Ensuring Quality Data for Deep Learning in Varied Application Domains: Data Collection, Curation and Annotation

Gaurav Singh
System Architect and Perception Lead
Nemo @ Ridecell

Talk outline

- Introduction and motivation
- Data collection
- Data curation
- Data annotation
- Conclusion

Introduction and motivation

Collection, curation and annotation key to good deep learning

Stage	Questions to address
Collection	Collect/buy, post-production data collection, synthetic data?
Curation	Active learning, tagging, curation tooling
Annotation	Instructions, auditing techniques, in-house vs managed service

Data collection

RIDECELL

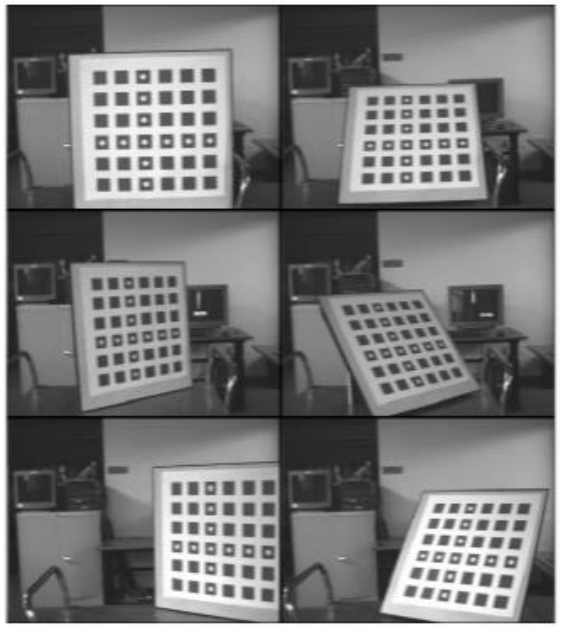
Data collection - datasets for deep learning

	Open and free datasets	*Paid Licensed datasets*	*Own datasets*	*Synthetic datasets*
Description	Free datasets under Creative Commons Attribution 4.0 licenses or similar	Paid licenses. Sold by companies for commercial gain	Collected on your own or with partners	Generated for specific use cases / corner cases by simulation companies.
Examples	DATA USA: EXPLORE, MAP, COMPARE, AND DOWNLOAD U.S. DATA Kinetics The Wikipedia Corpus	Mapillary Matterport MedICaT	Collected by your own company and annotated.	cognata: Simulating Autonomy ANYVERSE
Pros	Lots of research interest - publications and networks built on top of these datasets. Labeling done usually. Government datasets usually here. Can be relabeled with required class.	Some research interest from universities etc. Generally higher quality labels. Some filtering to pick interesting sequences.	Can be heavily filtered for use case, full flexibility in labeling. Quality under company collecting and using data.	Generated for specific corner cases where algorithms are not working well. Easily annotated.
Cons	Can be limited size, labeling accuracy can be suspect.	Cost, Can't be re-labelled.	Costly to collect and annotate	Reality - simulation gap

Data collection – key challenges for sensor data

Sensor data (Lidar, Camera, GPS etc.) collection challenges

- Sensor rig setup
- Sensor rig calibration
- Time syncing sensors
- Maintenance of rig – recalibration etc.



Data collection – key challenges contd.

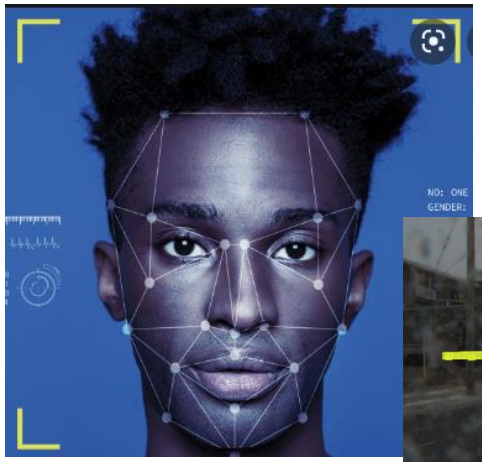
More challenges –

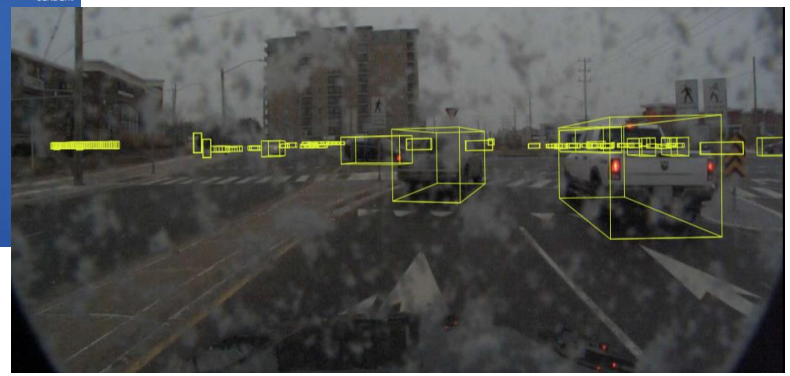
- Planning of collection routes
- Cost of operations
- Data pipelines and data infrastructure

