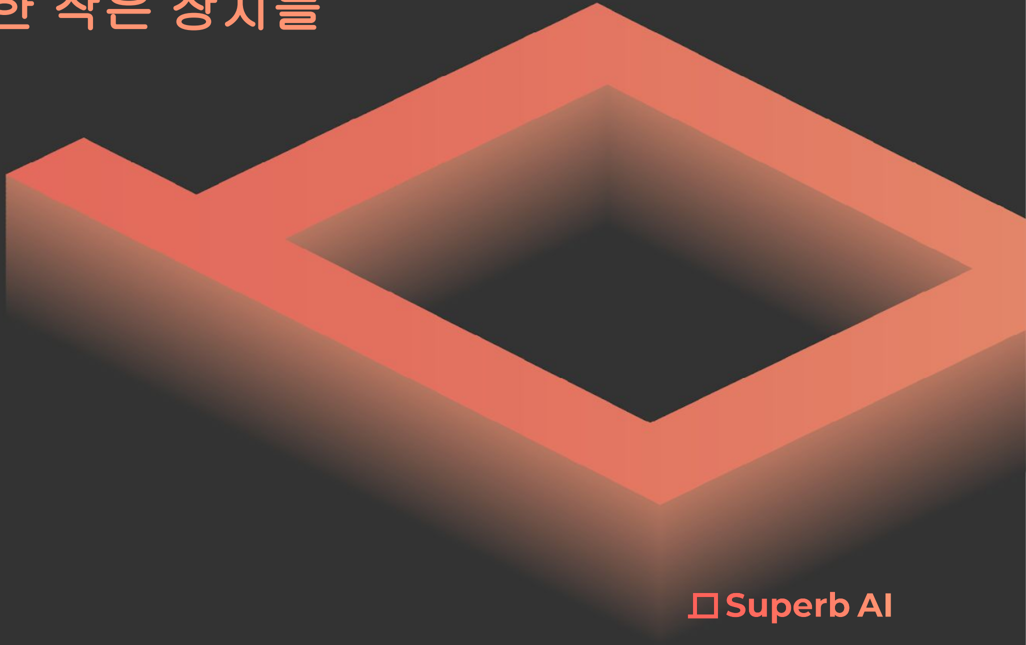
Data-centric MLOps

: 데이터 중심 MLOps를 돕기 위한 작은 장치들

Superb AI 이정권



Superb AI

AI / ML = Model + Data

AI / ML = Model + Data

Data centric?

A Chat with Andrew on MLOps: From Model-centric to Data-centric AI

Task

Baseline:

70% accuracy

↓

Target Performance:

90% accuracy

Should the team improve the code or the data?

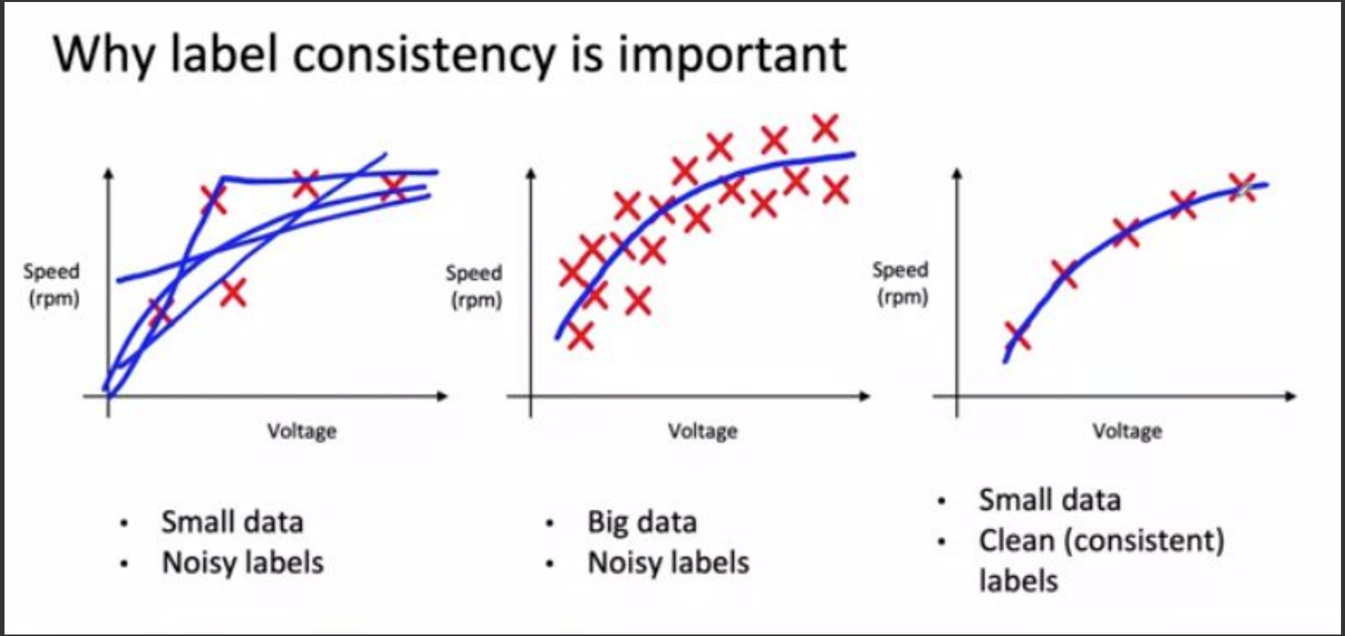
: code(20%), data(80%)

Superb AI

Improve AI →
Improve the quality of the data:

consistency
error rate
diversity
coverage
feedback frequency
size
...

