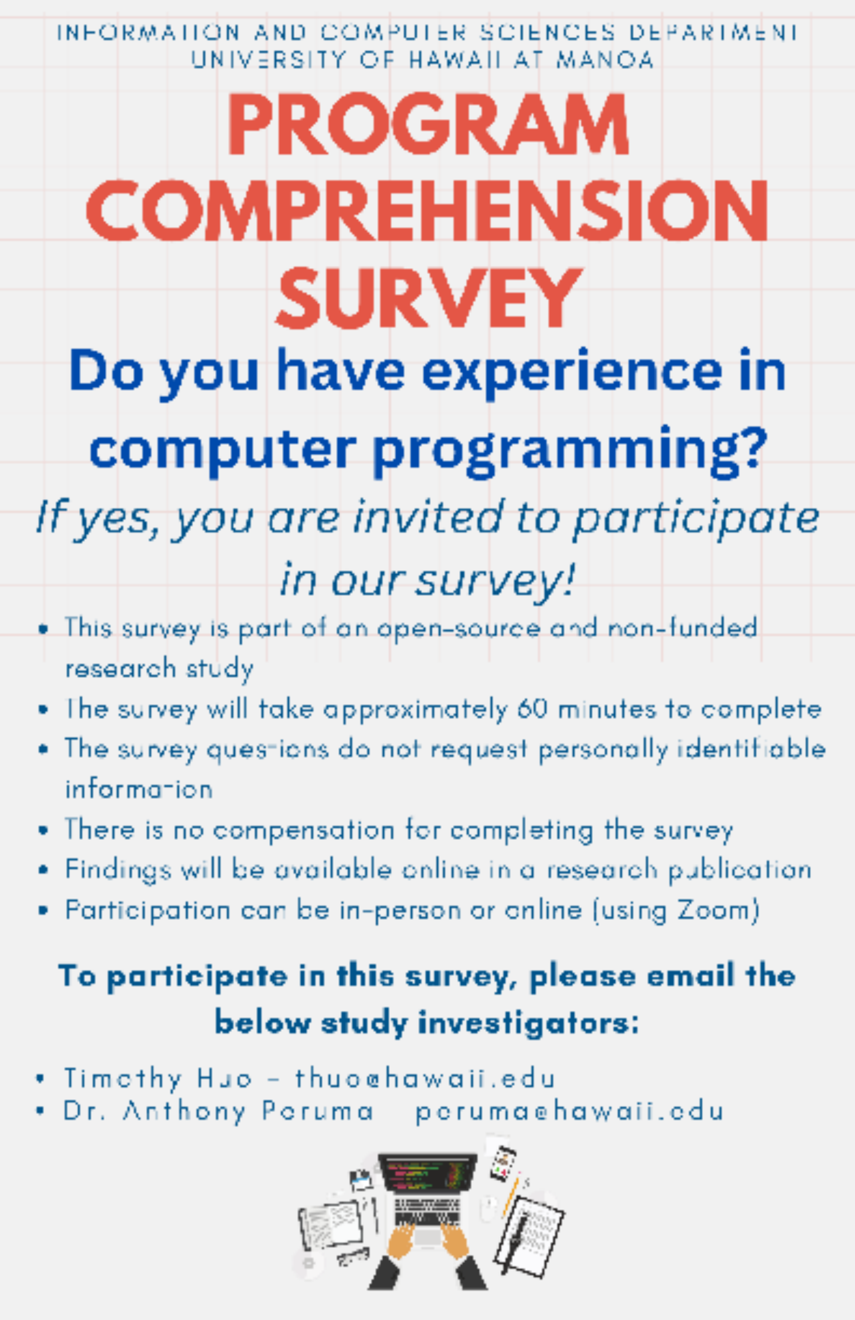
Data Labeling

ICS 491



INFORMATION AND COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF HAWAII AT MANOA

PROGRAM COMPREHENSION SURVEY

Do you have experience in computer programming?

If yes, you are invited to participate in our survey!

- This survey is part of an open-source and non-funded research study
- The survey will take approximately 60 minutes to complete
- The survey questions do not request personally identifiable information
- There is no compensation for completing the survey
- Findings will be available online in a research publication
- Participation can be in-person or online (using Zoom)

To participate in this survey, please email the below study investigators:

- Timothy Huo - thuo@hawaii.edu
- Dr. Anthony Peruma   peruma@hawaii.edu

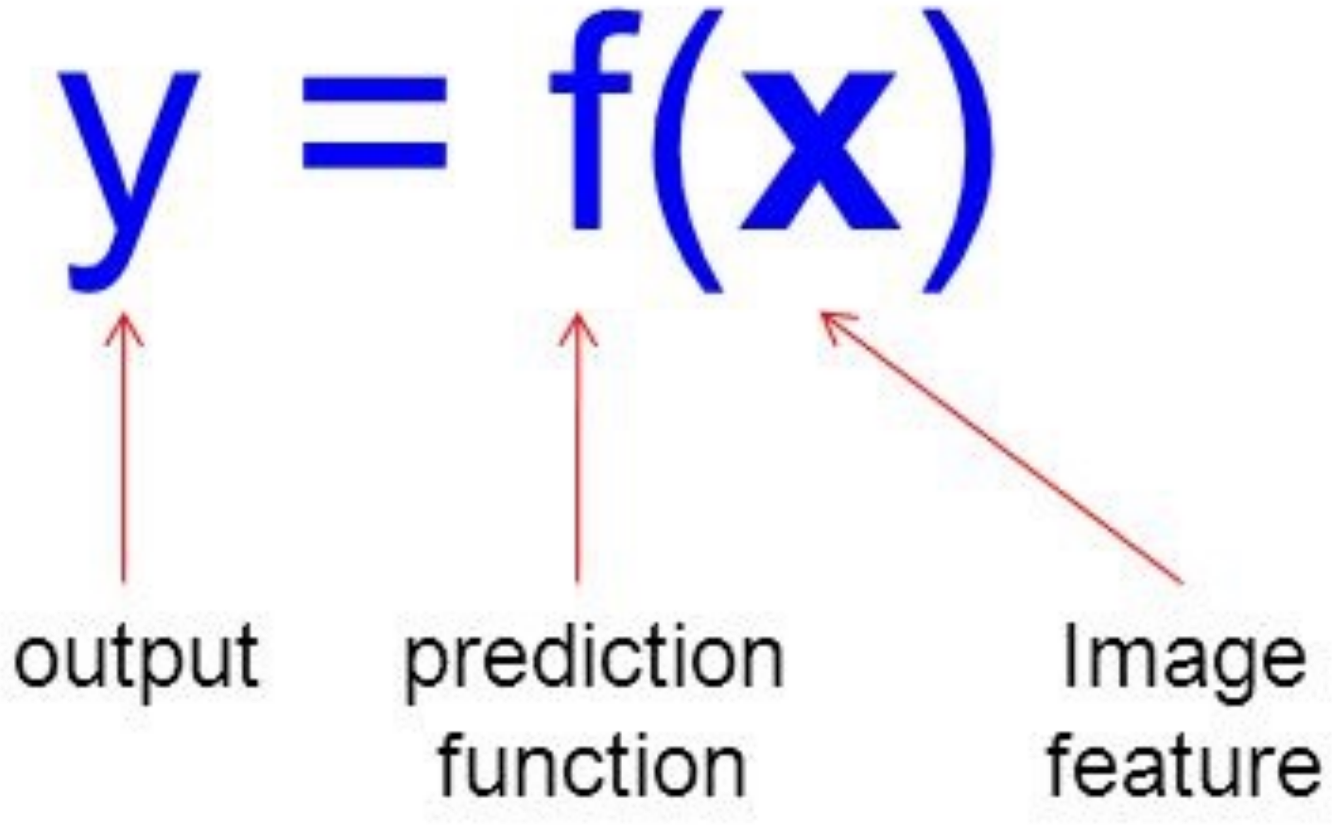
For Final Project Checkin #1 on Sep. 26, be prepared to discuss:

- General project idea
- Specific datasets you will use or recruitment strategy
- How your project goes beyond prior work

I will provide feedback in class for other students to learn. (Roughly 2.5-3 minutes for each student's discussion)

Why do we need data labels?

Machine Learning Setup:



$$y = f(\mathbf{x})$$

output

prediction function

Image feature

In machine learning, we learn from data AND labels

- You need to provide a bunch of input/output pairs to train a machine learning model

- The model learns to predict the correct output based on the input

- We talked about acquiring data last class, but for machine learning, we need to be able to make predictions from those data

Popular Data Labeling Platforms

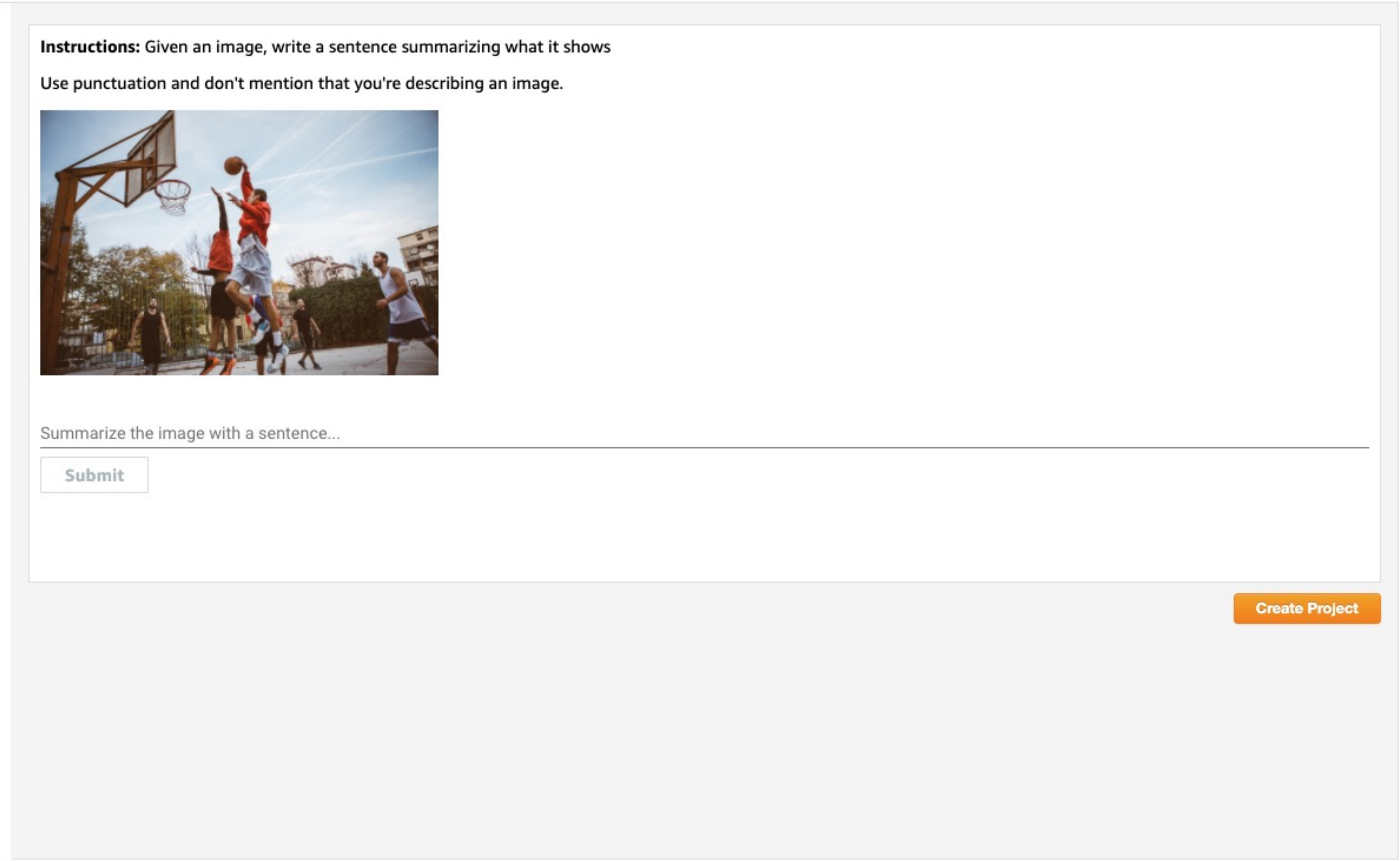
- Amazon Mechanical Turk
- Snorkel
- Labelbox
- Hive
- …