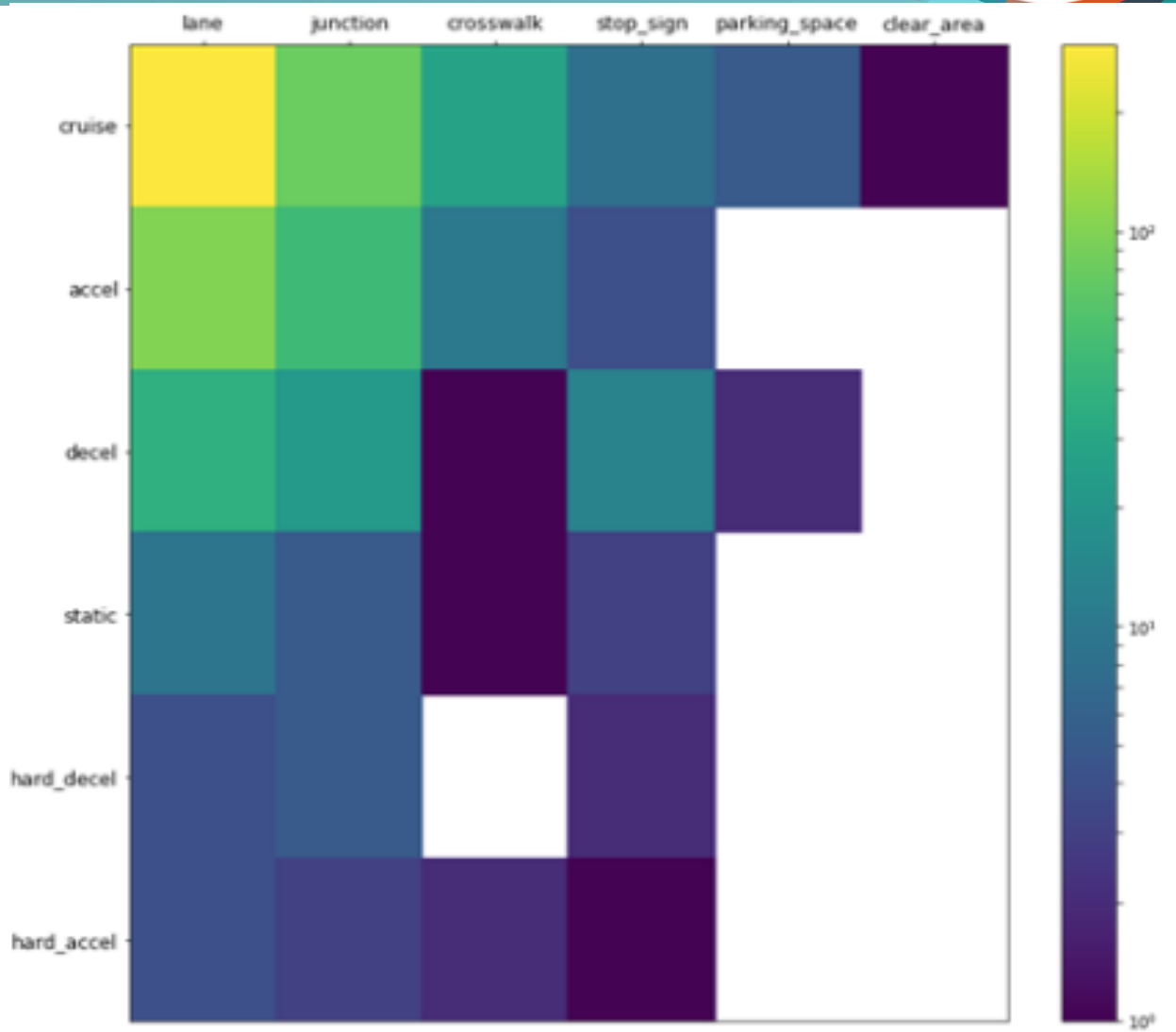
Data collection – pre-launch best practices

- Metadata important during data collection
  - Pre-trained networks used
  - Other sources - GPS or map data

- Metadata gives 3Ws (where, when and what) -
  - Help explain gaps in data collection
  - Target under-represented scenarios







Data collection – production best practices

- Use ML confidence scores.
- Use customer feedback, e.g., customer spoke twice to voice recognition.
- Use ambiguous data, e.g., traffic sign detected close. Earlier frames selected.

Synthetic data – for corner cases

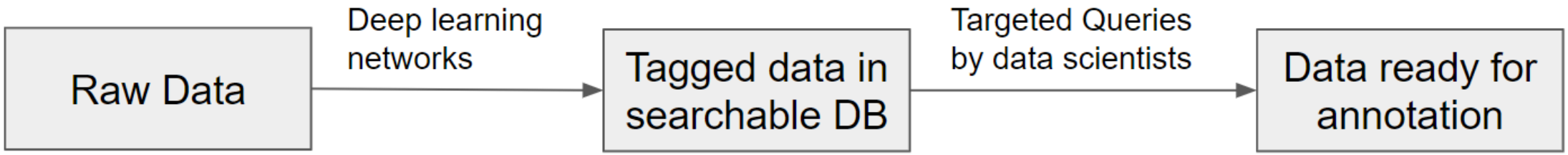
- Synthetic data for rarer conditions/difficult to capture real world scenarios.
- Synthetic data mixed with real data gives good performance.

RIDECELL

Data curation

RIDECELL

Data curation – prelaunch
Tagging data for targeted search



Deep learning networks/algos

Tags

car 3 | trafficcone 16 | cruise 2 | accel 1 | decel 2 | hard_decel 1 | lanelet 1 | vehicle 43 | unknown 5 | bicycle 2 | pedestrian 1

Metadata tags stored in databases with timestamps.
Exposed for querying through a GUI