A Chat with Andrew on MLOps: From Model-centric to Data-centric AI



Why label consistency is important

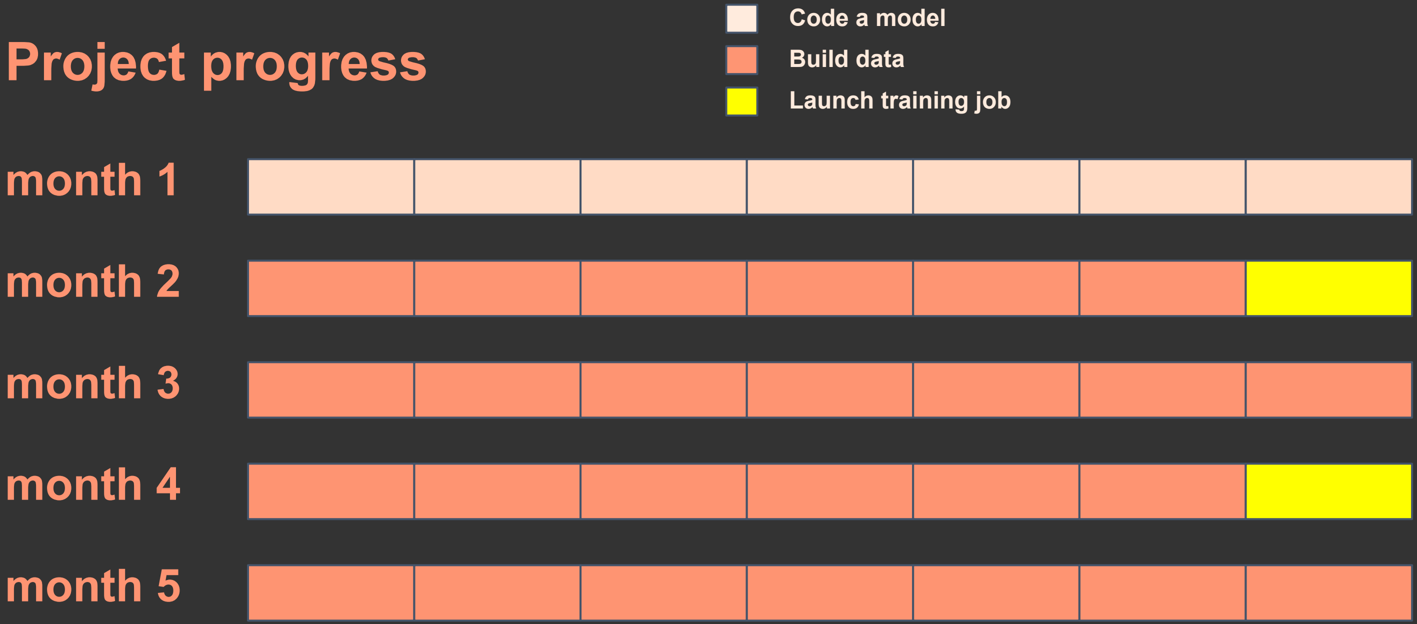
- Small data
- Noisy labels

- Big data
- Noisy labels

- Small data
- Clean (consistent) labels

slide credit: A Chat with Andrew on MLOps: From Model-centric to Data-centric AI (https://www.youtube.com/watch?v=06-AZXmwHjo)