Paid Data Labeling

Amazon Mechanical Turk

Crowdsourcing platform

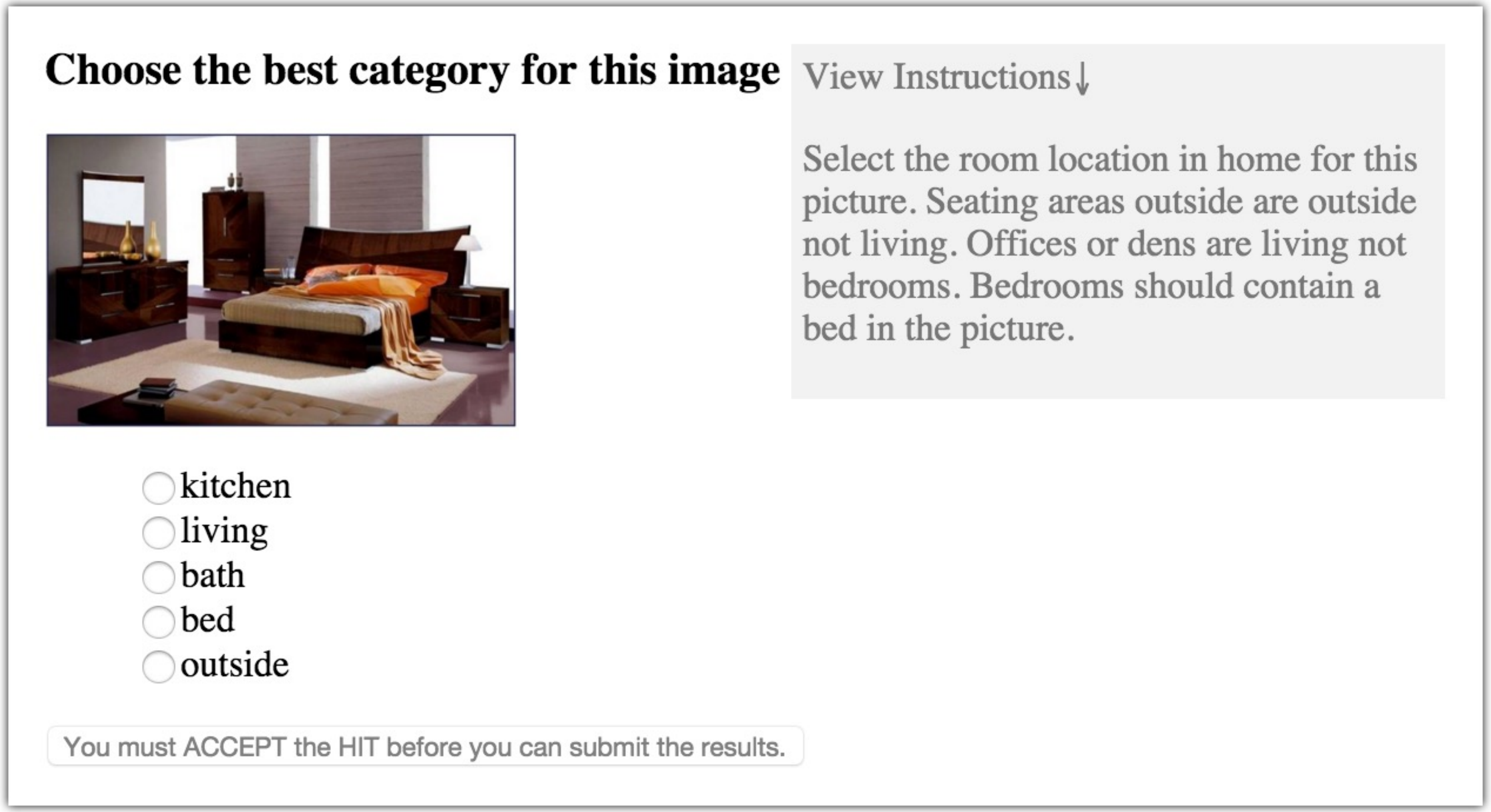
Amazon Mechanical Turk: Example Task



Instructions: Given an image, write a sentence summarizing what it shows

Use punctuation and don't mention that you're describing an image.

Amazon Mechanical Turk: Example Task



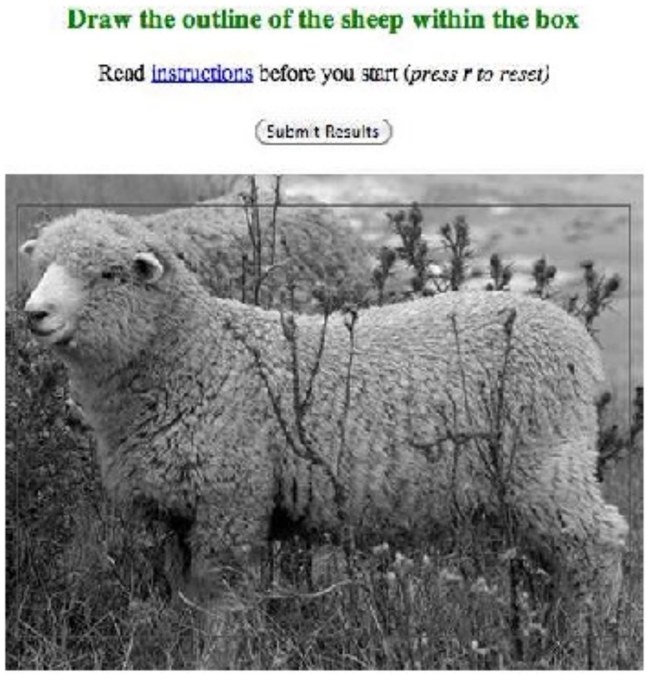
Choose the best category for this image

View Instructions↓

Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.

○ kitchen
○ living
○ bath
○ bed
○ outside

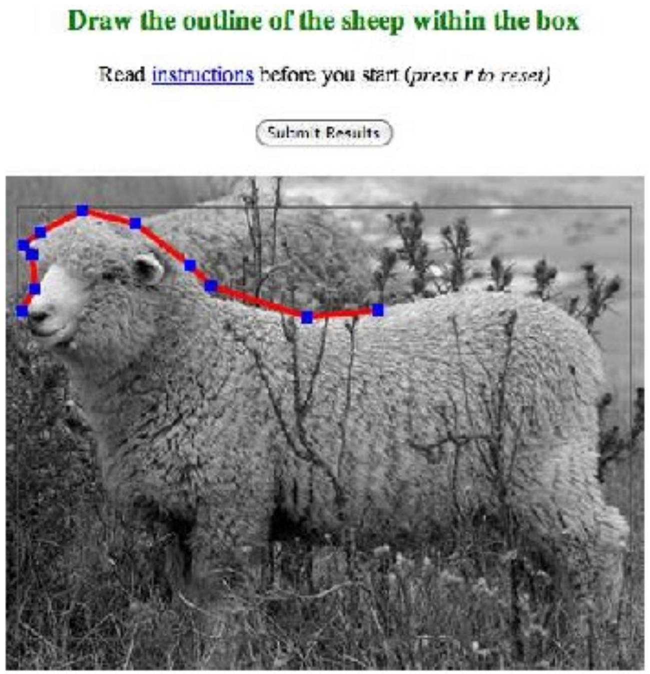
Amazon Mechanical Turk: Example Task



Draw the outline of the sheep within the box

Read instructions before you start (*press r to reset*)

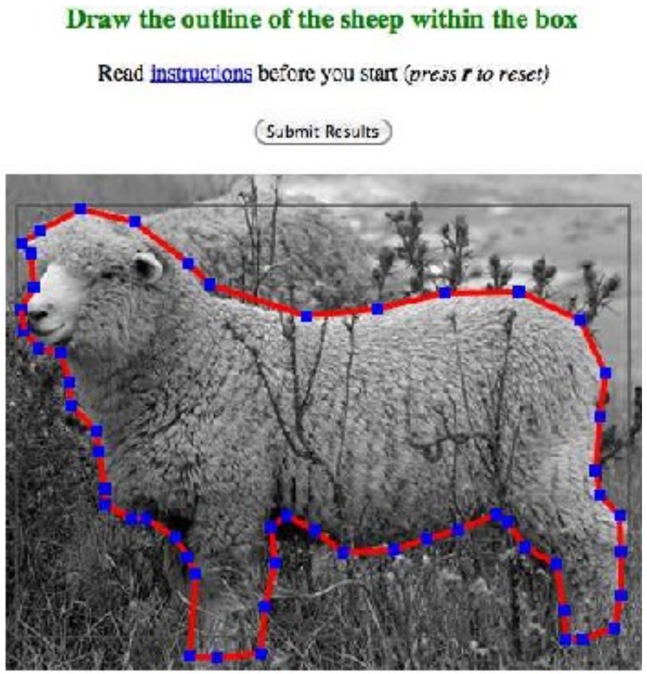
(a)



Draw the outline of the sheep within the box

Read instructions before you start (*press r to reset*)

(b)

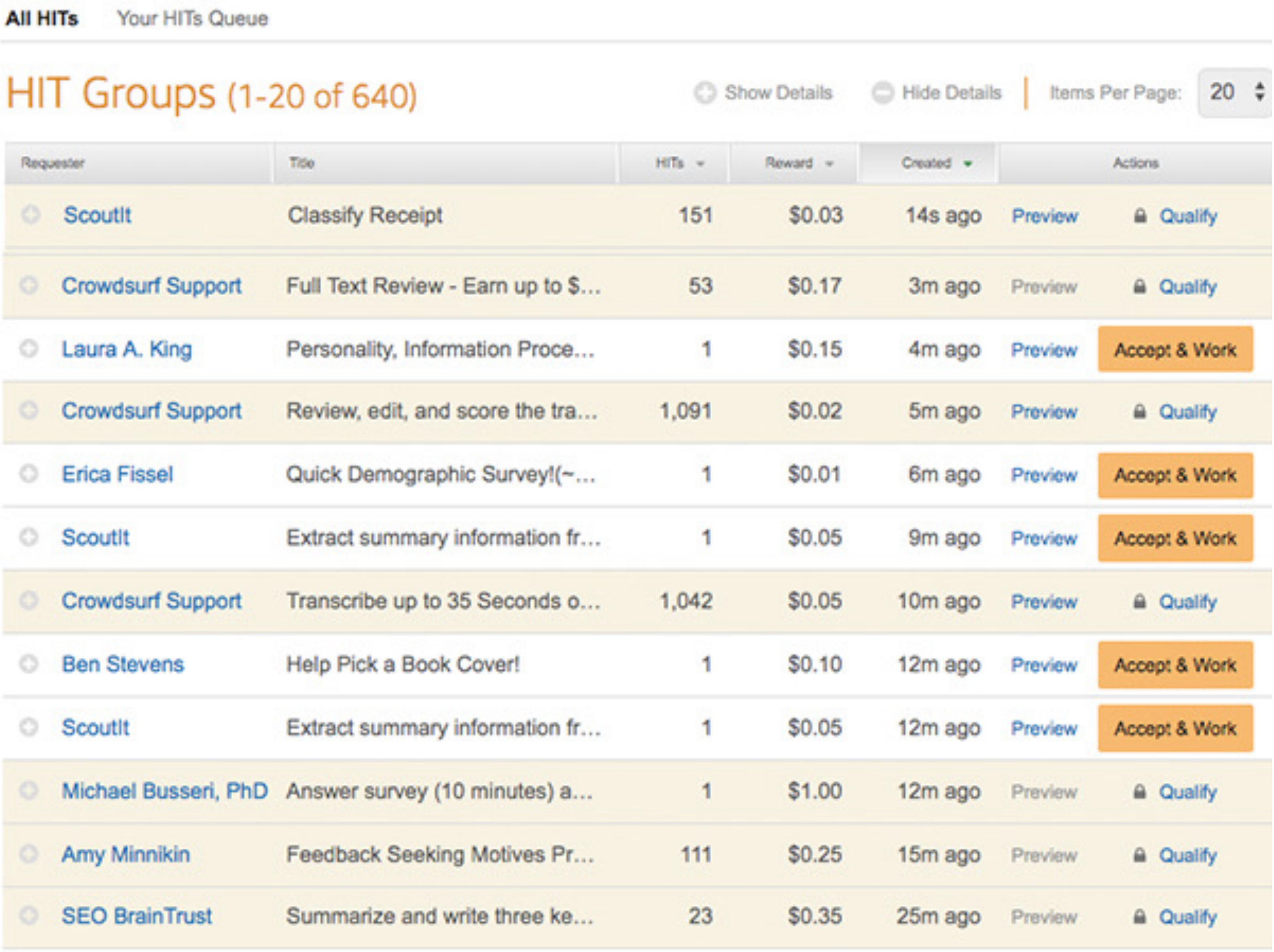


Draw the outline of the sheep within the box

Read instructions before you start (*press r to reset*)

(c)

Amazon Mechanical Turk



All HITs Your HITs Queue

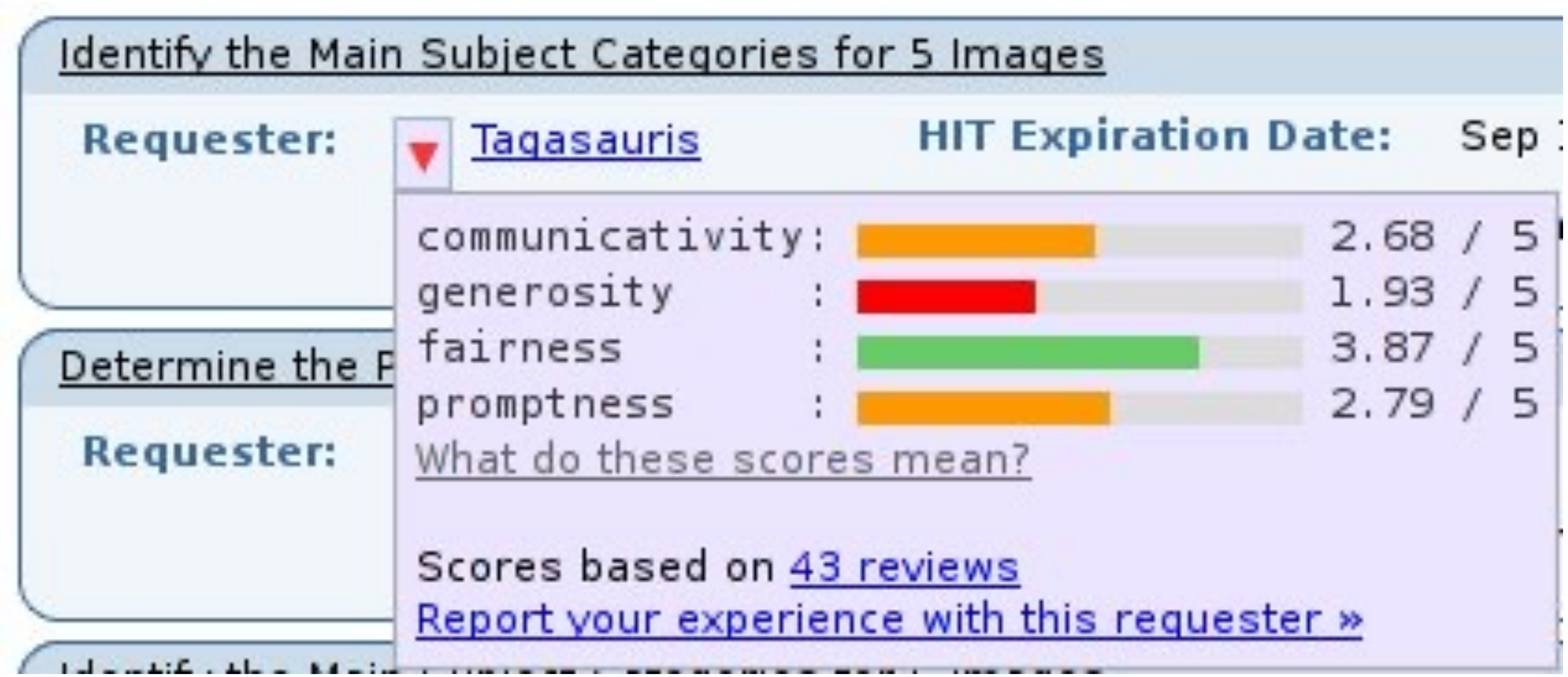
HIT Groups (1-20 of 640)