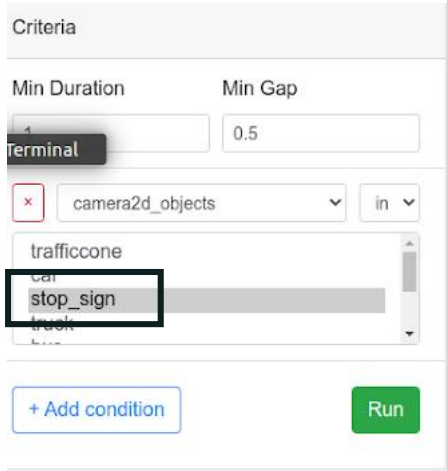
RIDECELL

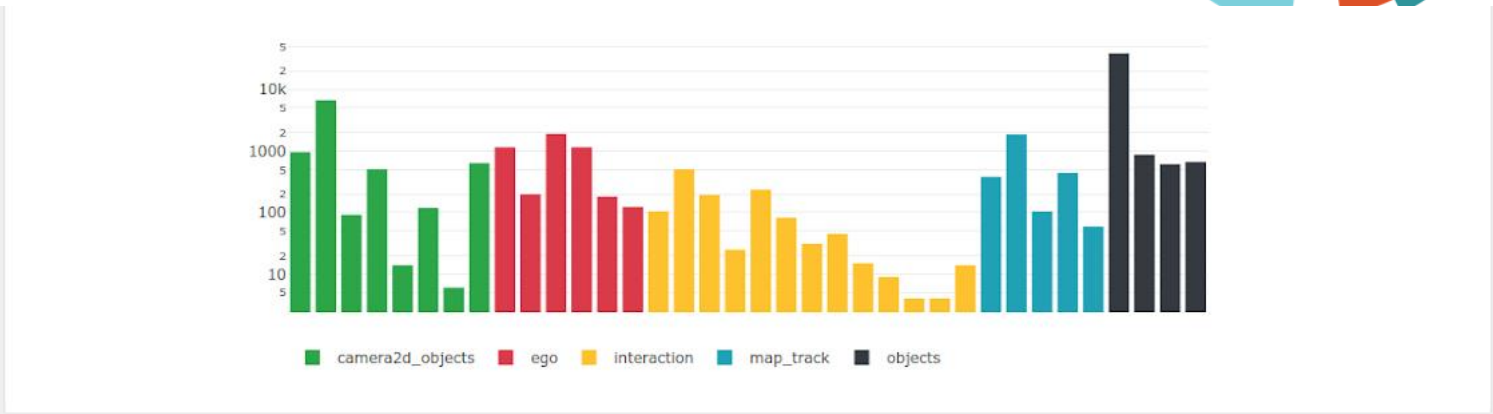
Data curation – prelaunch
Tools to select batch for annotation