사실은, 늘 해오던 일



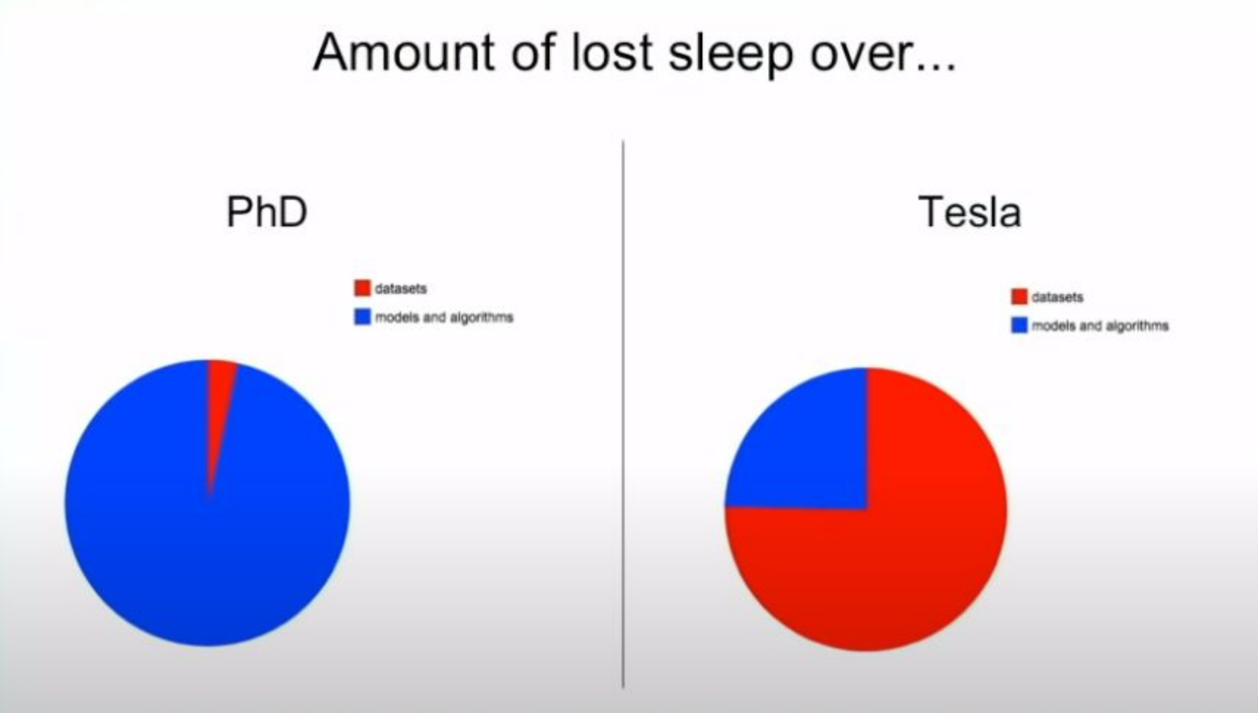
Project progress

- ⬜ Code a model
- 🟧 Build data
- 🟨 Launch training job

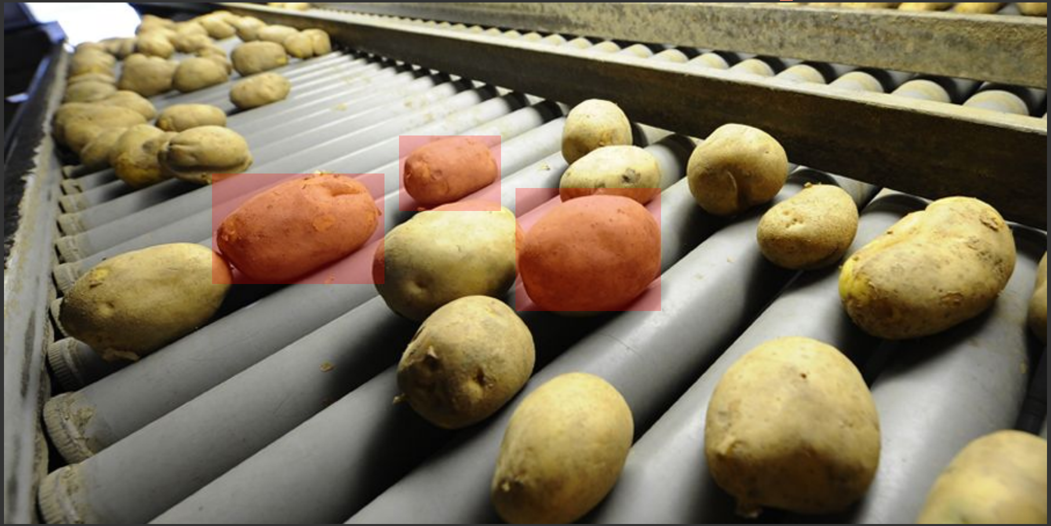
month 1	⬜	⬜	⬜	⬜	⬜	⬜	⬜
month 2	🟧	🟧	🟧	🟧	🟧	🟧	🟨
month 3	🟧	🟧	🟧	🟧	🟧	🟧	🟧
month 4	🟧	🟧	🟧	🟧	🟧	🟧	🟨
month 5	🟧	🟧	🟧	🟧	🟧	🟧	🟧

사실은, 늘 해오던 일

Building the Software 2.0 Stack (Andrej Karpathy, 2018)



Question:
How many labeled images are needed to solve this problem?



Answer:
100,000 images?

Superb AI

My Answer:
I don’t know. Let’s start from 5,000

WHY?

여전히, 잘 모른다
→ Data-centric MLOps

Systematic & iterative way to build Data for ML

단순히 지루한 작업을 자동화하는 과정이 아닌

ML 문제를 해결하기 위한 과정

저는 Superb AI라는 팀에서 이 문제를 풀고 있습니다.

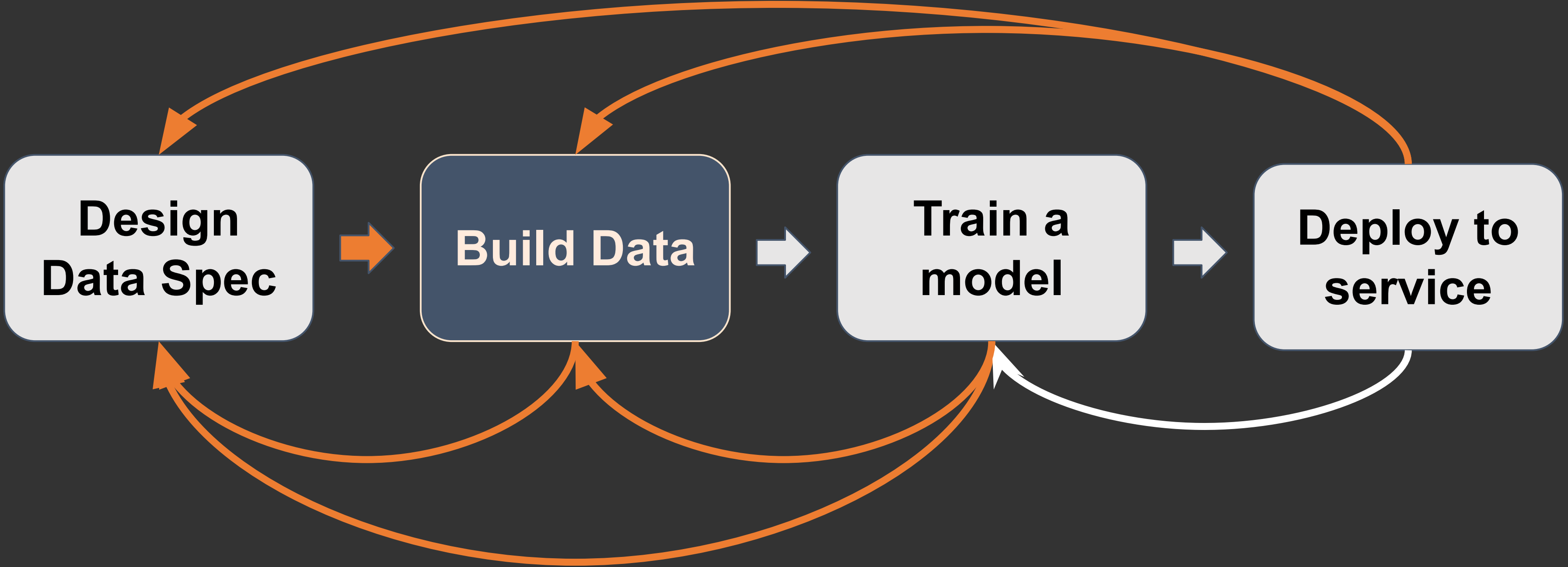
The Problem

<2달

<30명

<20,000 Images

The Meta Problem



Design Data Spec

Build Data

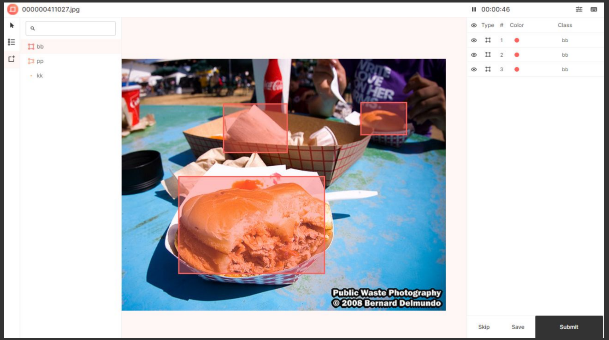
Train a model

Deploy to service

Superb AI

Starting Point

Labeling Tool



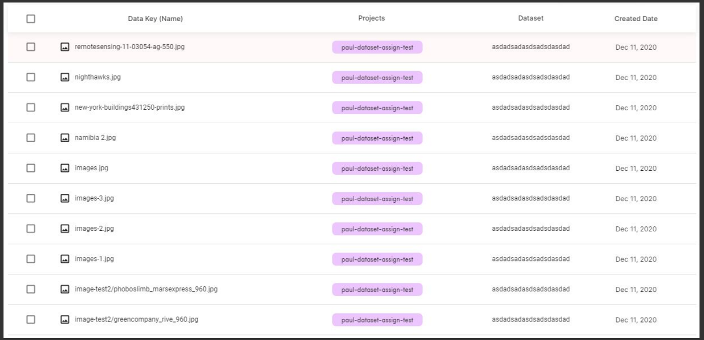
000000411027.jpg

00:00:46

- bb
- pp
- kk

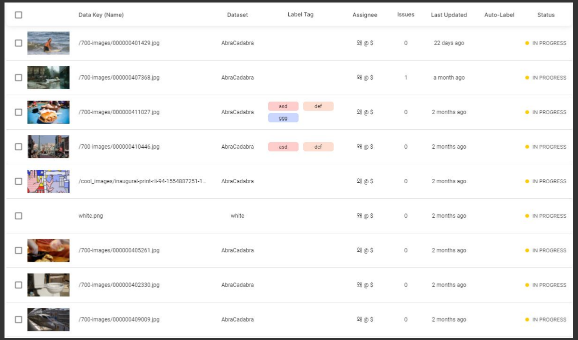
	Type	#	Color	Class
		1	●	bb
		2	●	bb
		3	●	bb