Show Details | Hide Details | Items Per Page: 20

Requester	Title	HITs	Reward	Created	Actions	
ScoutIt	Classify Receipt	151	$0.03	14s ago	Preview	🔒 Qualify
Crowdsurf Support	Full Text Review - Earn up to $…	53	$0.17	3m ago	Preview	🔒 Qualify
Laura A. King	Personality, Information Proce…	1	$0.15	4m ago	Preview	Accept & Work
Crowdsurf Support	Review, edit, and score the tra…	1,091	$0.02	5m ago	Preview	🔒 Qualify
Erica Fissel	Quick Demographic Survey!(~…	1	$0.01	6m ago	Preview	Accept & Work
ScoutIt	Extract summary information fr…	1	$0.05	9m ago	Preview	Accept & Work
Crowdsurf Support	Transcribe up to 35 Seconds o…	1,042	$0.05	10m ago	Preview	🔒 Qualify
Ben Stevens	Help Pick a Book Cover!	1	$0.10	12m ago	Preview	Accept & Work
ScoutIt	Extract summary information fr…	1	$0.05	12m ago	Preview	Accept & Work
Michael Busseri, PhD	Answer survey (10 minutes) a…	1	$1.00	12m ago	Preview	🔒 Qualify
Amy Minnikin	Feedback Seeking Motives Pr…	111	$0.25	15m ago	Preview	🔒 Qualify
SEO BrainTrust	Summarize and write three ke…	23	$0.35	25m ago	Preview	🔒 Qualify

Amazon Mechanical Turk: Worker Rights

Platforms like Turkopticon allow for rating Requesters rather than Workers



Identify the Main Subject Categories for 5 Images

Requester: ▼ Tagasauris HIT Expiration Date: Sep

Determine the P

Requester:

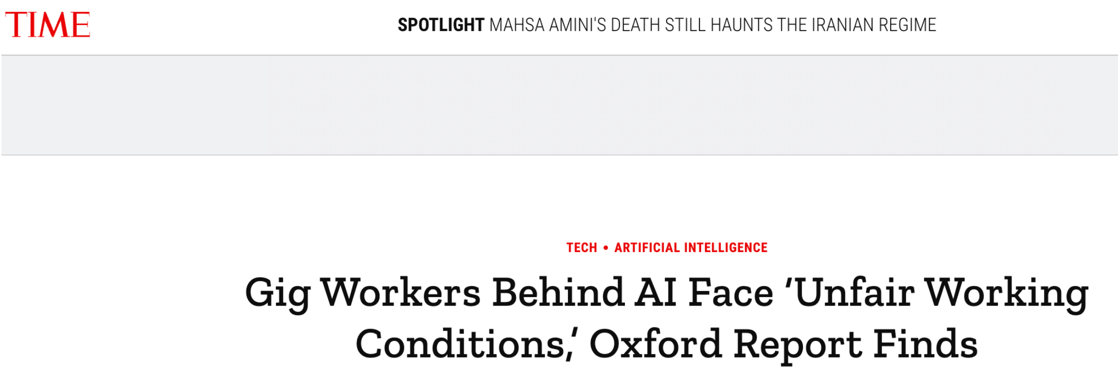
communicativity:	▬	2.68 / 5
generosity :	▬	1.93 / 5
fairness :	▬	3.87 / 5
promptness :	▬	2.79 / 5

What do these scores mean?

Scores based on 43 reviews

Report your experience with this requester »

Amazon Mechanical Turk: Worker Rights

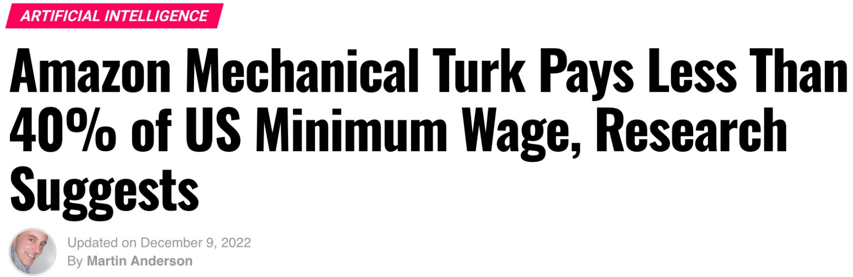


TIME

SPOTLIGHT MAHSA AMINI'S DEATH STILL HAUNTS THE IRANIAN REGIME

TECH • ARTIFICIAL INTELLIGENCE

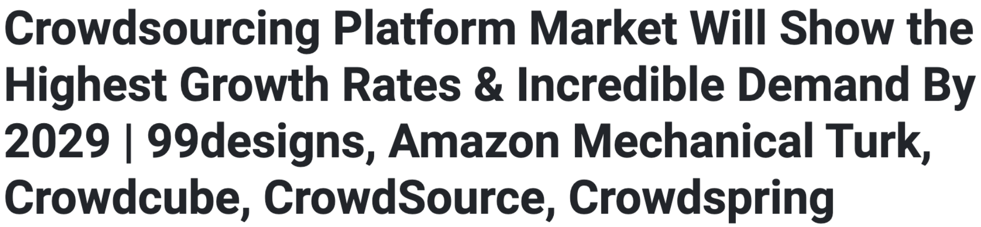
Gig Workers Behind AI Face 'Unfair Working Conditions,' Oxford Report Finds



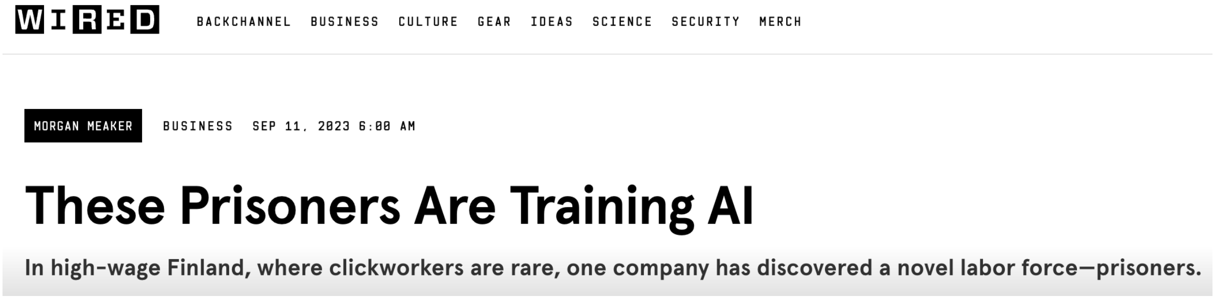
ARTIFICIAL INTELLIGENCE

Amazon Mechanical Turk Pays Less Than 40% of US Minimum Wage, Research Suggests

Updated on December 9, 2022
By Martin Anderson



Crowdsourcing Platform Market Will Show the Highest Growth Rates & Incredible Demand By 2029 | 99designs, Amazon Mechanical Turk, Crowdcube, CrowdSource, Crowdspring

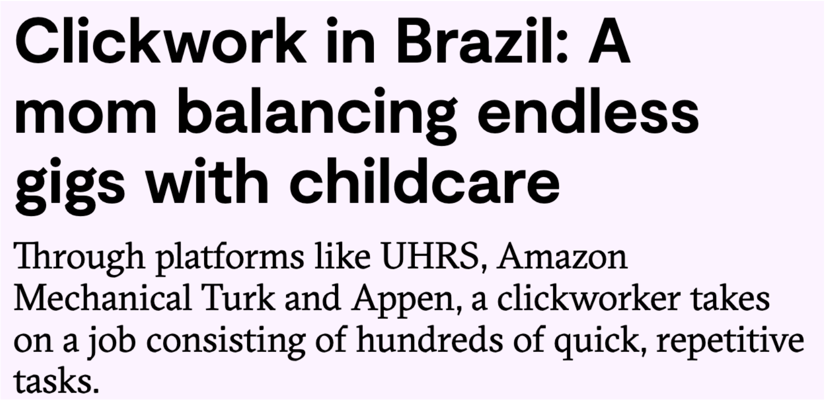


WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY MERCH

MORGAN MEAKER BUSINESS SEP 11, 2023 6:00 AM

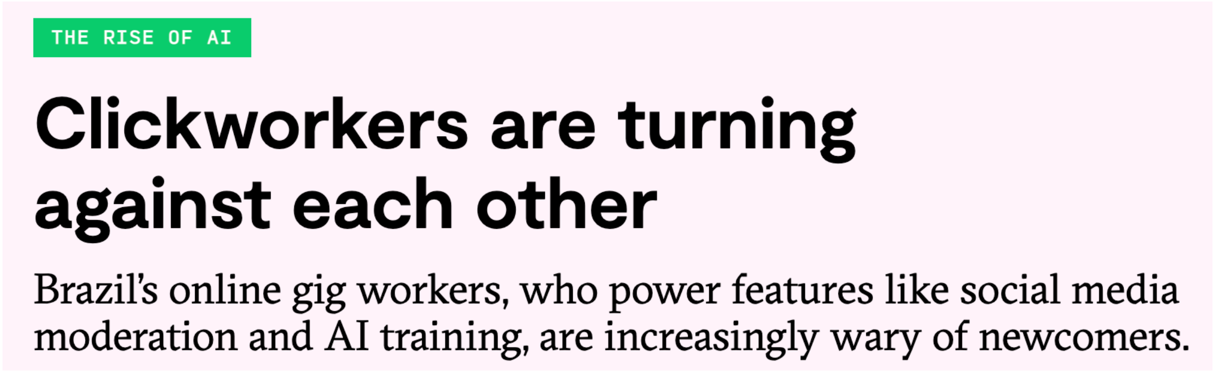
These Prisoners Are Training AI

In high-wage Finland, where clickworkers are rare, one company has discovered a novel labor force—prisoners.



Clickwork in Brazil: A mom balancing endless gigs with childcare

Through platforms like UHRS, Amazon Mechanical Turk and Appen, a clickworker takes on a job consisting of hundreds of quick, repetitive tasks.



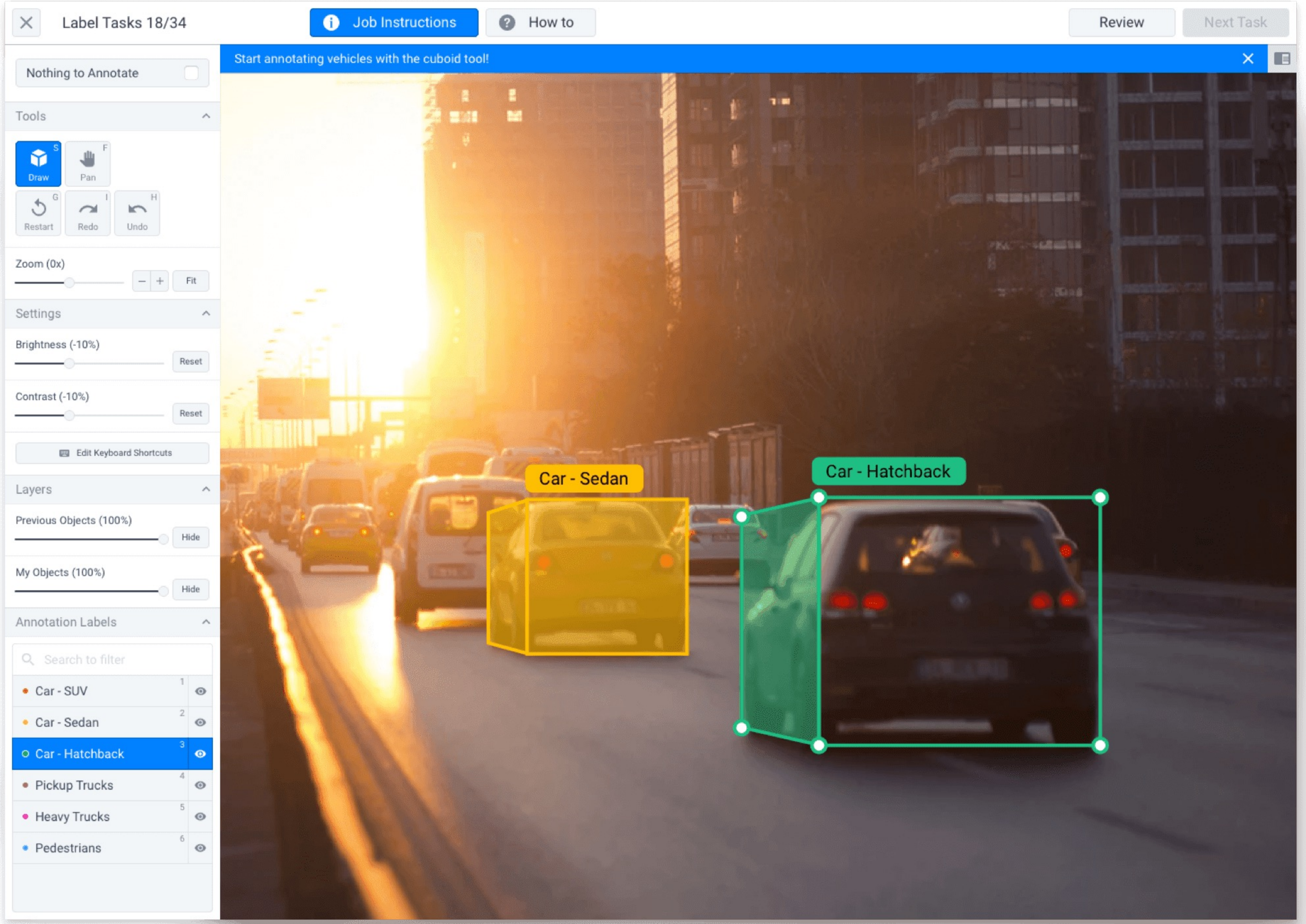
THE RISE OF AI

Clickworkers are turning against each other

Brazil's online gig workers, who power features like social media moderation and AI training, are increasingly wary of newcomers.