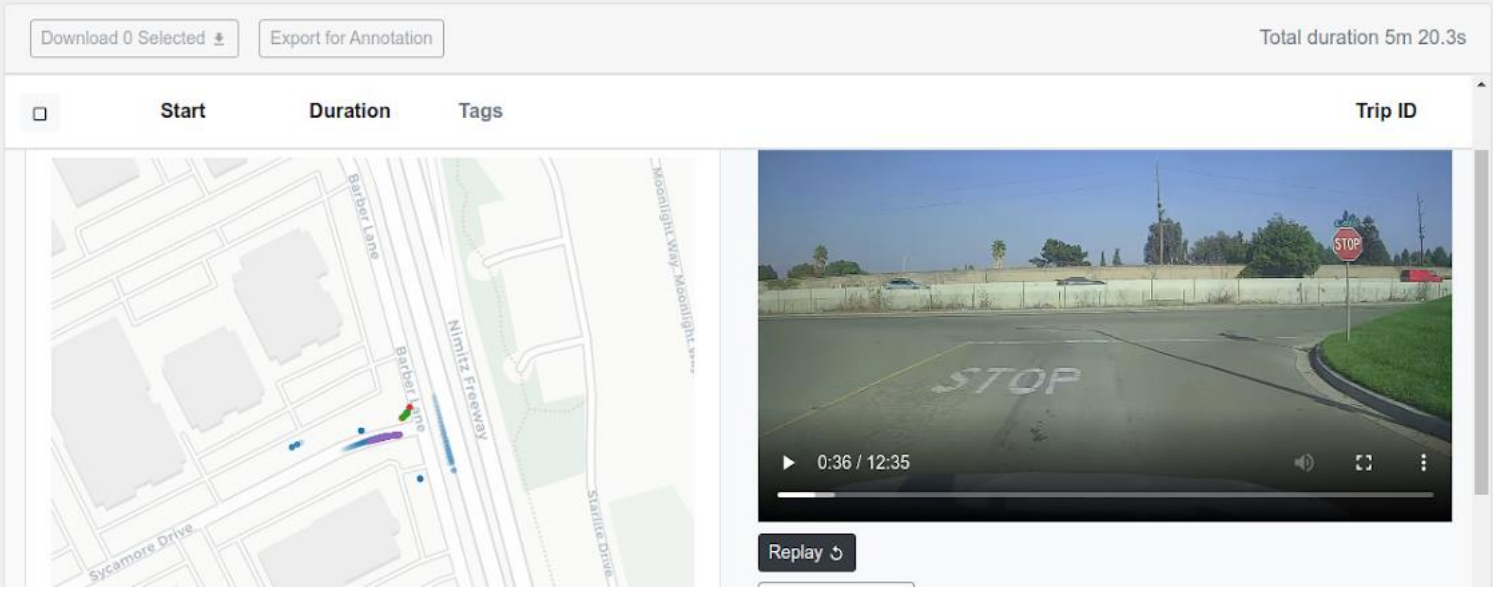
Tool to search tags

E.g., Search stop sign, construction sign and other key scenarios

Data scientists use such tools



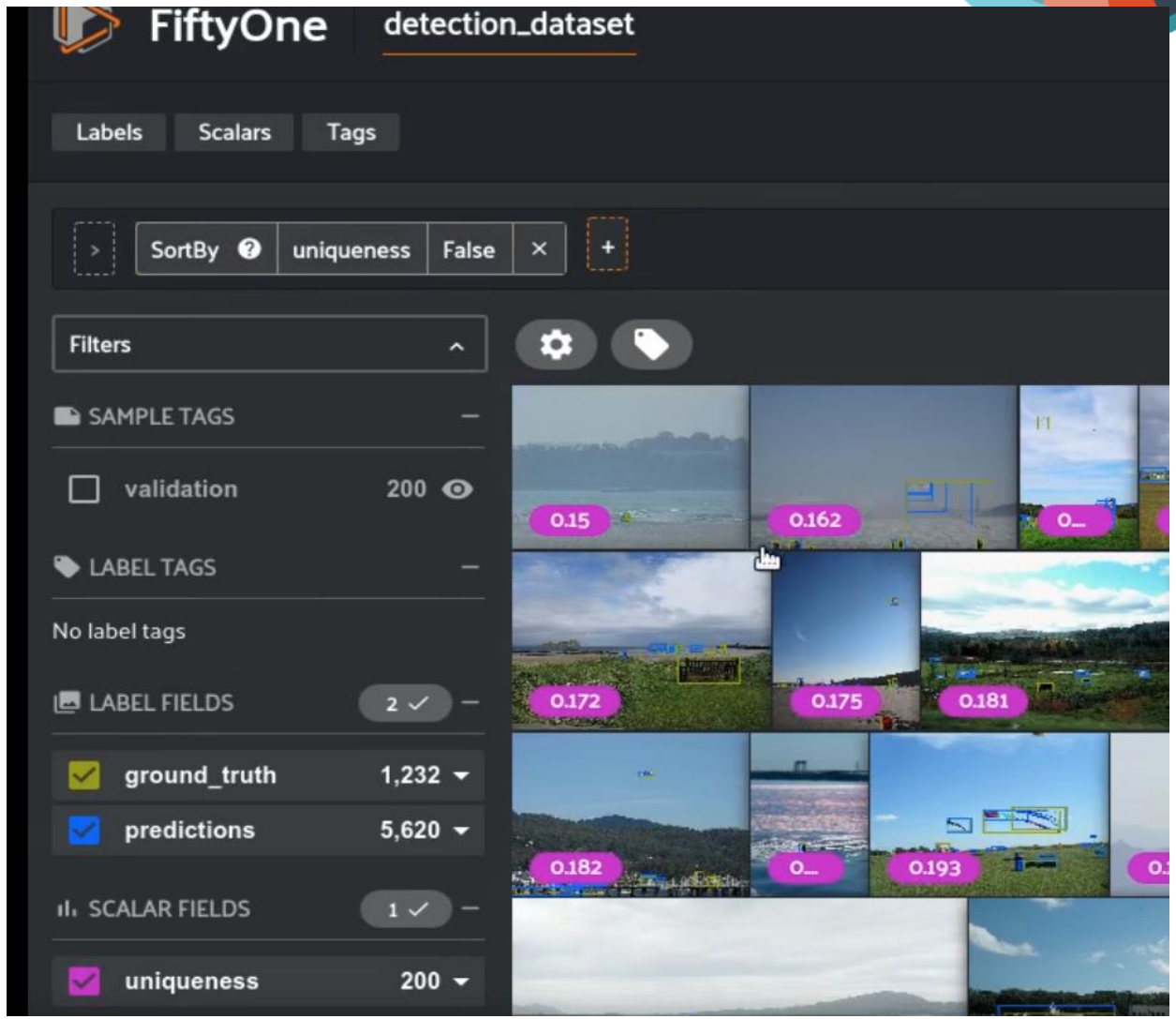




Data curation – prelaunch
Tools to select batch for annotation

Key features:

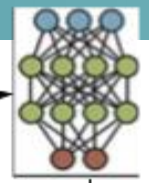
- Fast metadata queries – varied tags overlapping
- Re-run tagging with new models
- Support for text/image similarity
- Compare ground truth and ML inference visually



*Courtesy Voxel51 presentation