Data



□	Data Key (Name)	Projects	Dataset	Created Date
□	remotesensing-11-03054-ag-550.jpg	paul-dataset-assign-test	asdadsadasdsadsdasdad	Dec 11, 2020
□	nighthawks.jpg	paul-dataset-assign-test	asdadsadasdsadsdasdad	Dec 11, 2020
□	new-york-buildings431250-prints.jpg	paul-dataset-assign-test	asdadsadasdsadsdasdad	Dec 11, 2020
□	namibia 2.jpg	paul-dataset-assign-test	asdadsadasdsadsdasdad	Dec 11, 2020
□	images.jpg	paul-dataset-assign-test	asdadsadasdsadsdasdad	Dec 11, 2020
□	images-3.jpg	paul-dataset-assign-test	asdadsadasdsadsdasdad	Dec 11, 2020
□	images-2.jpg	paul-dataset-assign-test	asdadsadasdsadsdasdad	Dec 11, 2020
□	images-1.jpg	paul-dataset-assign-test	asdadsadasdsadsdasdad	Dec 11, 2020
□	image-test2/phoboslimb_marsexpress_960.jpg	paul-dataset-assign-test	asdadsadasdsadsdasdad	Dec 11, 2020
□	image-test2/greencompany_rive_960.jpg	paul-dataset-assign-test	asdadsadasdsadsdasdad	Dec 11, 2020

Label

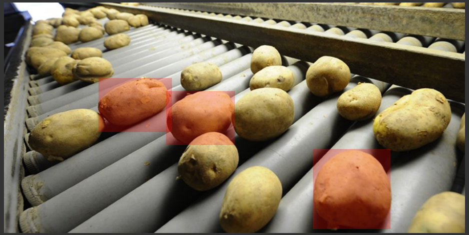


□	Data Key (Name)	Dataset	Label Tag	Assignee	Issues	Last Updated	Auto-Label	Status
□	/700-images/000000401429.jpg	AbraCadabra		왜 @ $	0	22 days ago		● IN PROGRESS
□	/700-images/000000407368.jpg	AbraCadabra		왜 @ $	1	a month ago		● IN PROGRESS
□	/700-images/000000411027.jpg	AbraCadabra	asd def ggg	왜 @ $	0	2 months ago		● IN PROGRESS
□	/700-images/000000410446.jpg	AbraCadabra	asd def	왜 @ $	0	2 months ago		● IN PROGRESS
□	/cool_images/inaugural-print-rli-94-1554887251-1...	AbraCadabra		왜 @ $	0	2 months ago		● IN PROGRESS
□	white.png	white		왜 @ $	0	2 months ago		● IN PROGRESS
□	/700-images/000000405261.jpg	AbraCadabra		왜 @ $	0	2 months ago		● IN PROGRESS
□	/700-images/000000402330.jpg	AbraCadabra		왜 @ $	0	2 months ago		● IN PROGRESS
□	/700-images/000000409009.jpg	AbraCadabra		왜 @ $	0	2 months ago		● IN PROGRESS

Reusable Data Spec

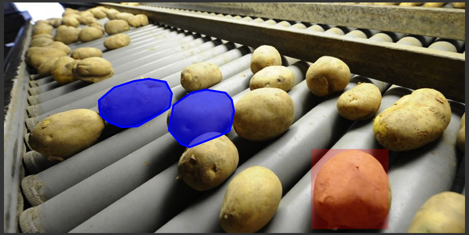
```
{
  project_name:
   potato_detect_1

  data_spec:
   good_potato:
    box:
     color: red
     condition: ...
   bad_potato:
    box:
}
```



```
{
  project_name:
   potato_detect_2

  data_spec:
   good_potato:
    polygon:
     color: red
     condition: ...
   bad_potato:
    box:
}
```

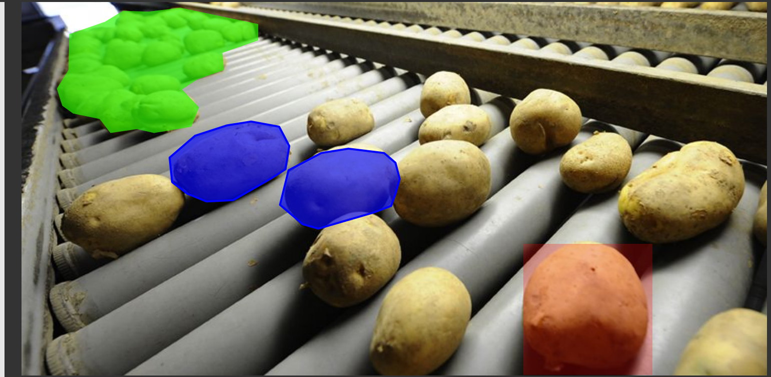


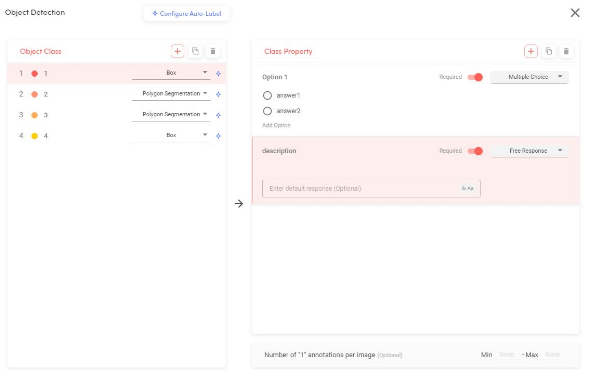
Reusable Data Spec

```
{
   project_name:
     potato_detect_13

   data_spec:
     best_potato:
       polygon:
       direction:
       options: ...
     good_potato: {}
     normal_potato: {}
     bad_potato: {}


}
```