Other Platforms: Hive

HIVE



Start annotating vehicles with the cuboid tool!

Zoom (0x)

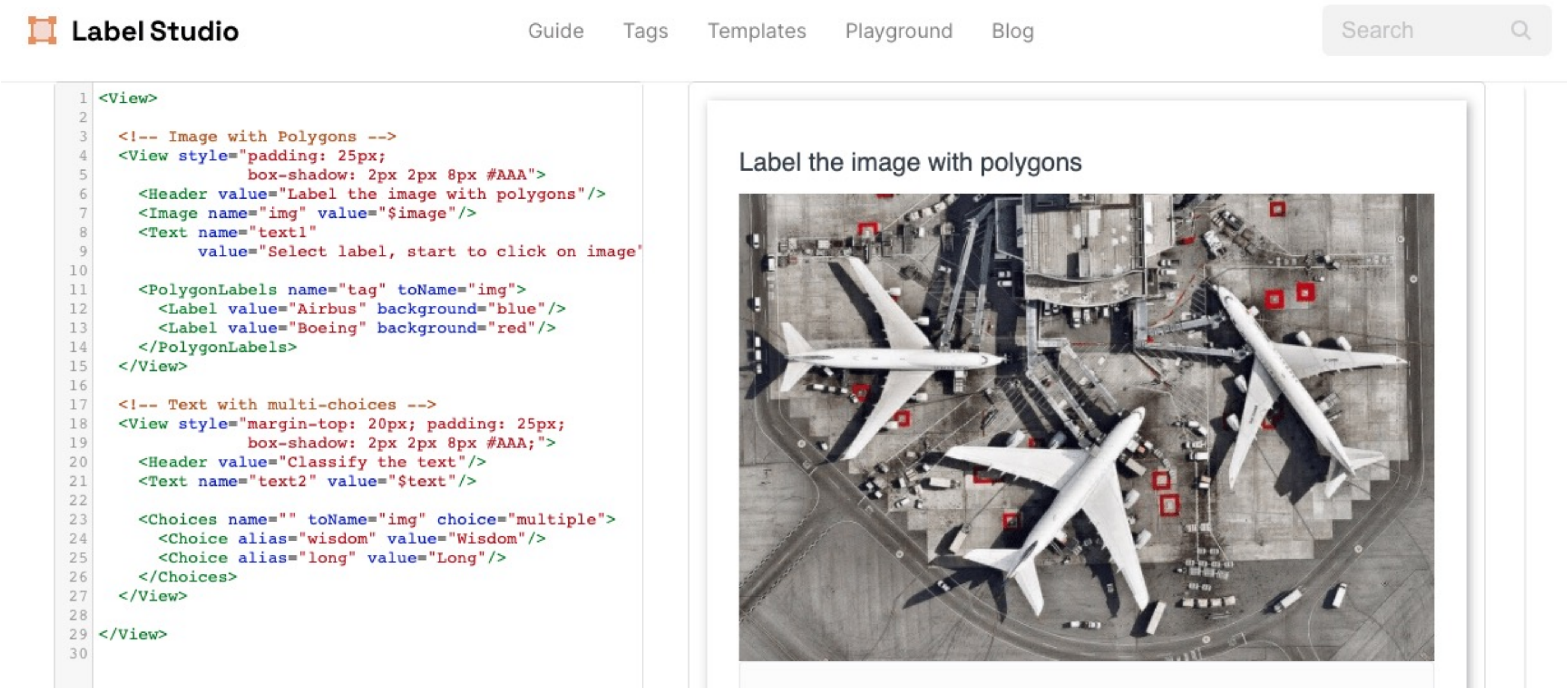
Brightness (-10%)

Contrast (-10%)

Previous Objects (100%)

My Objects (100%)

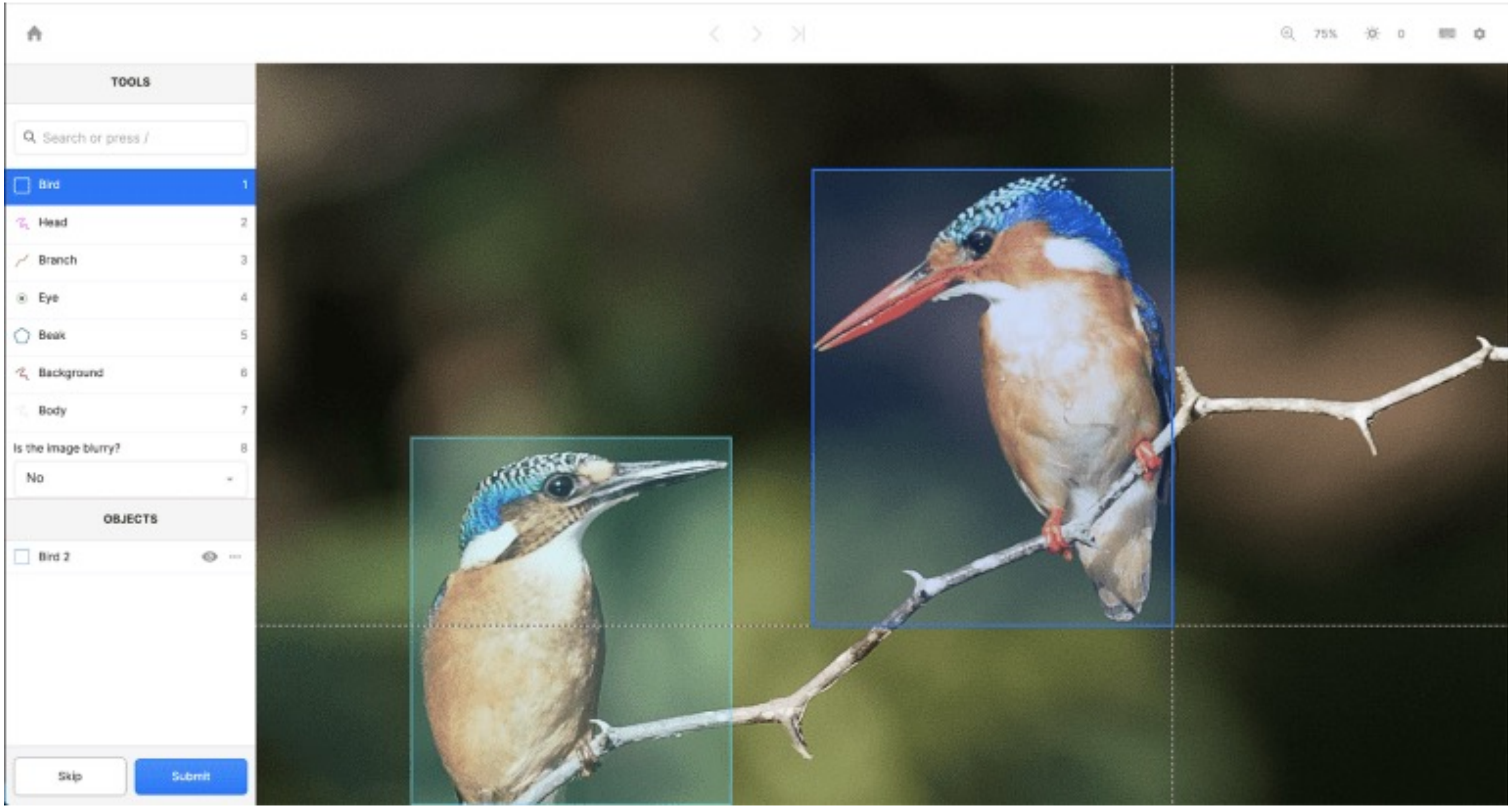
Other Platforms: Label Studio



```
<View>

  <!-- Image with Polygons -->
  <View style="padding: 25px;
               box-shadow: 2px 2px 8px #AAA">
    <Header value="Label the image with polygons"/>
    <Image name="img" value="$image"/>
    <Text name="text1"
          value="Select label, start to click on image'

    <PolygonLabels name="tag" toName="img">
      <Label value="Airbus" background="blue"/>
      <Label value="Boeing" background="red"/>
    </PolygonLabels>
  </View>

  <!-- Text with multi-choices -->
  <View style="margin-top: 20px; padding: 25px;
               box-shadow: 2px 2px 8px #AAA;">
    <Header value="Classify the text"/>
    <Text name="text2" value="$text"/>

    <Choices name="" toName="img" choice="multiple">
      <Choice alias="wisdom" value="Wisdom"/>
      <Choice alias="long" value="Long"/>
    </Choices>
  </View>

</View>
```