Data curation – active learning in production systems

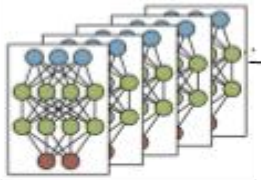
trains



Customer target deep learning network

Fleet

Labeled data



Other high performance models

Active Learning inference module M

Raw unlabeled drive data pool

D_{pool}

High entropy between customer & other high performance models

Flickering/uncertain customer model output on raw data stream snippets

Sudden appearance of undetected objects

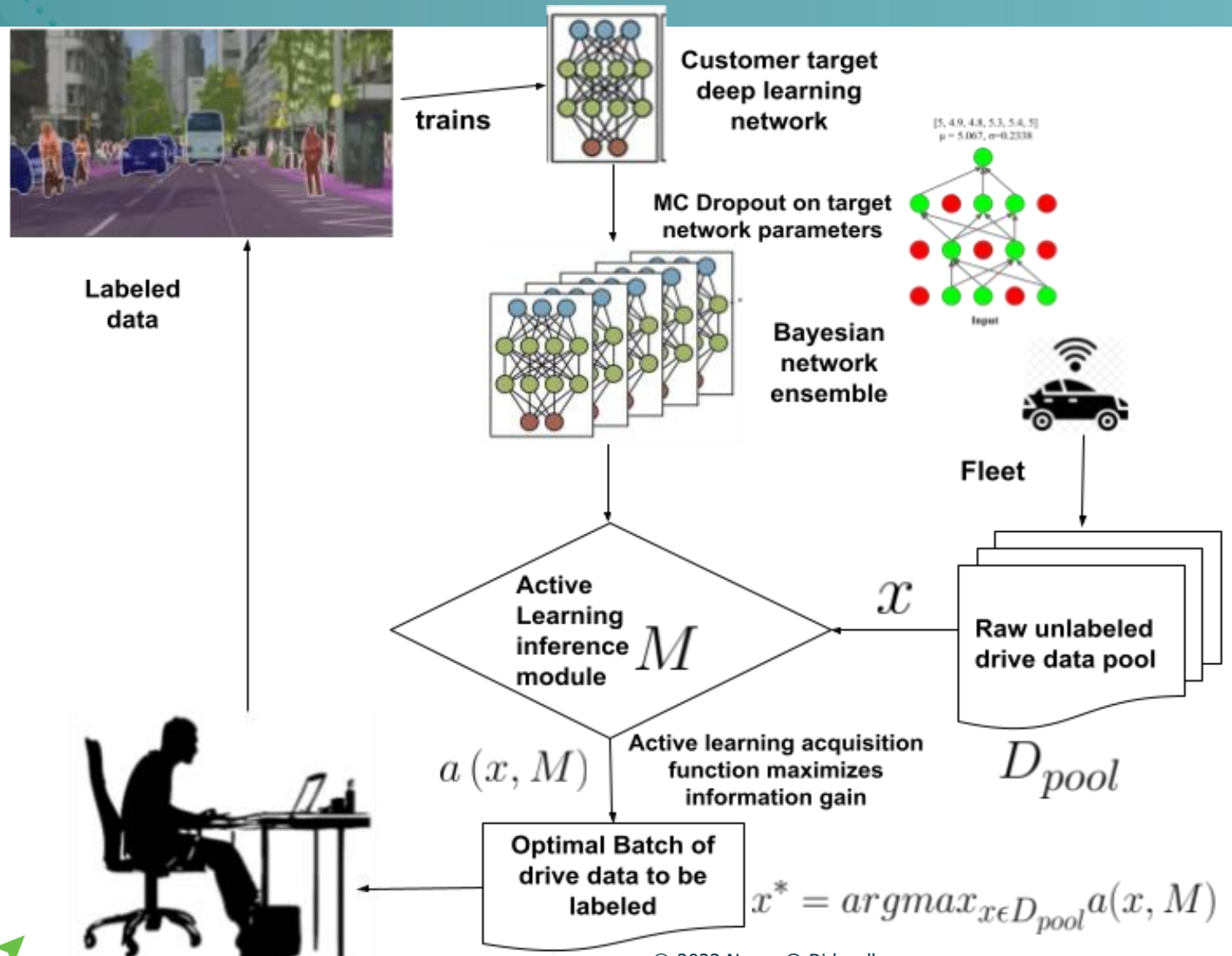
Optimal Batch of drive data to be labeled

Example is image data from production car fleet.