Object Detection

Object Class

1	🟥	1	Box
2	●	2	Polygon Segmentation
3	●	3	Polygon Segmentation
4	●	4	Box

Class Property

Option 1 Required Multiple Choice

○ answer1

○ answer2

description Required Free Response

Enter default response (Optional)

Number of "1" annotations per image (Optional) Min None - Max None

Goal ≠ Task

ALWAYS configured repeatedly

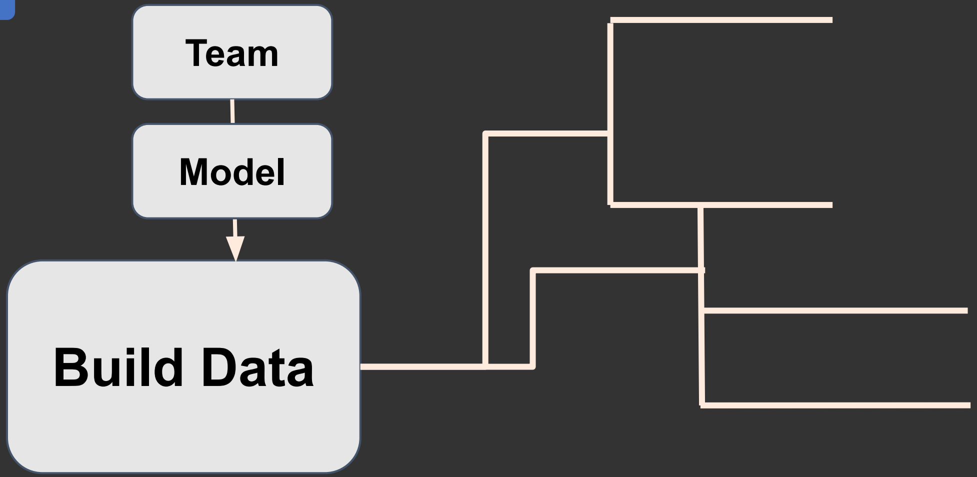
name,
color,
type,
conditions,
options,
property,
ROI Info,
...