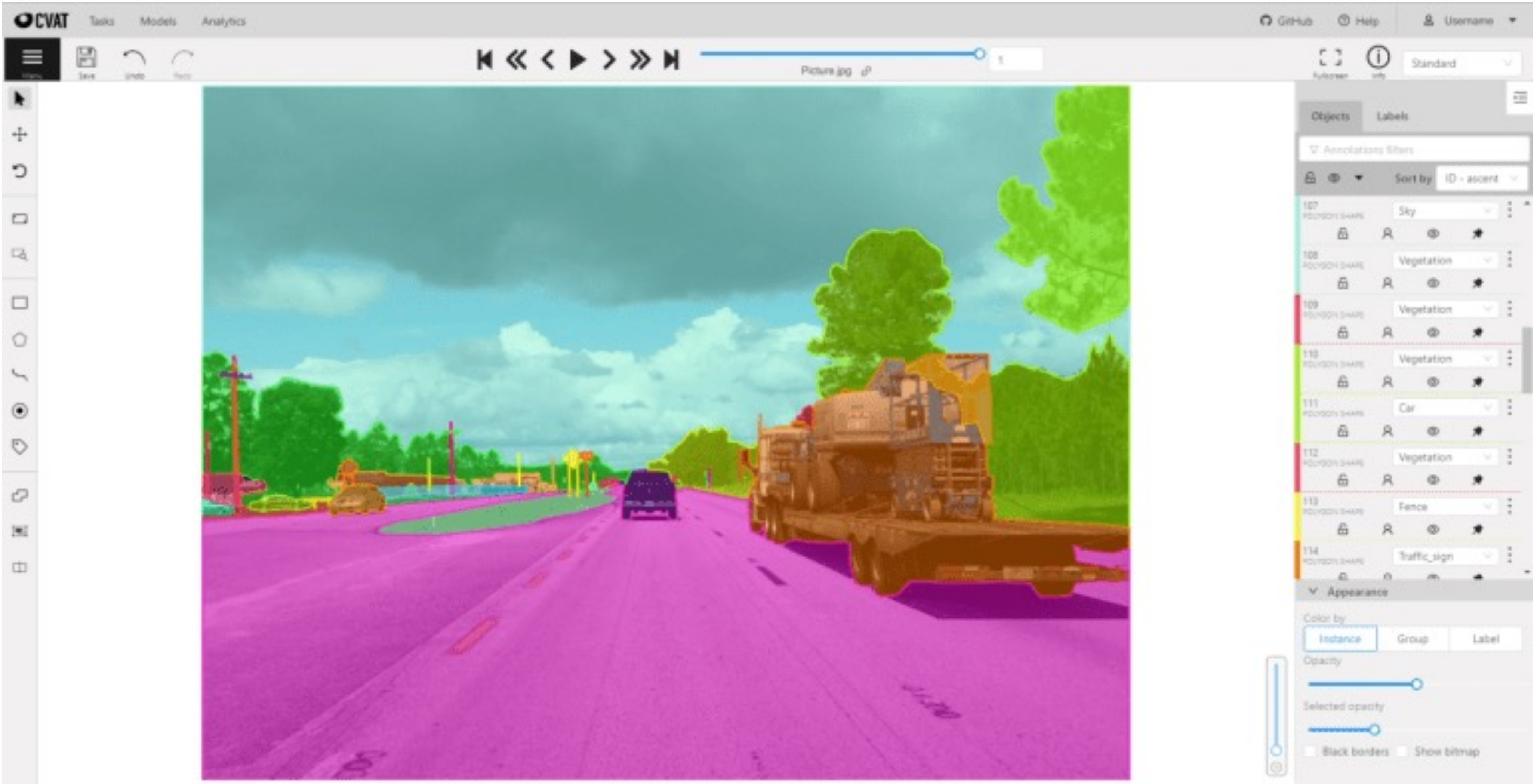
Label the image with polygons

Other Platforms: Label Box



Is the image blurry? 8
No

Other Platforms: CVAT



107 POLYGON SHAPE Sky

108 POLYGON SHAPE Vegetation

109 POLYGON SHAPE Vegetation

110 POLYGON SHAPE Vegetation

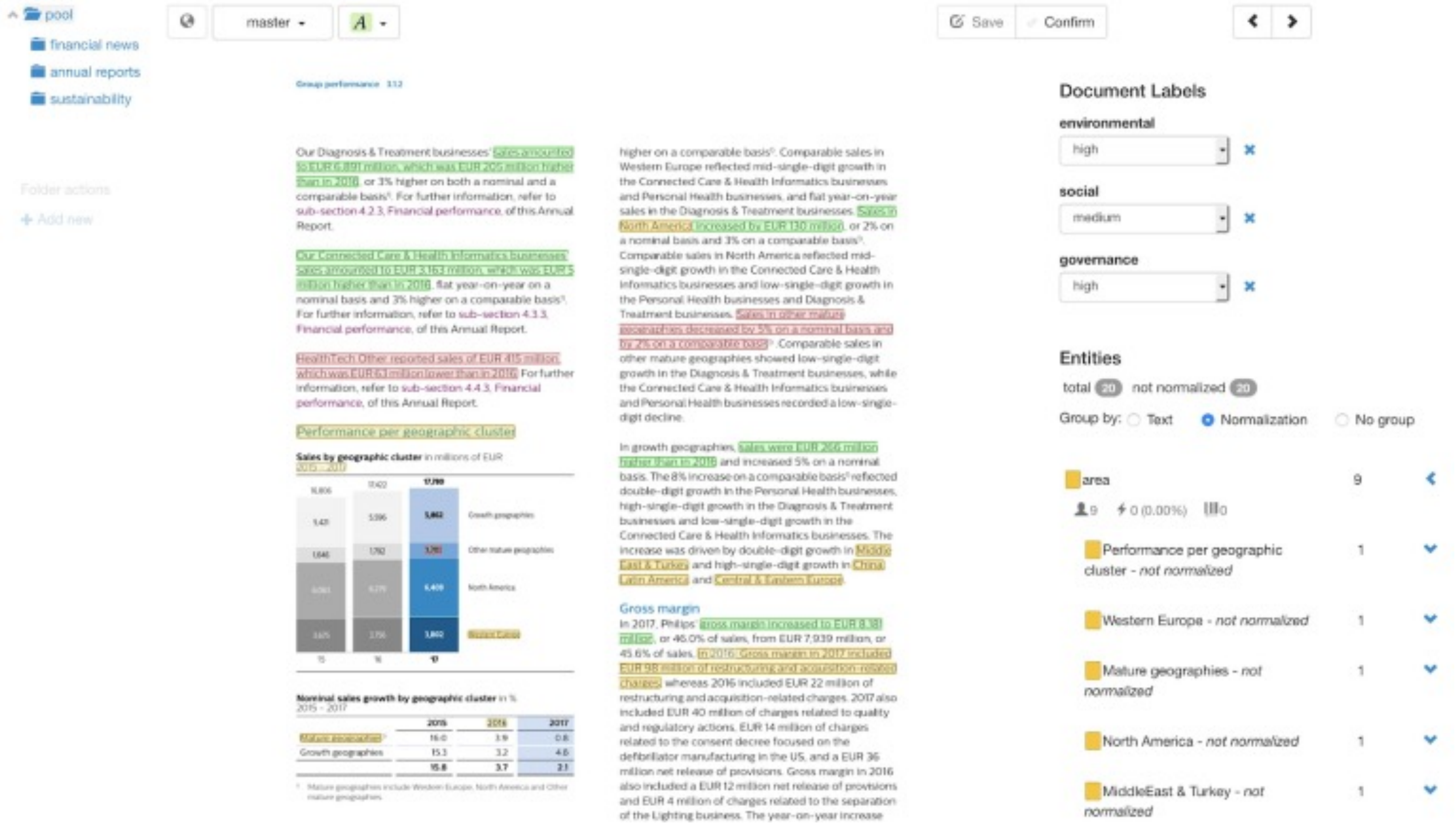
111 POLYGON SHAPE Car

112 POLYGON SHAPE Vegetation

113 POLYGON SHAPE Fence

114 POLYGON SHAPE Traffic_sign

Other Platforms: Tag Tot



Group performance 112

Our Diagnosis & Treatment businesses' sales amounted to EUR 6,891 million, which was EUR 205 million higher than in 2016, or 3% higher on both a nominal and a comparable basis[1]. For further information, refer to sub-section 4.2.3, Financial performance, of this Annual Report.

Our Connected Care & Health Informatics businesses' sales amounted to EUR 3,163 million, which was EUR 5 million higher than in 2016, flat year-on-year on a nominal basis and 3% higher on a comparable basis[1]. For further information, refer to sub-section 4.3.3, Financial performance, of this Annual Report.

HealthTech Other reported sales of EUR 415 million, which was EUR 63 million lower than in 2016. For further information, refer to sub-section 4.4.3, Financial performance, of this Annual Report.

Performance per geographic cluster

Sales by geographic cluster in millions of EUR
2015 - 2017

Nominal sales growth by geographic cluster in %
2015 - 2017

	2015	2016	2017
Mature geographies[1]	16.0	3.9	0.8
Growth geographies	15.3	3.2	4.6
	15.8	3.7	2.1

[1] Mature geographies include Western Europe, North America and Other mature geographies

higher on a comparable basis[1]. Comparable sales in Western Europe reflected mid-single-digit growth in the Connected Care & Health Informatics businesses and Personal Health businesses, and flat year-on-year sales in the Diagnosis & Treatment businesses. Sales in North America increased by EUR 130 million, or 2% on a nominal basis and 3% on a comparable basis[1]. Comparable sales in North America reflected mid-single-digit growth in the Connected Care & Health Informatics businesses and low-single-digit growth in the Personal Health businesses and Diagnosis & Treatment businesses. Sales in other mature geographies decreased by 5% on a nominal basis and by 2% on a comparable basis[1]. Comparable sales in other mature geographies showed low-single-digit growth in the Diagnosis & Treatment businesses, while the Connected Care & Health Informatics businesses and Personal Health businesses recorded a low-single-digit decline.

In growth geographies, sales were EUR 266 million higher than in 2016 and increased 5% on a nominal basis. The 8% increase on a comparable basis[1] reflected double-digit growth in the Personal Health businesses, high-single-digit growth in the Diagnosis & Treatment businesses and low-single-digit growth in the Connected Care & Health Informatics businesses. The increase was driven by double-digit growth in Middle East & Turkey and high-single-digit growth in China, Latin America and Central & Eastern Europe.

Gross margin

In 2017, Philips' gross margin increased to EUR 8,181 million, or 46.0% of sales, from EUR 7,939 million, or 45.6% of sales, in 2016. Gross margin in 2017 included EUR 98 million of restructuring and acquisition-related charges, whereas 2016 included EUR 22 million of restructuring and acquisition-related charges. 2017 also included EUR 40 million of charges related to quality and regulatory actions, EUR 14 million of charges related to the consent decree focused on the defibrillator manufacturing in the US, and a EUR 36 million net release of provisions. Gross margin in 2016 also included a EUR 12 million net release of provisions and EUR 4 million of charges related to the separation of the Lighting business. The year-on-year increase

Document Labels

environmental: high ✖

social: medium ✖

governance: high ✖

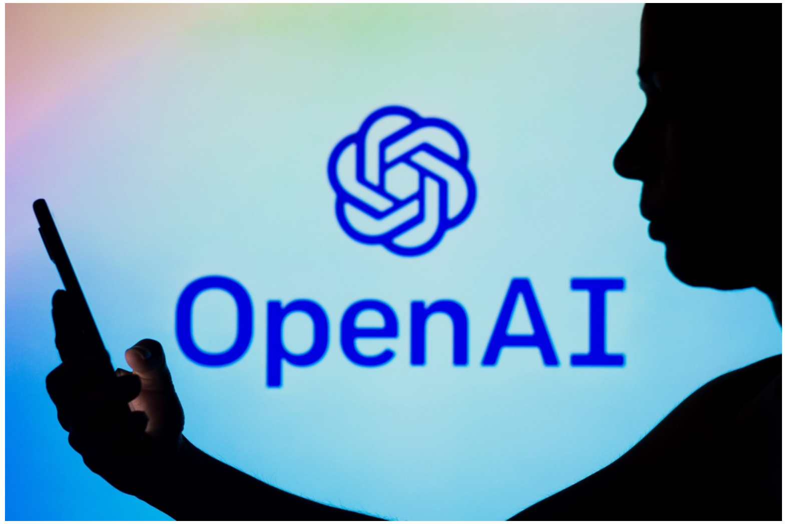
Entities

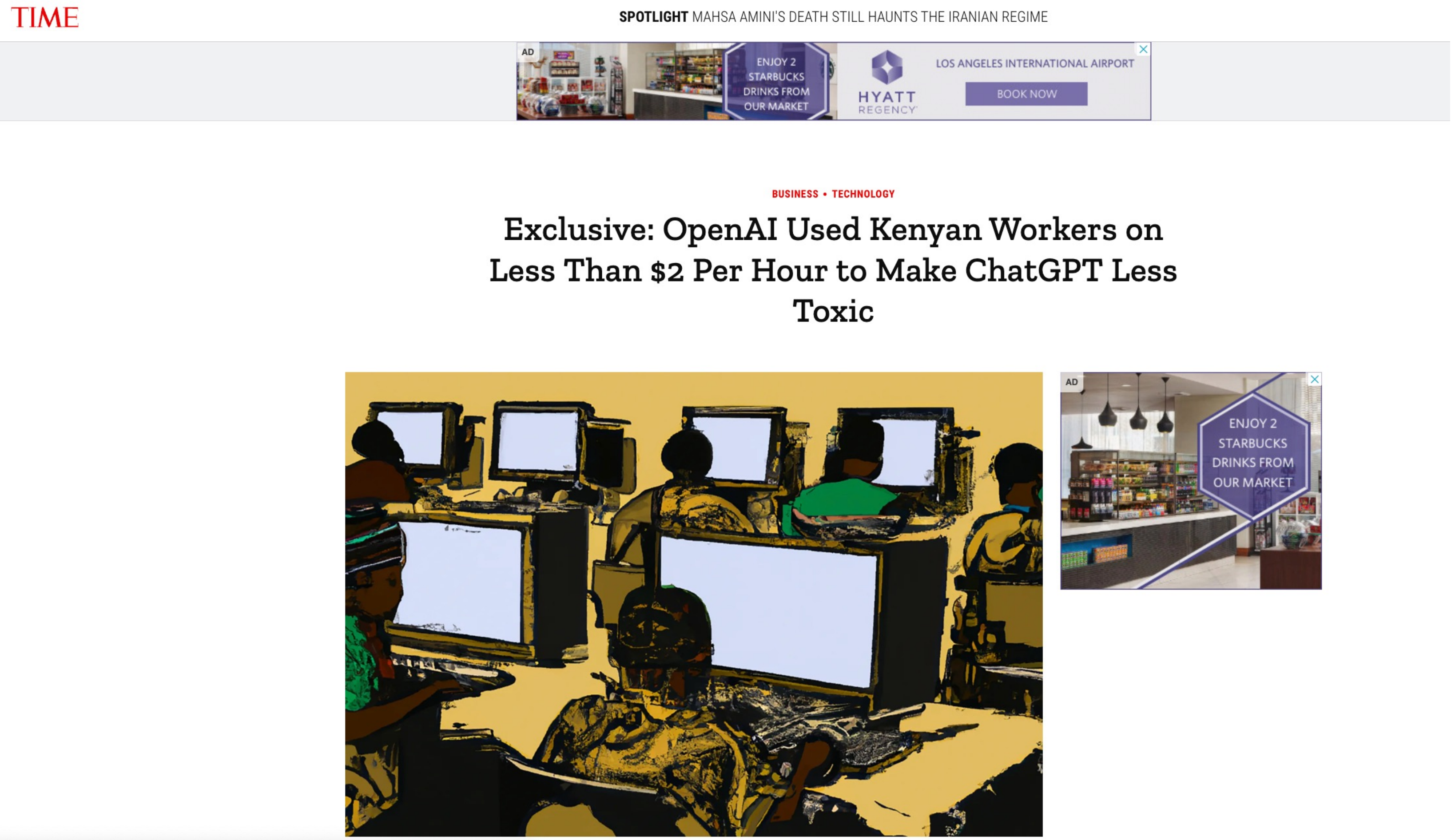
total 20 not normalized 20

Group by: ○ Text ● Normalization ○ No group

🟨 area	9	‹
👤9 ⚡0 (0.00%) ▥0		
🟨 Performance per geographic cluster - *not normalized*	1	⌄
🟨 Western Europe - *not normalized*	1	⌄
🟨 Mature geographies - *not normalized*	1	⌄
🟨 North America - *not normalized*	1	⌄
🟨 MiddleEast & Turkey - *not normalized*	1	⌄