This workflow can be modified for any type of data.

RIDECELL

Data curation – active Bayesian learning



Deep learning confidence scores not reliable.

Convert a network to Bayesian network using MC Dropout.

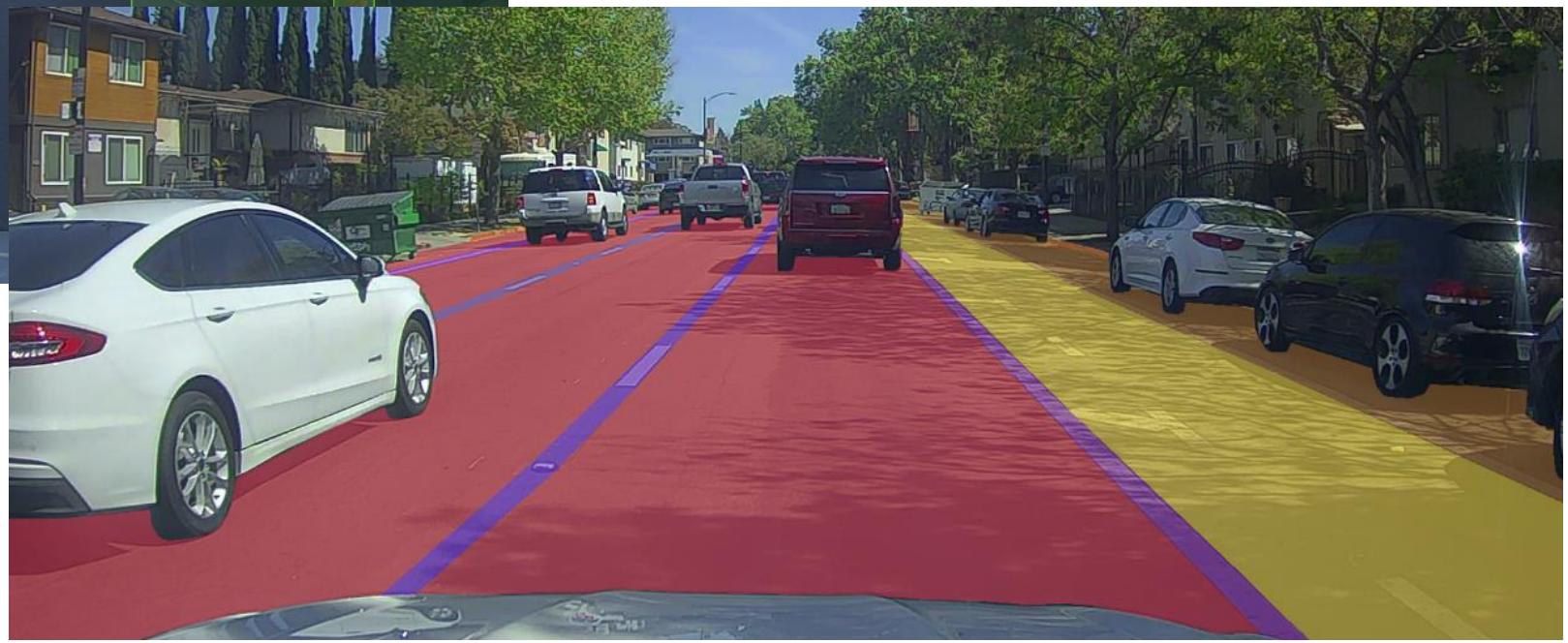
Data annotation

RIDECELL

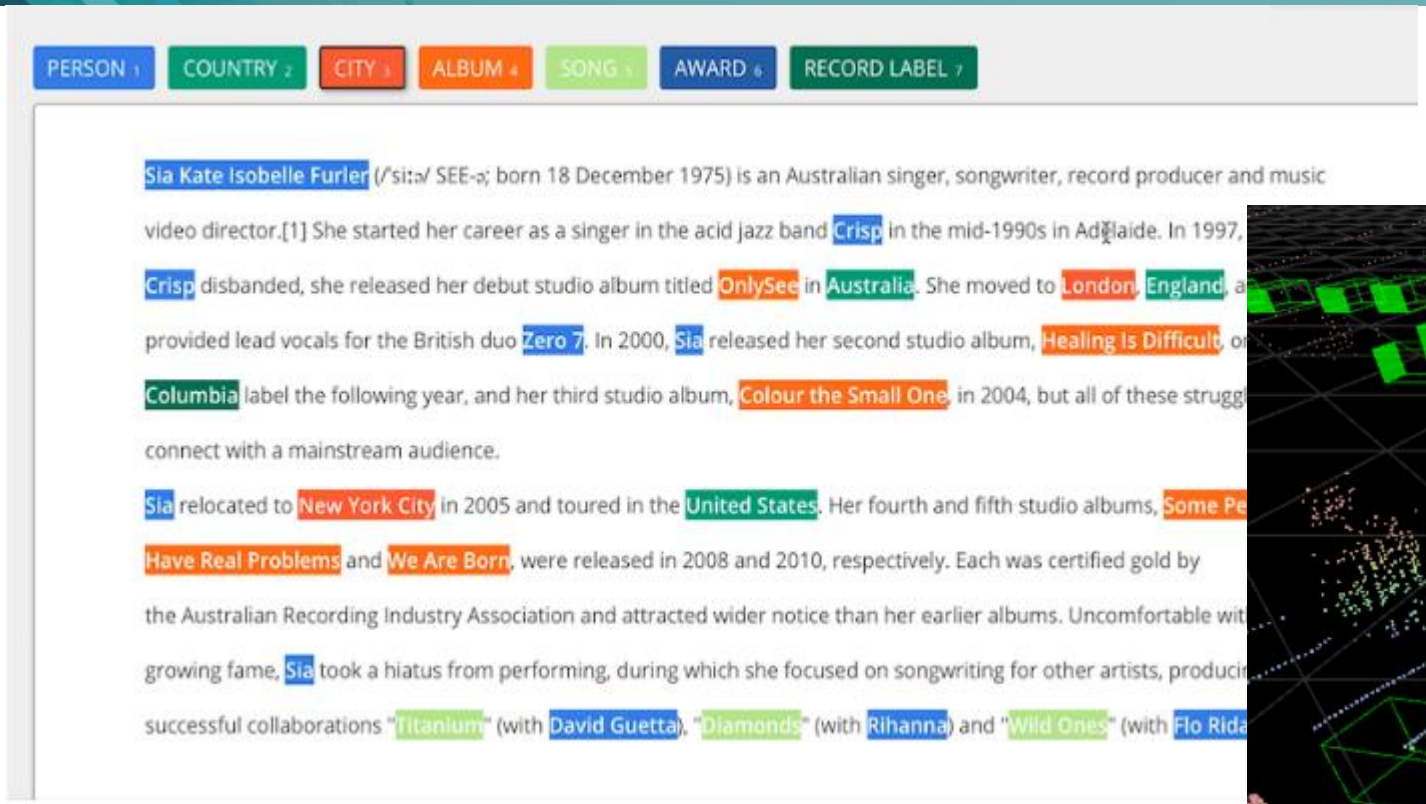
Data annotation – types of annotation
Image annotation