Support flexible pipeline

WORKING SUBMITTED REVIEWED

100 different problems,
100 different datasets,
100 different ways

To support
flexible pipeline

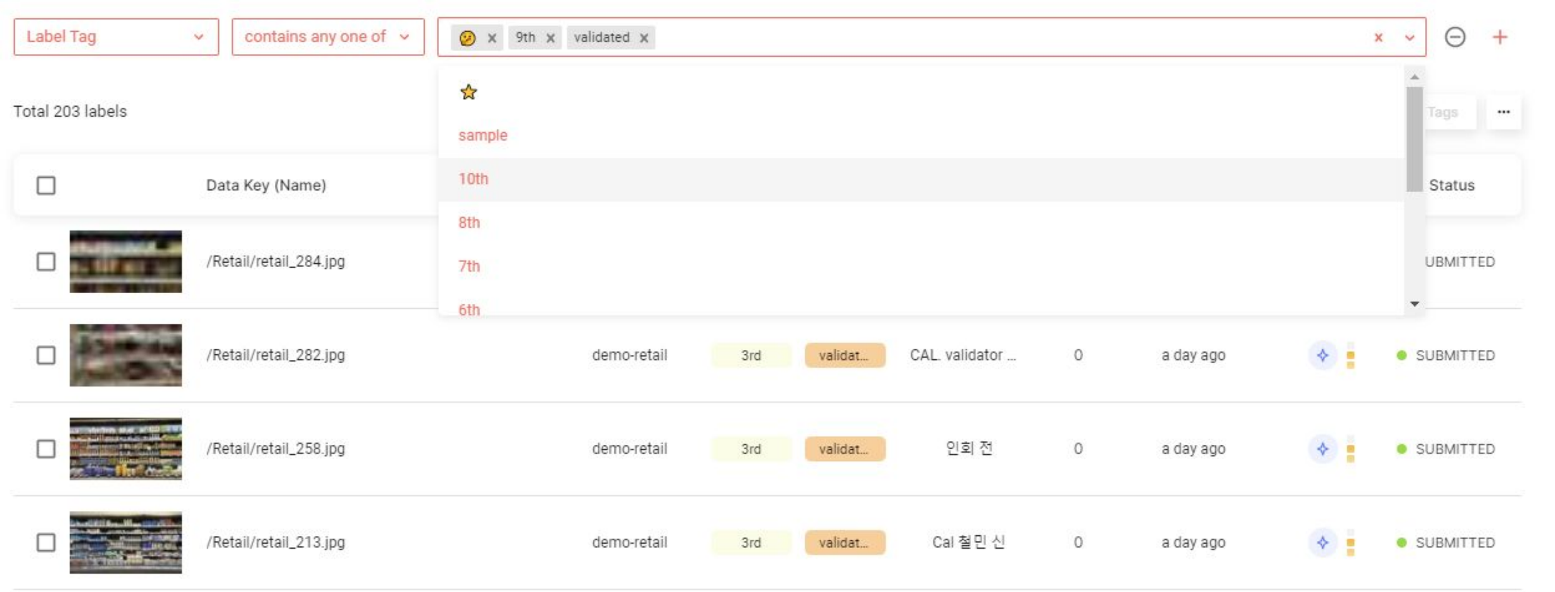


Team

Model

Build Data

Support flexible pipeline



Label Tag ⌄ | contains any one of ⌄ | 🤔 × | 9th × | validated × | × ⌄ | ⊖ | +

★

sample

10th

8th

7th

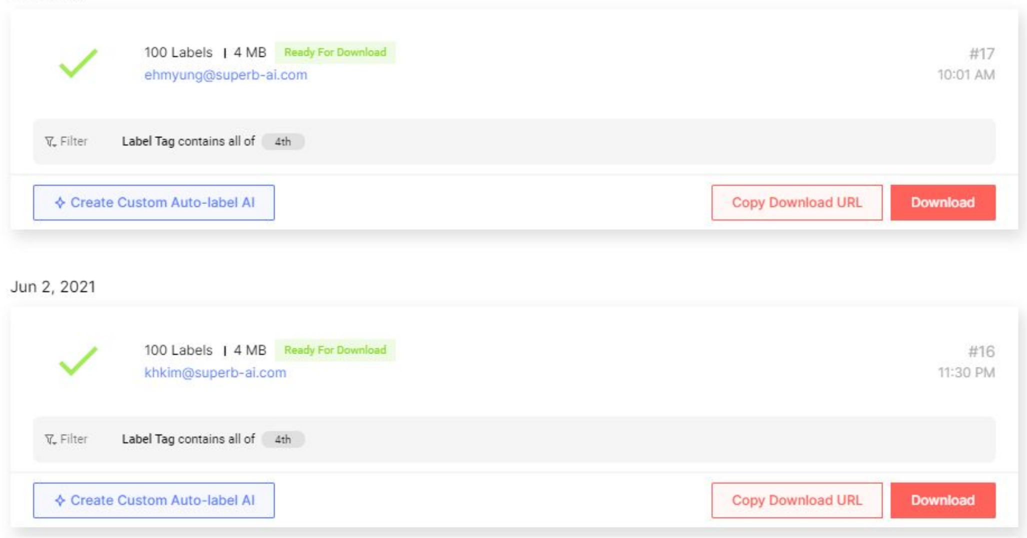
6th

Total 203 labels

□		Data Key (Name)									Status
□		/Retail/retail_284.jpg									UBMITTED
□		/Retail/retail_282.jpg	demo-retail	3rd	validat...	CAL. validator ...	0	a day ago	✧		● SUBMITTED
□		/Retail/retail_258.jpg	demo-retail	3rd	validat...	인회 전	0	a day ago	✧		● SUBMITTED
□		/Retail/retail_213.jpg	demo-retail	3rd	validat...	Cal 철민 신	0	a day ago	✧		● SUBMITTED

Versioning

Set 단위, 실험 당



100 Labels | 4 MB Ready For Download

ehmyung@superb-ai.com

#17

10:01 AM

Filter Label Tag contains all of 4th

Jun 2, 2021

100 Labels | 4 MB Ready For Download

khkim@superb-ai.com

#16

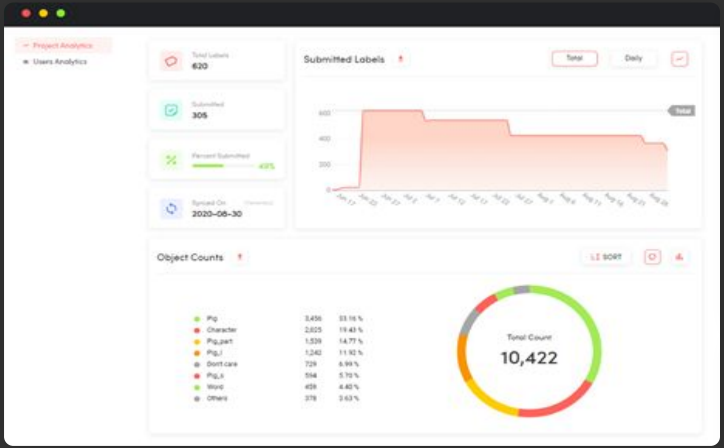
11:30 PM

Filter Label Tag contains all of 4th

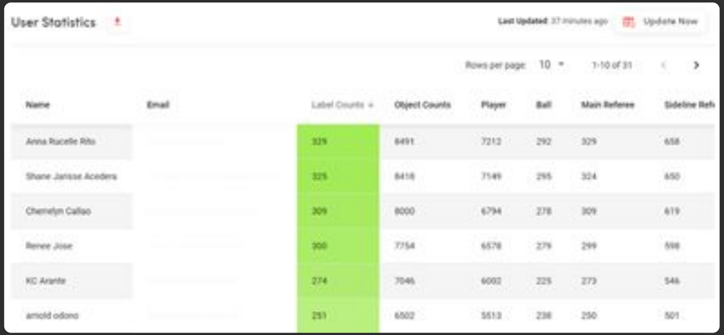
Superb AI

Detailed Statistics & Report

ML Engineer를 위해 … ?



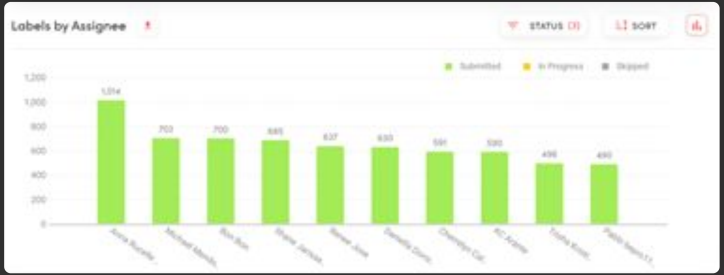
● Pig	3,456	33.16 %
● Character	2,025	19.43 %
● Pig_part	1,539	14.77 %
● Pig_l	1,242	11.92 %
● Dont care	729	6.99 %
● Pig_s	594	5.70 %
● Word	459	4.40 %
● Others	378	3.63 %