Another Solution: Hire Your Own Trained Labelers



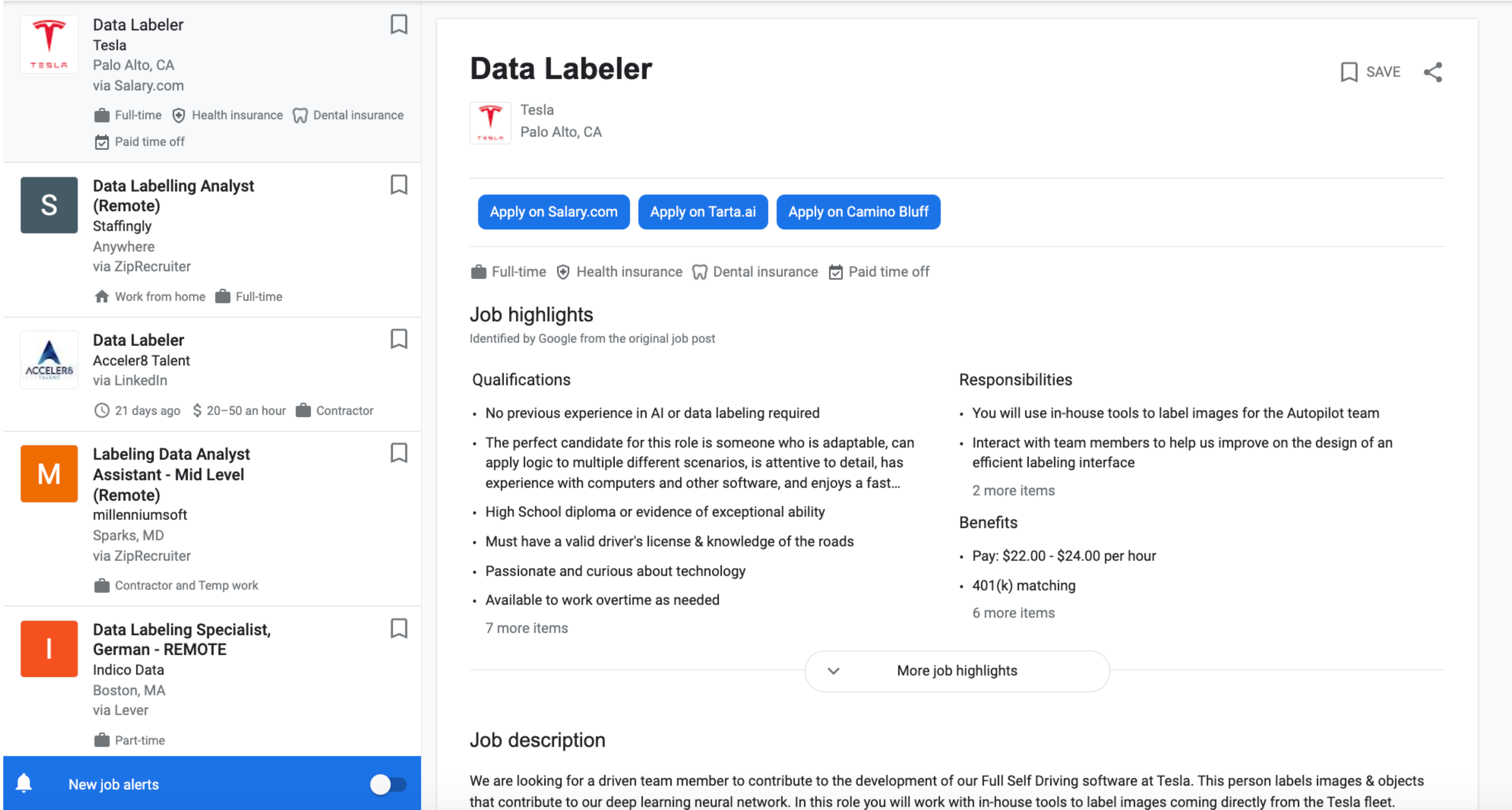

TIME

SPOTLIGHT MAHSA AMINI'S DEATH STILL HAUNTS THE IRANIAN REGIME

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

Another Solution: Hire Your Own Trained Labelers



Data Labeler
Tesla
Palo Alto, CA
via Salary.com
Full-time · Health insurance · Dental insurance · Paid time off

Data Labelling Analyst (Remote)
Staffingly
Anywhere
via ZipRecruiter
Work from home · Full-time

Data Labeler
Acceler8 Talent
via LinkedIn
21 days ago · 20–50 an hour · Contractor

Labeling Data Analyst Assistant - Mid Level (Remote)
millenniumsoft
Sparks, MD
via ZipRecruiter
Contractor and Temp work

Data Labeling Specialist, German - REMOTE
Indico Data
Boston, MA
via Lever
Part-time

New job alerts

Data Labeler

Tesla
Palo Alto, CA

Full-time · Health insurance · Dental insurance · Paid time off

Job highlights

Identified by Google from the original job post

Qualifications

- No previous experience in AI or data labeling required
- The perfect candidate for this role is someone who is adaptable, can apply logic to multiple different scenarios, is attentive to detail, has experience with computers and other software, and enjoys a fast...
- High School diploma or evidence of exceptional ability
- Must have a valid driver's license & knowledge of the roads
- Passionate and curious about technology
- Available to work overtime as needed

7 more items

Responsibilities

- You will use in-house tools to label images for the Autopilot team
- Interact with team members to help us improve on the design of an efficient labeling interface

2 more items

Benefits

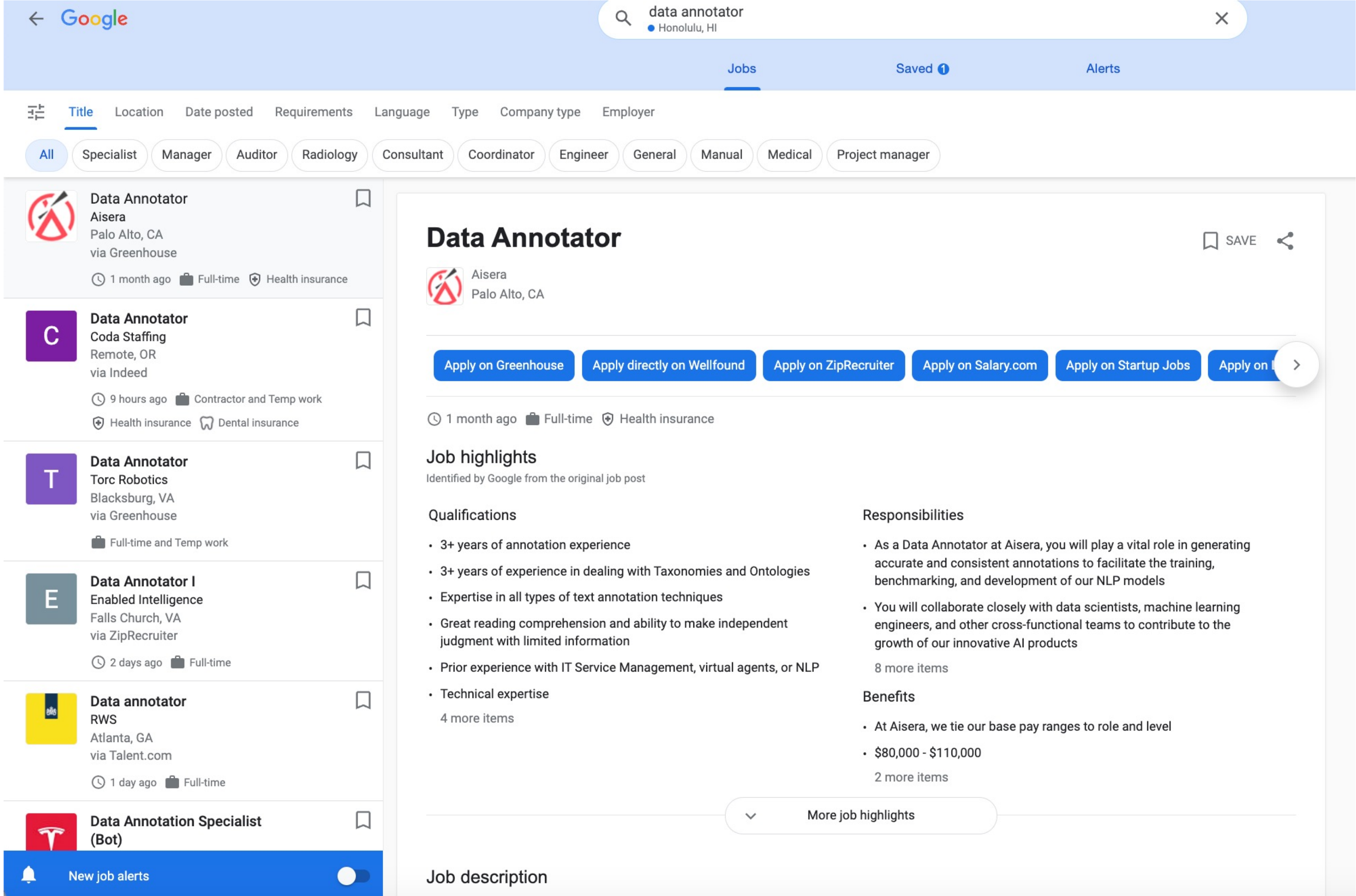
- Pay: $22.00 - $24.00 per hour
- 401(k) matching

6 more items

Job description

We are looking for a driven team member to contribute to the development of our Full Self Driving software at Tesla. This person labels images & objects that contribute to our deep learning neural network. In this role you will work with in-house tools to label images coming directly from the Tesla fleet.

Another Solution: Hire Your Own Trained Labelers



Data Annotator
Aisera
Palo Alto, CA
via Greenhouse
1 month ago · Full-time · Health insurance

Data Annotator
Coda Staffing
Remote, OR
via Indeed
9 hours ago · Contractor and Temp work
Health insurance · Dental insurance

Data Annotator
Torc Robotics
Blacksburg, VA
via Greenhouse
Full-time and Temp work

Data Annotator I
Enabled Intelligence
Falls Church, VA
via ZipRecruiter
2 days ago · Full-time

Data annotator
RWS
Atlanta, GA
via Talent.com
1 day ago · Full-time

Data Annotation Specialist (Bot)

Data Annotator

Aisera
Palo Alto, CA

1 month ago · Full-time · Health insurance

Job highlights

Identified by Google from the original job post

Qualifications

- 3+ years of annotation experience
- 3+ years of experience in dealing with Taxonomies and Ontologies
- Expertise in all types of text annotation techniques
- Great reading comprehension and ability to make independent judgment with limited information
- Prior experience with IT Service Management, virtual agents, or NLP
- Technical expertise

4 more items

Responsibilities

- As a Data Annotator at Aisera, you will play a vital role in generating accurate and consistent annotations to facilitate the training, benchmarking, and development of our NLP models
- You will collaborate closely with data scientists, machine learning engineers, and other cross-functional teams to contribute to the growth of our innovative AI products