Bounding box

Semantic segmentation annotation

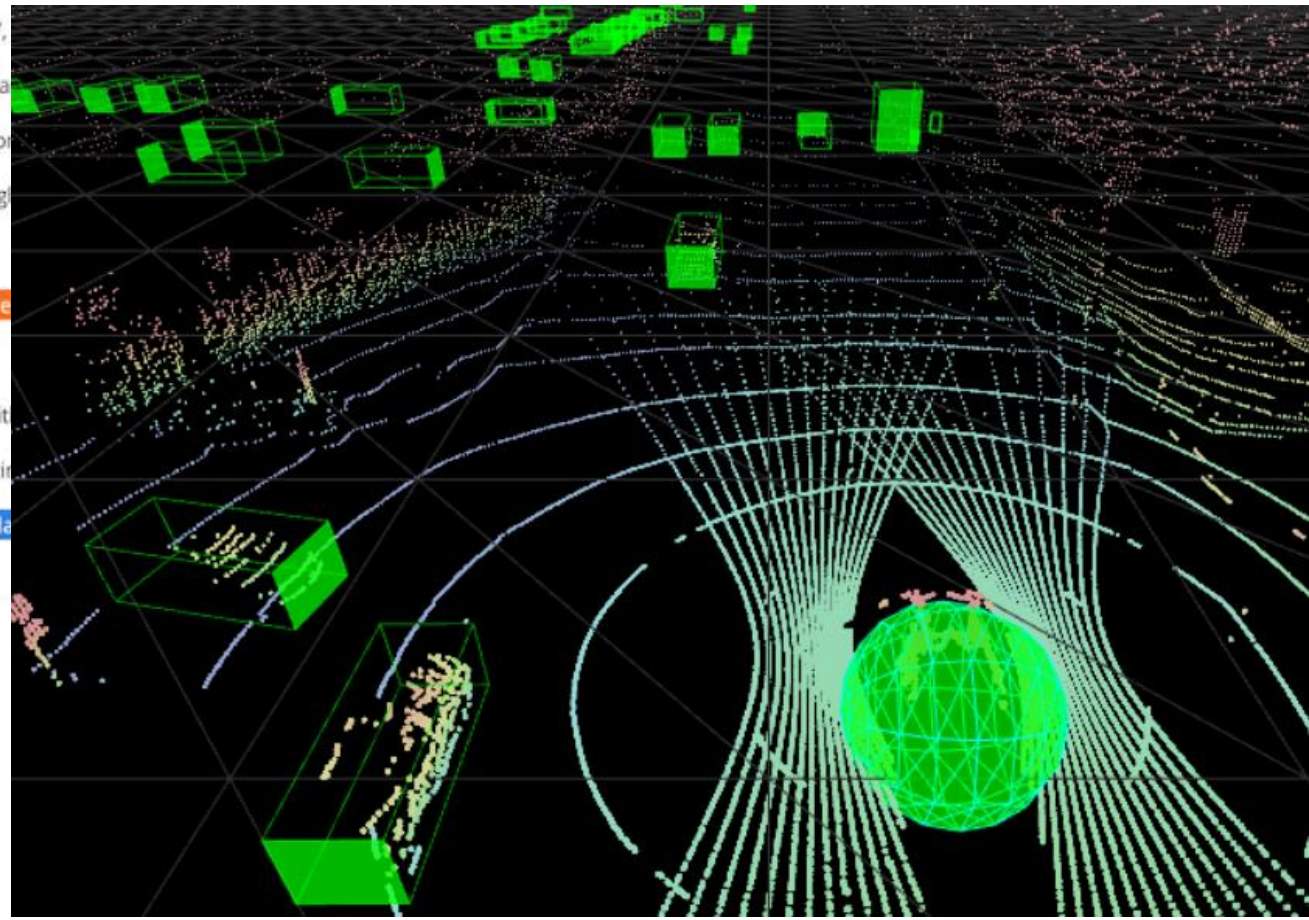


Data annotation – types of annotation
Other types

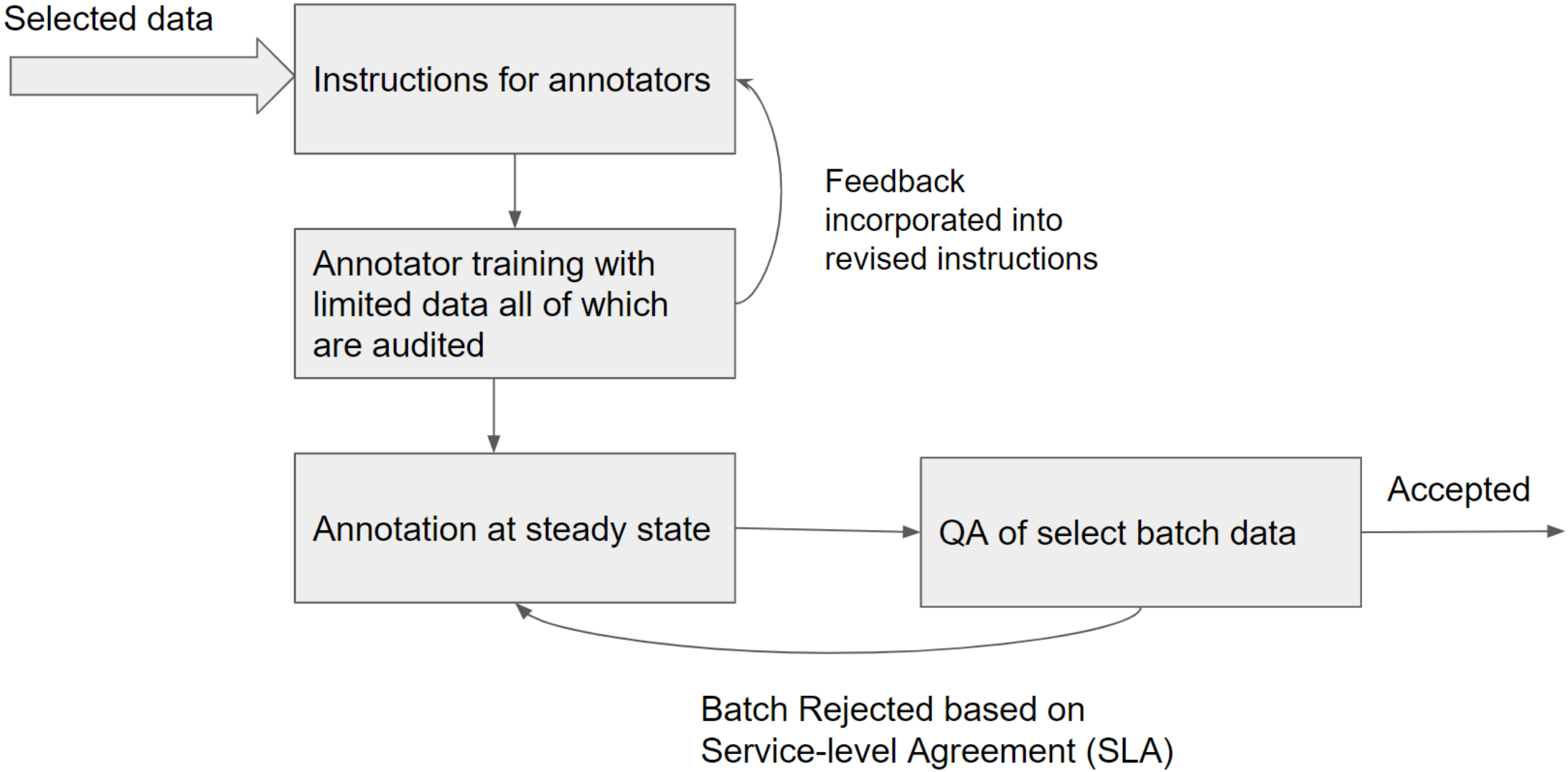


Text annotation image from Anolytics

Lidar 3D bounding box annotation



Data annotation – annotation process



Data annotation – annotation instructions

Ambiguous situations explained

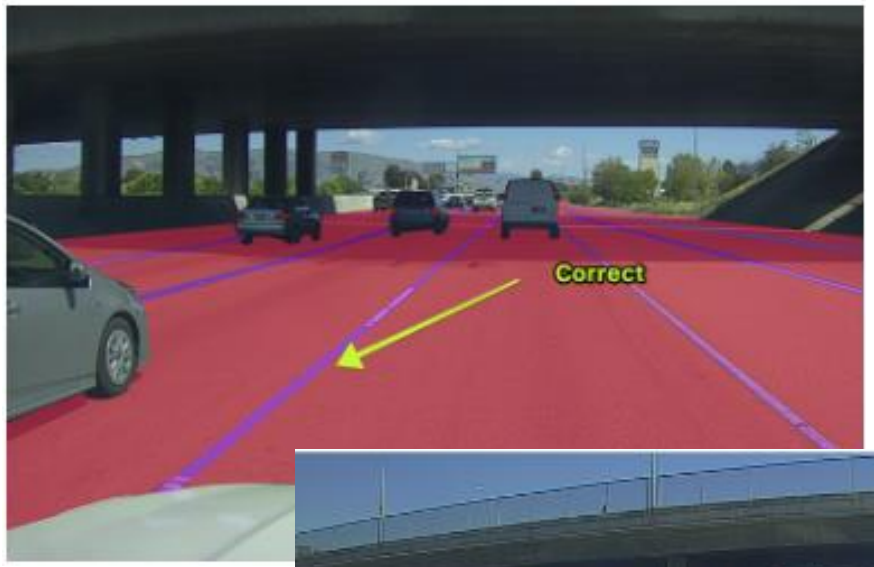
Correct sun glare

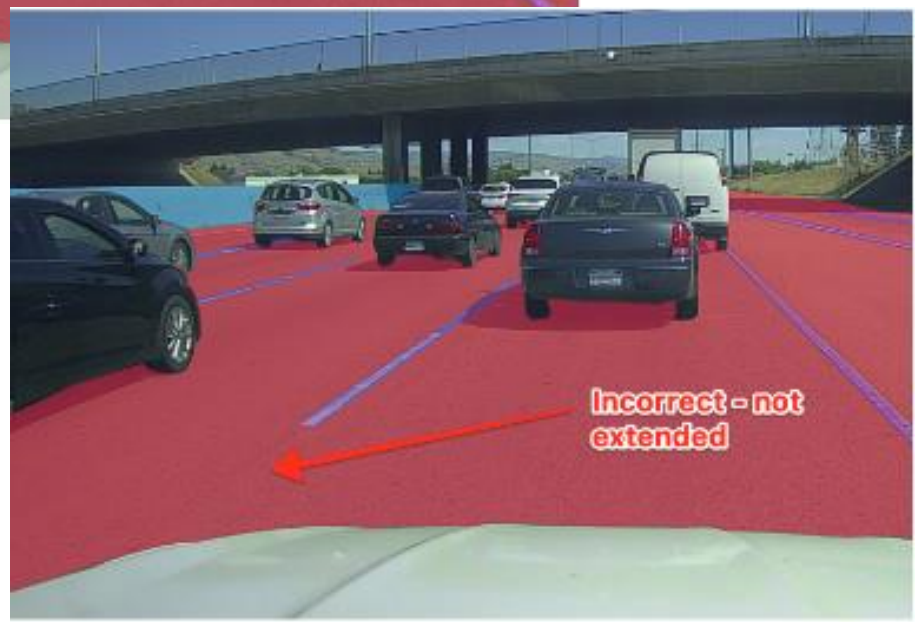
No sun glare – even though there is slight glare and flaring

RIDECELL

Data annotation – annotation instructions

Ambiguous situations explained





RIDECELL

Data annotation – annotation instructions

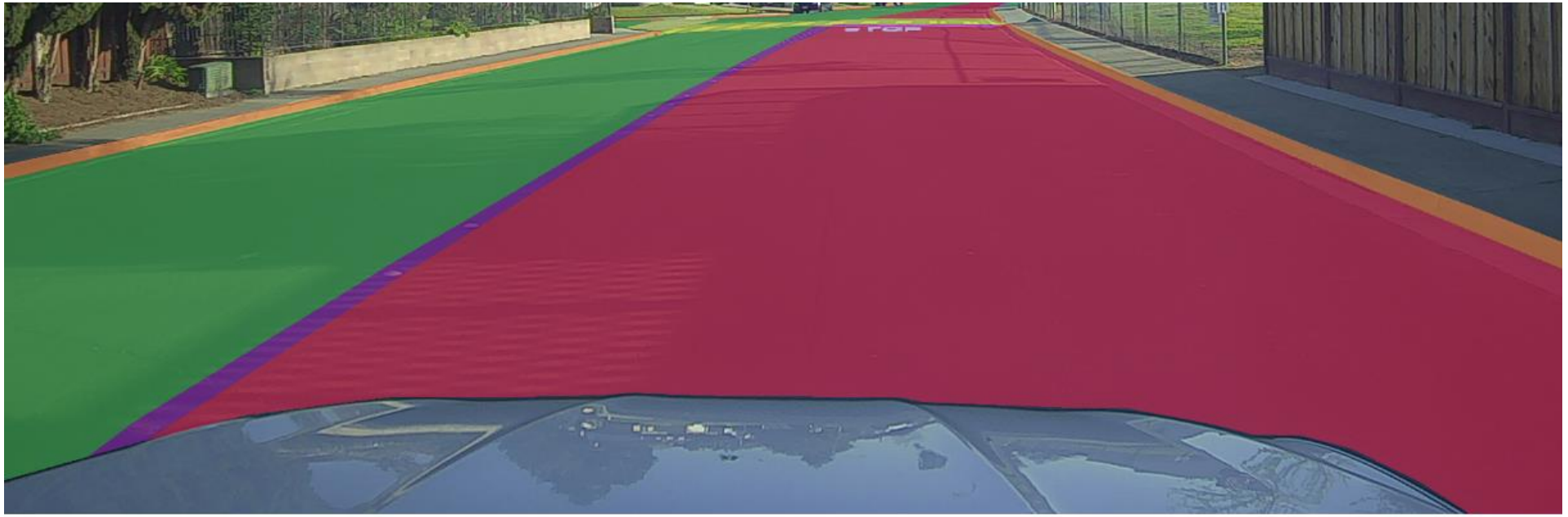
Scenario instructions example

Interaction event	Description	Ego green/Vehicle blue
interaction:anti_parallel:cut_in:from_left	Object in adjacent lane, moving in opposite direction to ego and cutting into ego lane	
interaction:anti_parallel:turn_across:to_right	Object in adjacent lane, moving in opposite direction to ego and turning across ego lane to right	
interaction:anti_parallel:head_on	Object in adjacent lane, moving in opposite direction to ego and head on moving towards ego vehicle	
interaction:anti_parallel:cut_out:to_left	Object in adjacent lane, moving in opposite direction to ego and cutting out to left of ego	
interaction:anti_parallel:turn_out:to_left	Object in adjacent lane, moving in opposite direction to ego and turning out to perpendicular lane to left of ego	

Scenario annotation instructions example.

Clear visualizations and text description for annotators.