User Statistics

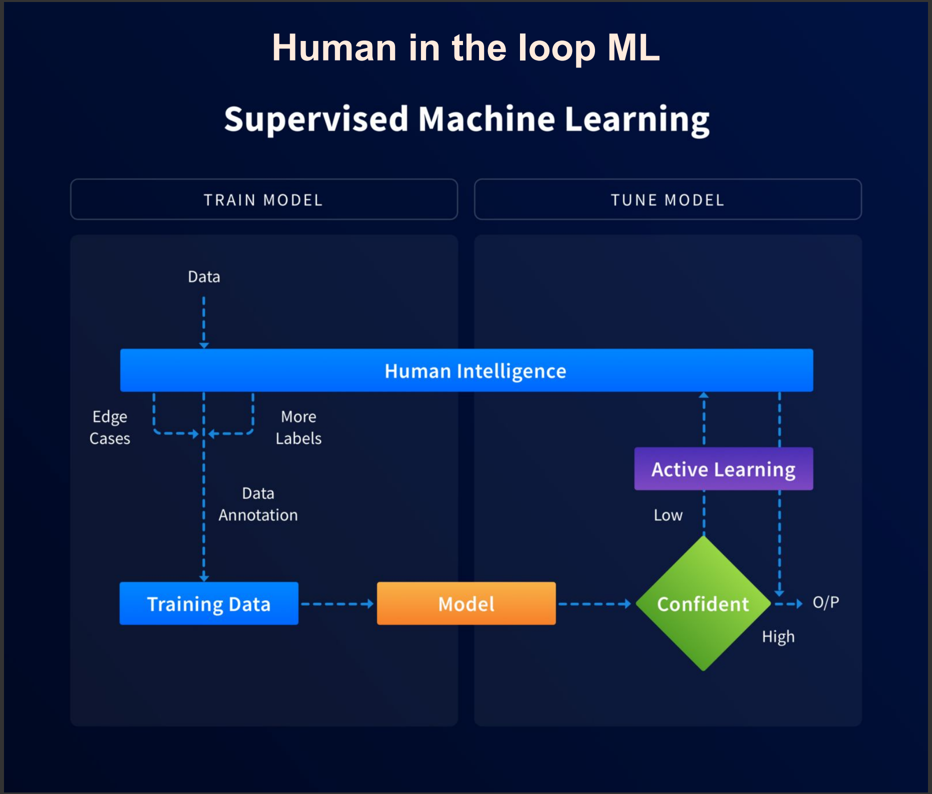
Last Updated 37 minutes ago Update Now

Rows per page: 10 1-10 of 31

Name	Email	Label Counts	Object Counts	Player	Ball	Main Referee	Sideline Refe
Anna Rucelle Rito		329	8491	7212	292	329	658
Shane Janisse Acedera		325	8418	7149	295	324	650
Chemelyn Callao		309	8000	6794	278	309	619
Renee Jose		300	7754	6578	279	299	598
KC Arante		274	7046	6002	225	273	546
arnold odono		251	6502	5513	238	250	501



Human in the loop ^ 2

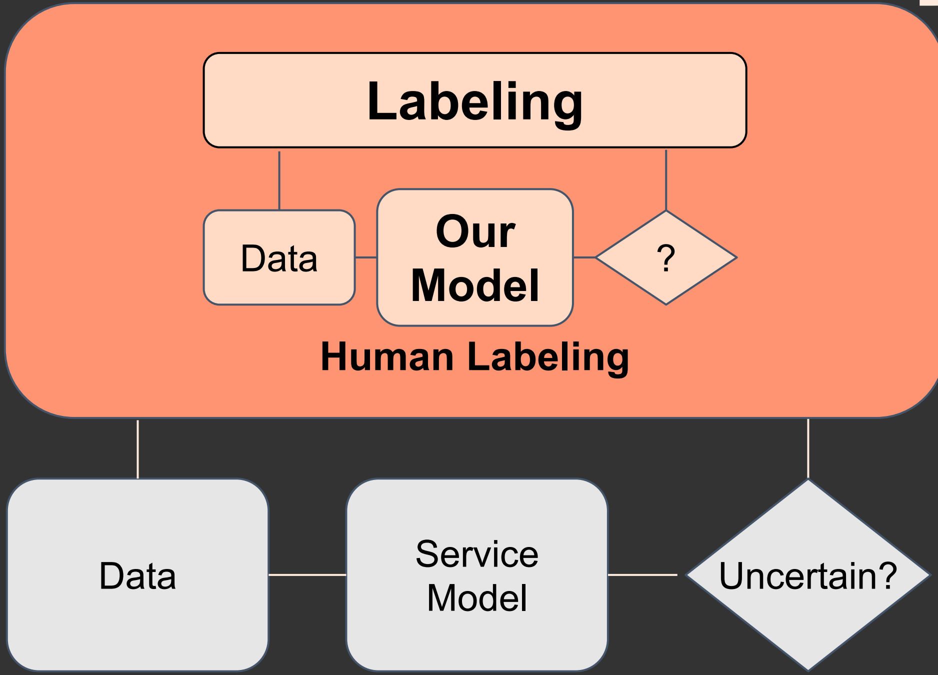


Superb AI

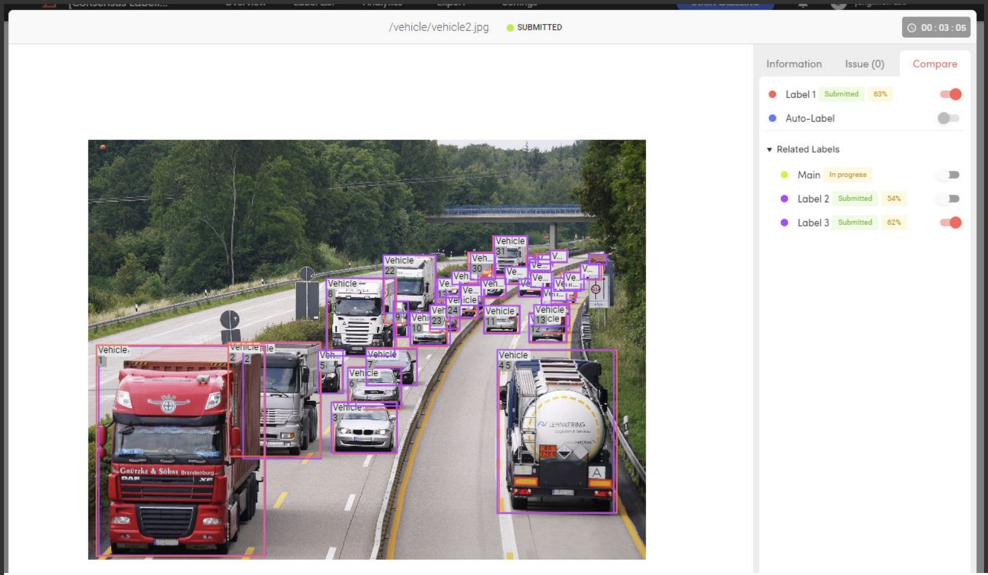
Human in the loop ^ 2

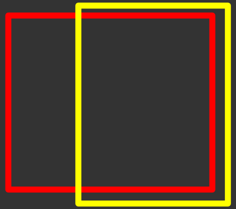
Inside Human Labeling

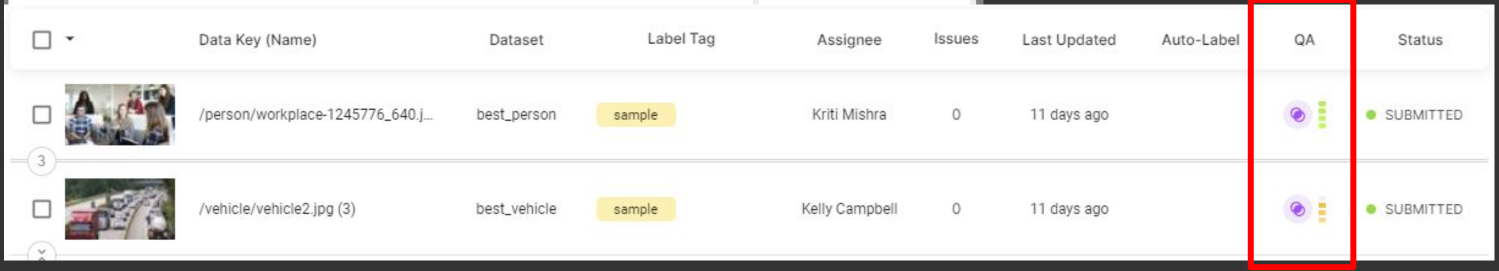
Label-wise Confidence
Overall Set Confidence
User performance estimate
Boost Labeling
...