8 more items

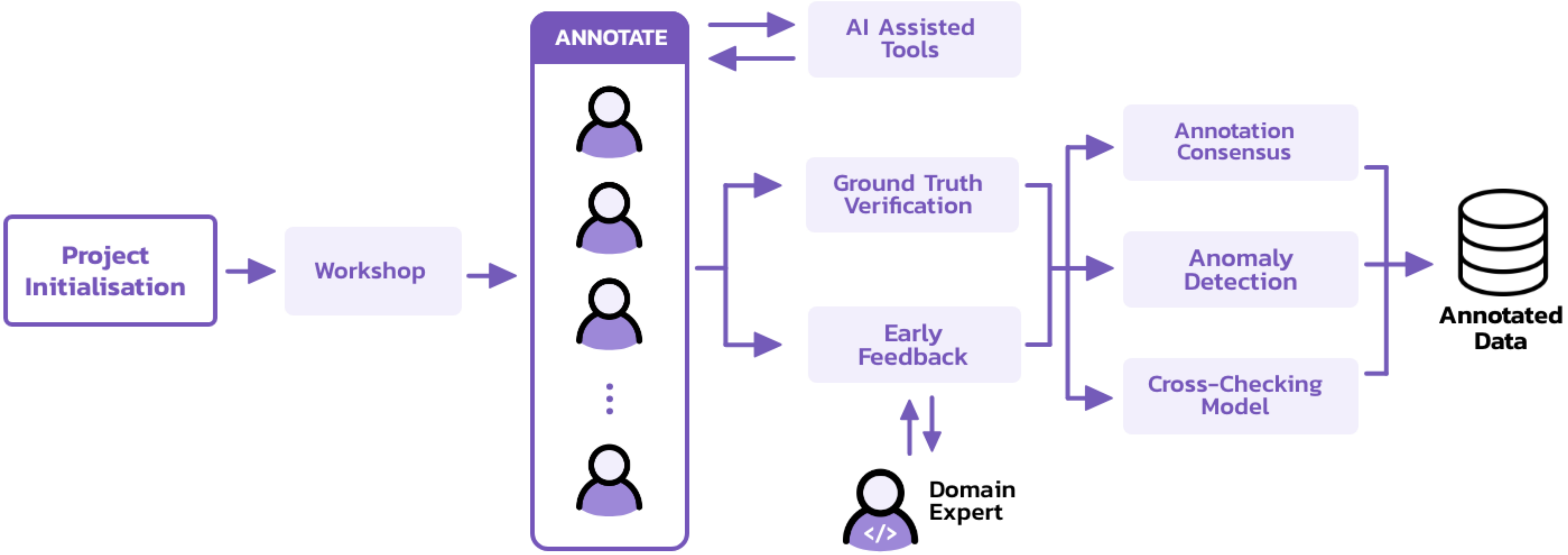
Benefits

- At Aisera, we tie our base pay ranges to role and level
- $80,000 - $110,000

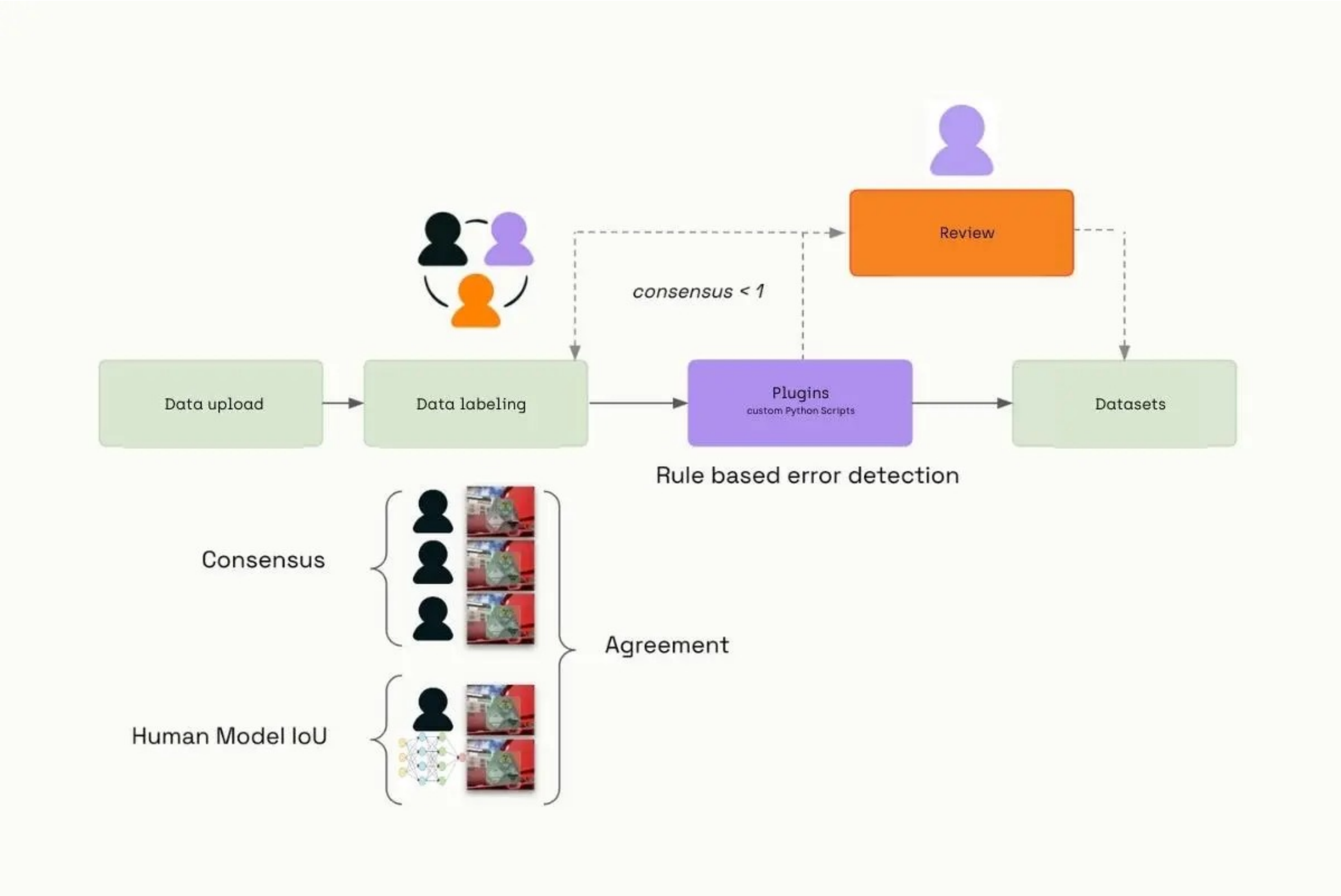
2 more items

Job description

Quality Control



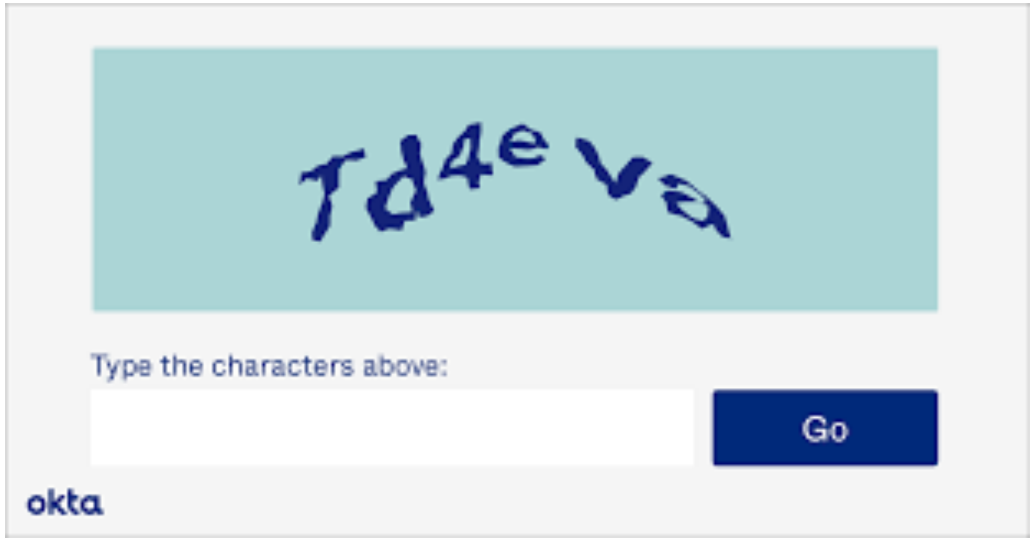
Quality Control



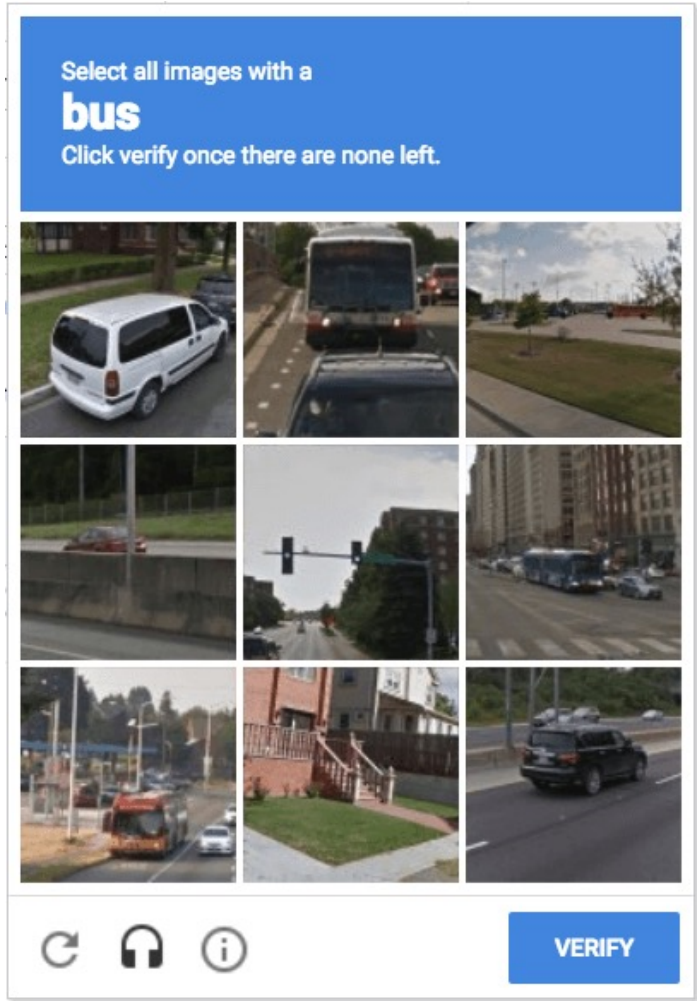
Creative Data Labeling Strategies

Captchas are for crowdsourced data labeling

They know the right label of some but not all figures. Your test is a subset of the figures and the rest is training data.

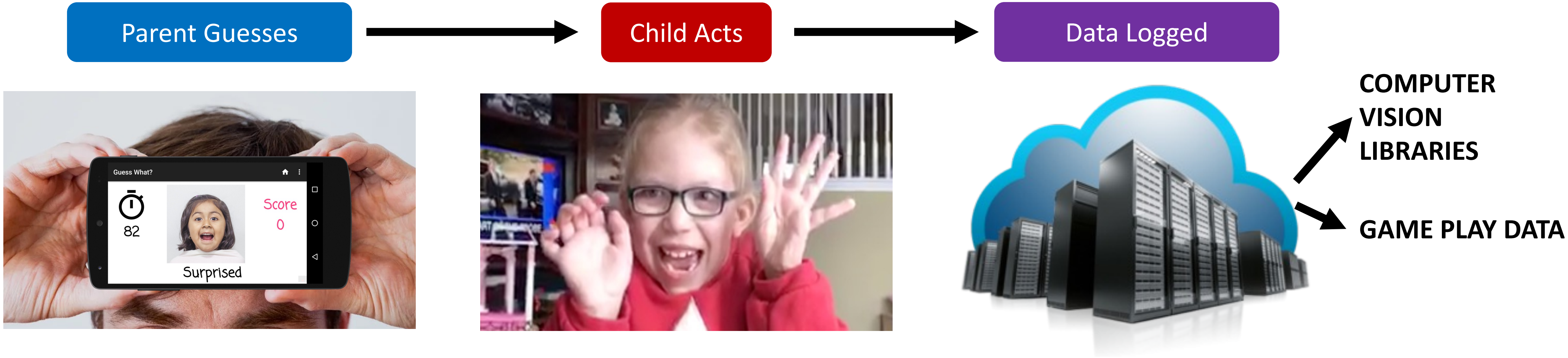


Type the characters above:



Select all images with a
bus
Click verify once there are none left.

Example from my research



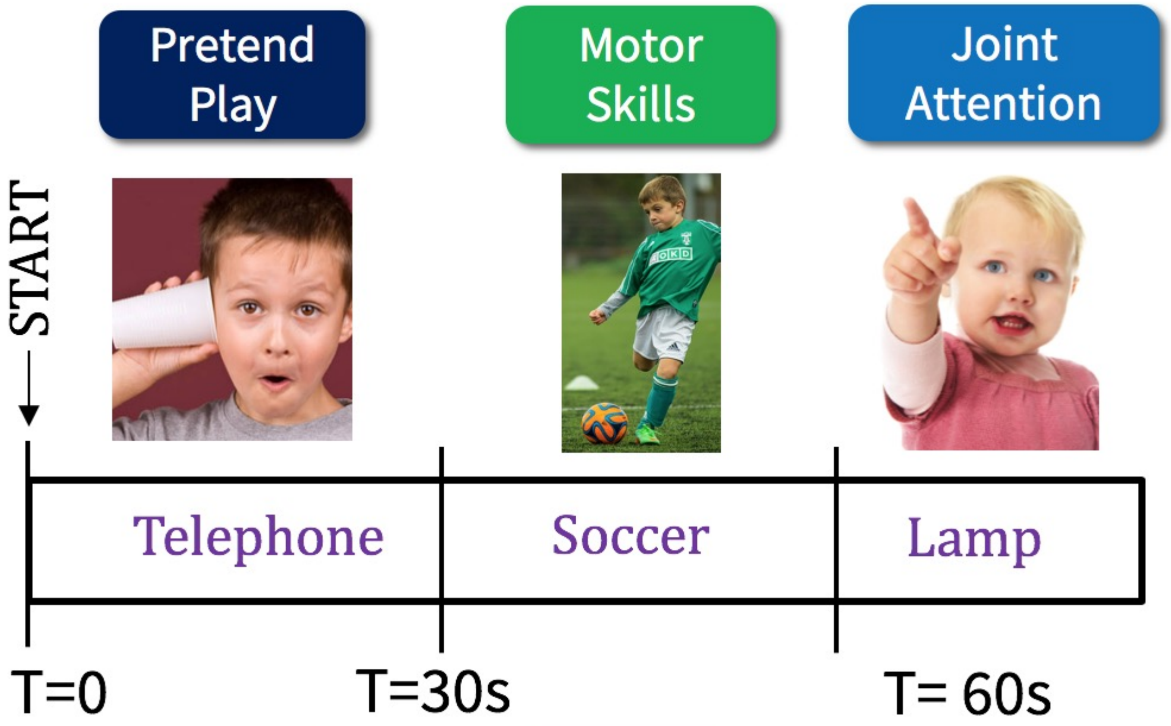
Parent Guesses → Child Acts → Data Logged

COMPUTER VISION LIBRARIES

GAME PLAY DATA

Kalantarian, **Washington** et al. *IEEE Conference on Healthcare Informatics.* 2018.
Kalantarian, **Washington** et al. *IEEE Transactions on Games.* 2018.
Kalantarian, …, **Washington** et al. *Journal of Healthcare Informatics Research.* 2019.
Kalantarian, …, **Washington** et al. *Artificial Intelligence in Medicine.* 2019.
Penev, …, **Washington** et al. *Applied Clinical Informatics.* 2021.

Example from my research



Pretend Play

Motor Skills

Joint Attention

START

Telephone

Soccer

Lamp

T=0

T=30s

T= 60s

Kalantarian, **Washington** et al. *IEEE Conference on Healthcare Informatics.* 2018.

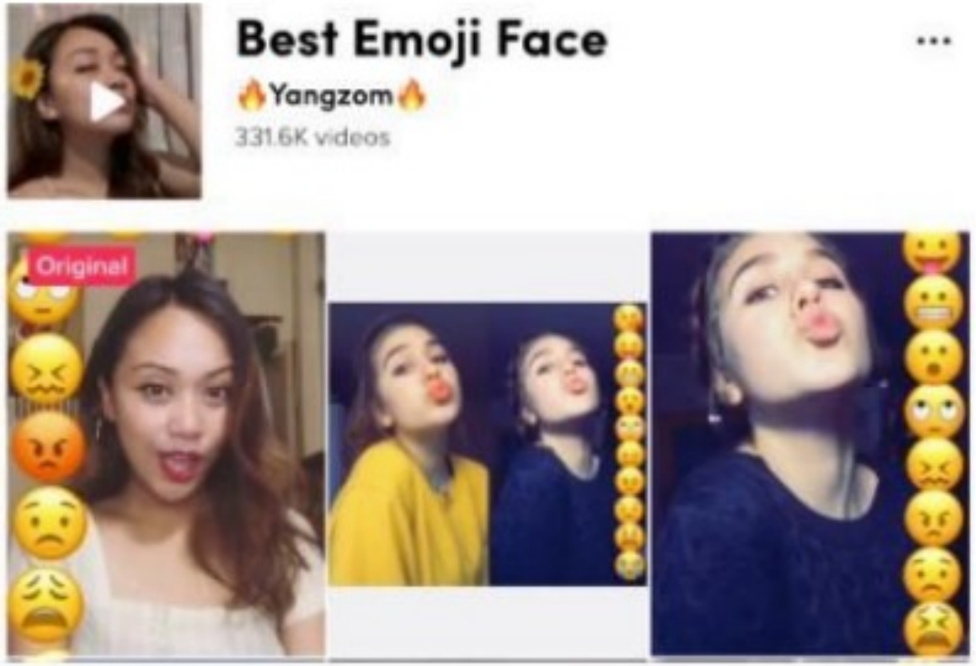
Kalantarian, **Washington** et al. *IEEE Transactions on Games.* 2018.