Data annotation – auditing annotations
Clear and precise feedback



Status: **REJECTED** Audited: **7 MONTHS AGO**
Comments: **STOP LINE 2 IS WRONGLY MARKED.**

Data annotation – auditing annotations
Clear and precise feedback

us: **REJECTED** Audited: **A YEAR AGO**
ments: **BOUNDING BOX FOR CAR B771 IS WRONG.. OTHER BOUNDING BOXES HAVE MINOR ERRORS**

RIDECELL

Data annotation – annotation in-house vs managed

	In-house annotation	Managed service
Quality	Generally, can be maintained higher	Depends on provider. If done with freelancers can be low
Time to setup team	Longer time to setup team.	Can be faster with managed service provider.
Tooling	Needs to purchased / developed in-house. Good options are now available	Good tooling through annotation companies.
Scaling up	Difficult	Easier as providers have bigger workforce.

Data annotation – common issues

Quality annotation is HARD!

- Tooling issues
- ML assistance to annotation – can reduce quality!
- Bad annotation instructions
- QA in managed services is automated
- Annotator burn-out

Conclusion

Conclusion

Key Takeaways!

- Invest specifically in tooling for collection, tagging, annotation.
- Active learning is key during the curation stage.
  - Continuous improvement using data from production systems.
- Synthetic data for those hard-to-find corner cases.
- Pay close attention to annotation instructions

Resources to learn more

Resources

Active learning for deep learning

https://jacobgil.github.io/deeplearning/activelearning

Tools for ML data management

https://scale.com/nucleus

https://nemosearch.ai/

https://voxel51.com/