Keep labels consistent

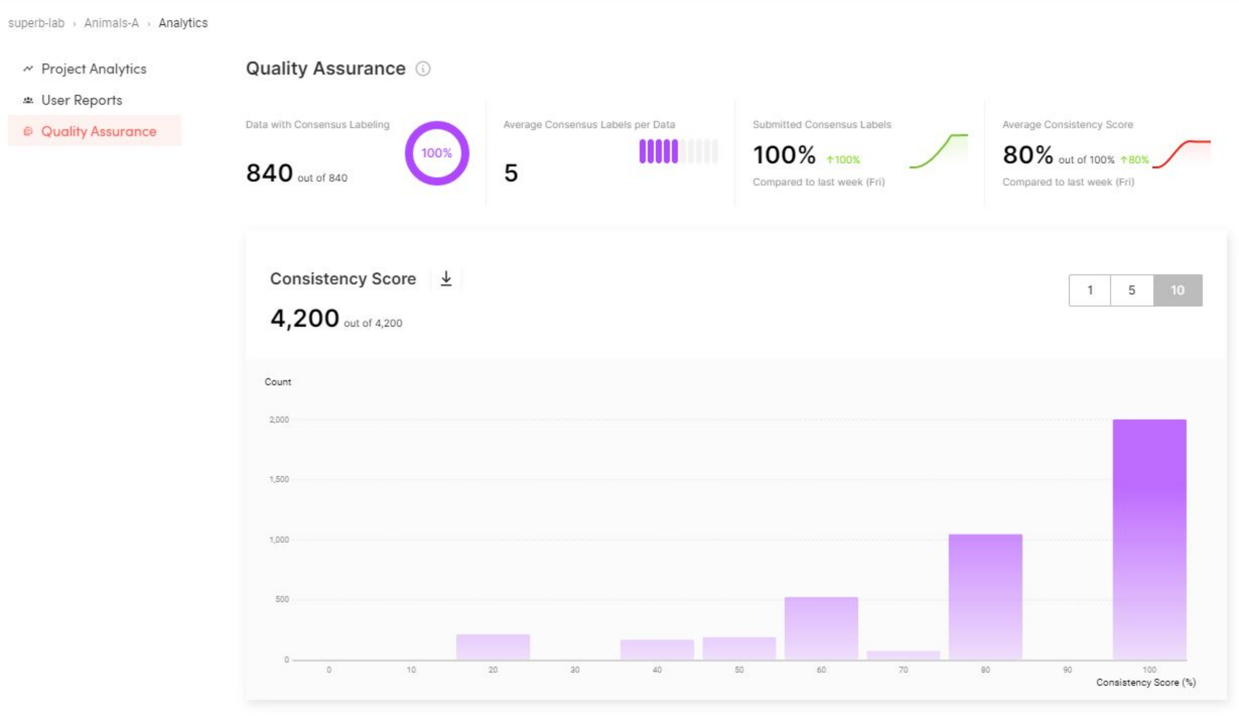






□		Data Key (Name)	Dataset	Label Tag	Assignee	Issues	Last Updated	Auto-Label	QA	Status
□		/person/workplace-1245776_640.j...	best_person	sample	Kriti Mishra	0	11 days ago			● SUBMITTED
□		/vehicle/vehicle2.jpg (3)	best_vehicle	sample	Kelly Campbell	0	11 days ago			● SUBMITTED

Keep labels consistent



Quality Assurance

Data with Consensus Labeling
840 out of 840

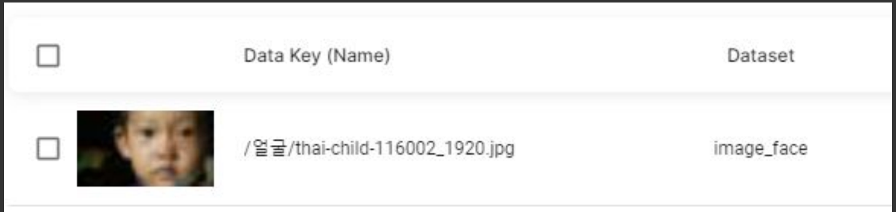
Average Consensus Labels per Data
5

Submitted Consensus Labels
100% ↑100%
Compared to last week (Fri)

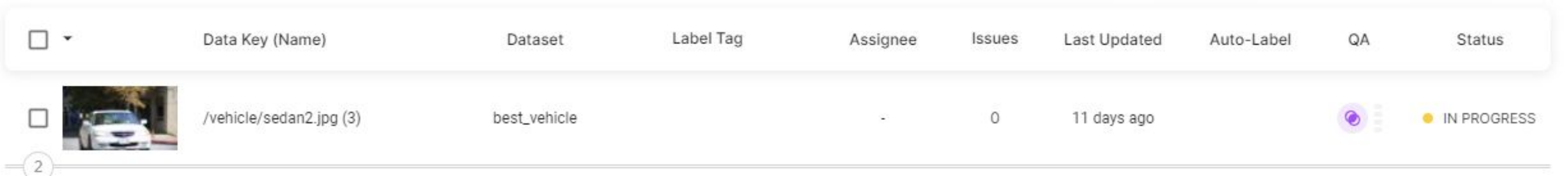
Average Consistency Score
80% out of 100% ↑80%
Compared to last week (Fri)

Consistency Score
4,200 out of 4,200

요약



□	Data Key (Name)	Dataset
□	/얼굴/thai-child-116002_1920.jpg	image_face



□ ▾	Data Key (Name)	Dataset	Label Tag	Assignee	Issues	Last Updated	Auto-Label	QA	Status
□	/vehicle/sedan2.jpg (3)	best_vehicle		-	0	11 days ago			● IN PROGRESS

2

마무리

Source data analysis, User analysis,
Log, Task matching, etc
여전히 할일이 정말 많다.

SDK를 이용한 사용 예제!는 다음에
https://github.com/superb-AI-Suite/

Full-pipeline MLOps
https://ai-infrastructure.org/

Valohai

Pachyderm
version-controlled data science