Kalantarian, …, **Washington** et al. *Journal of Healthcare Informatics Research.* 2019.

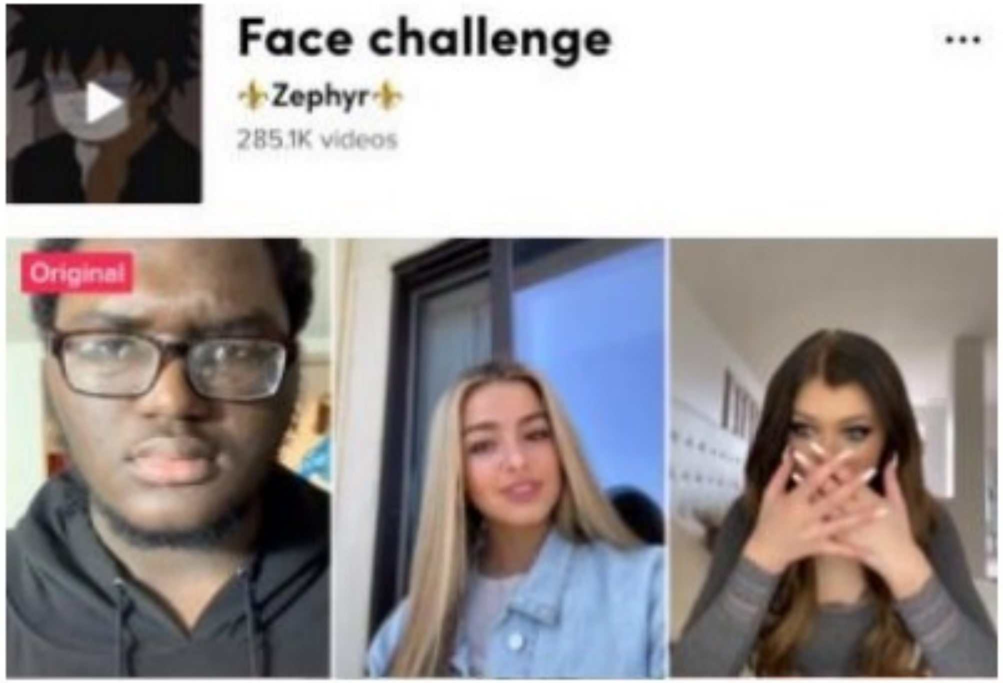
Kalantarian, …, **Washington** et al. *Artificial Intelligence in Medicine.* 2019.

Penev, …, **Washington** et al. *Applied Clinical Informatics.* 2021.

Another example from my research



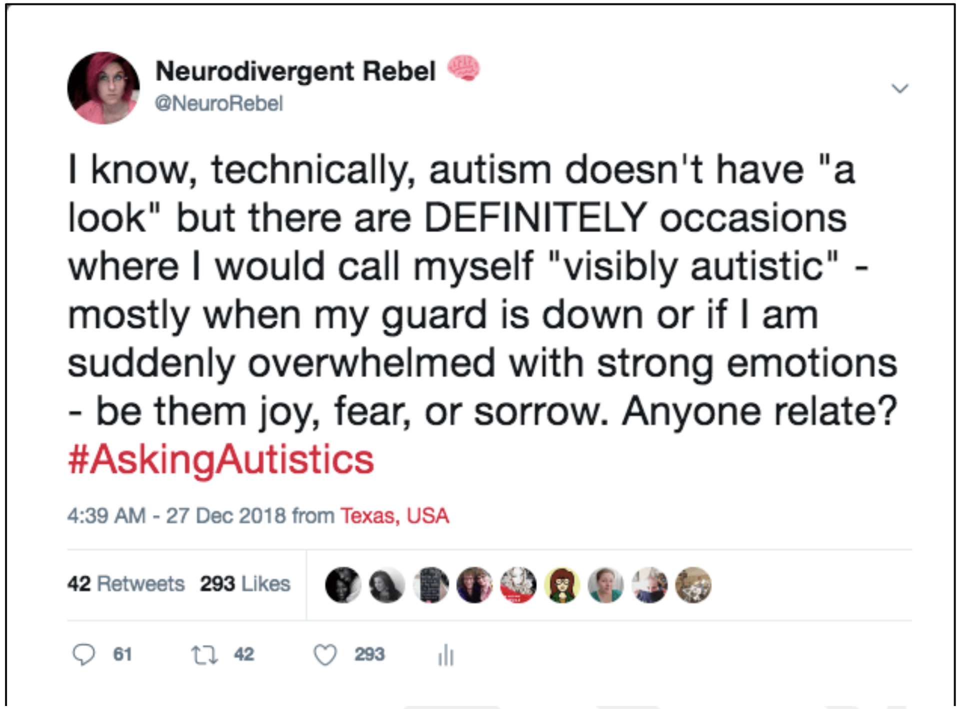
(a) Best Emoji Face Challenge.



(b) Face Challenge.

Surabhi, …, **Washington** et al. *CVPR Workshop.* 2022.

Example from my PhD student



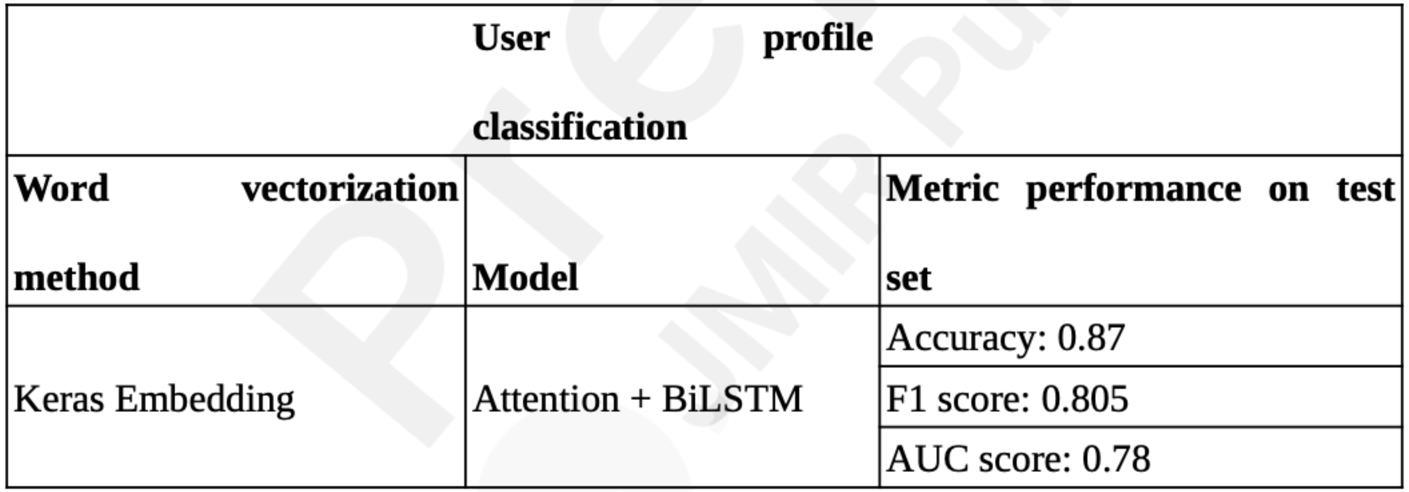
Neurodivergent Rebel 🧠
@NeuroRebel

I know, technically, autism doesn't have "a look" but there are DEFINITELY occasions where I would call myself "visibly autistic" - mostly when my guard is down or if I am suddenly overwhelmed with strong emotions - be them joy, fear, or sorrow. Anyone relate? #AskingAutistics

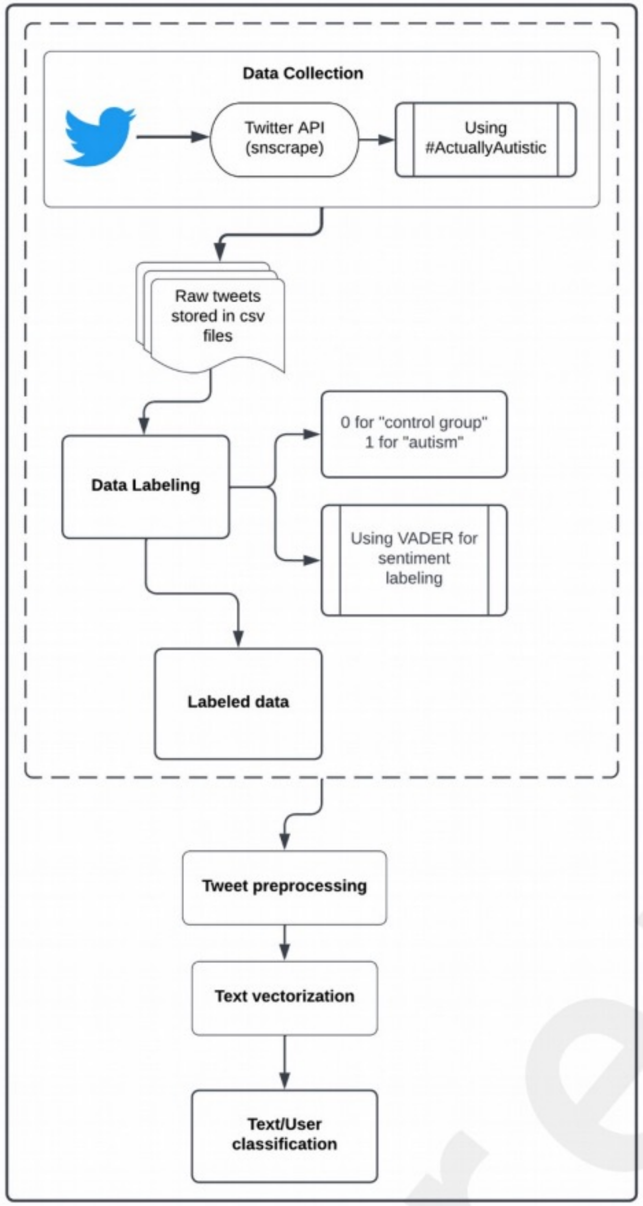
4:39 AM - 27 Dec 2018 from Texas, USA

42 Retweets **293** Likes

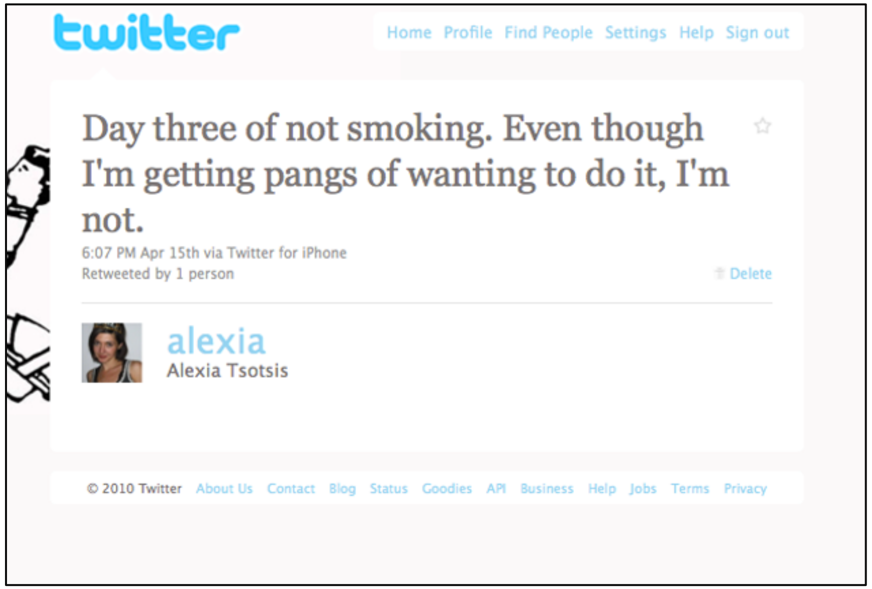
61 42 293



User profile classification		
Word vectorization method	**Model**	**Metric performance on test set**
Keras Embedding	Attention + BiLSTM	Accuracy: 0.87
		F1 score: 0.805
		AUC score: 0.78



One more example from my lab's research

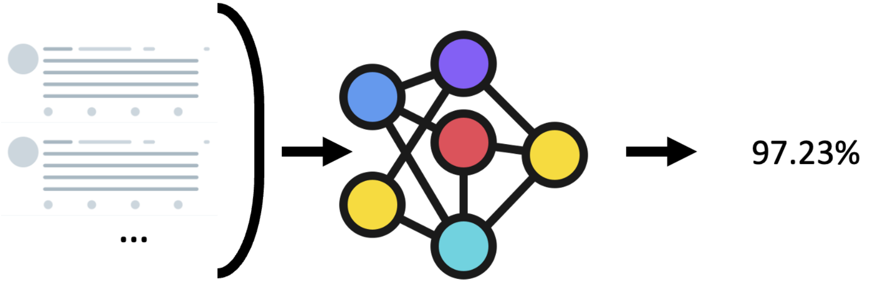

twitter
Home Profile Find People Settings Help Sign out

Day three of not smoking. Even though I'm getting pangs of wanting to do it, I'm not.

6:07 PM Apr 15th via Twitter for iPhone
Retweeted by 1 person

alexia
Alexia Tsotsis

© 2010 Twitter About Us Contact Blog Status Goodies API Business Help Jobs Terms Privacy



Input:
Twitter Data

Model:
Neural Network

Output:
Probability of Cigarette Use

Aditi Jaiswal

Dr. Pokhrel

UNIVERSITY OF HAWAI'I
CANCER CENTER

Dr. Amin

Class Exercise

- Spend 5-10 minutes thinking through the (1) recruitment, (2) data collection, (3) data labeling, and (4) data preprocessing strategies that you will need to conduct for your project
- We will discuss people’s strategies together as a class