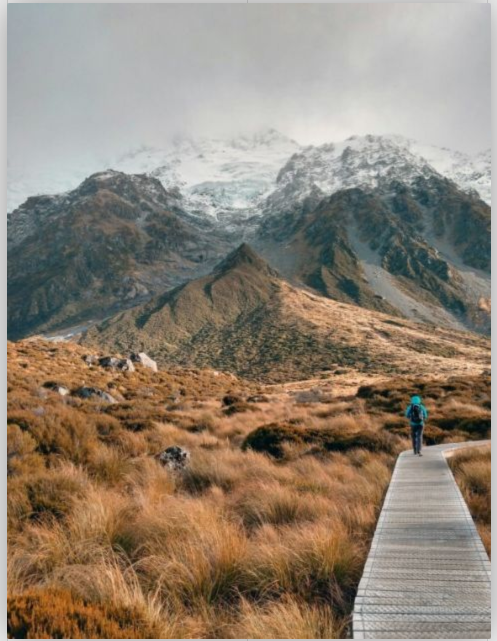
Saturday, 15 Nov | Online

AI Data Labeling: The Bottleneck in AI Development

Puneet Jindal
Founder & CEO, **Labellerr**

Session Guidelines



01 Post queries in the chat

I'll address them as we go or at the end

02 Keep Cameras On

for better engagement

03 Ask questions

stay focused on today's topic

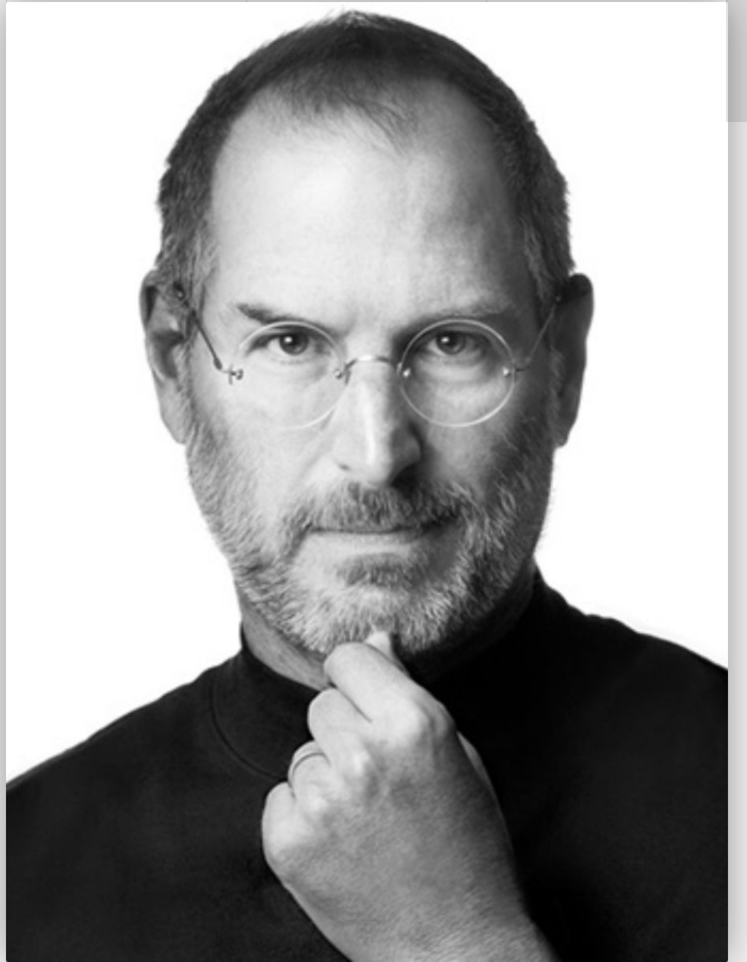
04 Stay on Mute

Unmute only when speaking to minimize background noise.

You’ve got to start with the customer experience and work backward to the technology. You can’t start with the technology then try to figure out where to sell it.

- Steve Jobs

CEO, Apple



My Journey to Labellerr

01
graduate from Thapar
2010

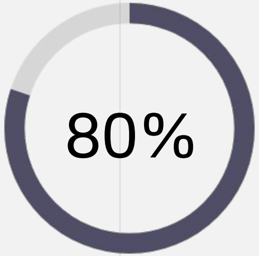
02
Startup journey with Delhivery
2012

03
Started my data and AI journey
2014

04
Indian School of Business
2015

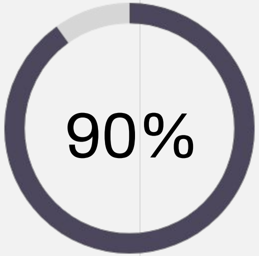
05
Walmart, Lybrate, Ephesoft, RateGain
2016-2019

The Key Insights!



AI projects failure

Its important to deliver ROI



Unstructured data

Such as images, videos, audio, etc

Labellerr in 2024: Serving Diverse AI Needs

Large Enterprises

- Toyota Research Institute *(Robotics Learning)*

Innovative Startups

- SpotAI *(Surveillance)*
- Spare-it *(Waste Management)*
- Mythos AI *(Self-driving)*
- Wadhwani AI

Leading Academic Institutions

- University of Maryland
- Baylor College of Medicine
- Stanford University *(Ecology)*

Robotic Learning case - Boston Dynamics and Toyota Research Institute https://vimeo.com/1023820181

CCTV based Surveillance case - SpotAI

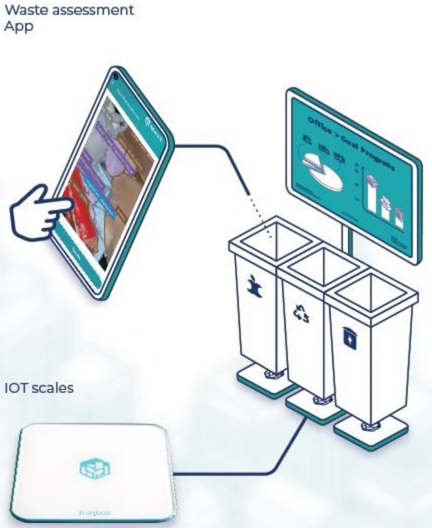
spare-it

Sustainable Building Operations Award Oct 2023

Waste Intelligence Platform

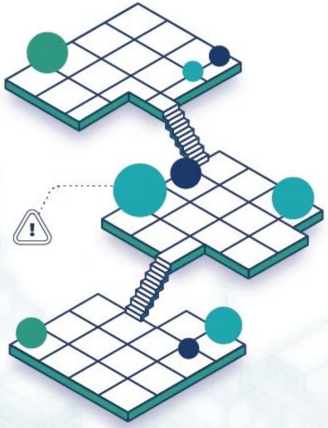
Monitor

Connect your bins



Discover

Floor heatmap



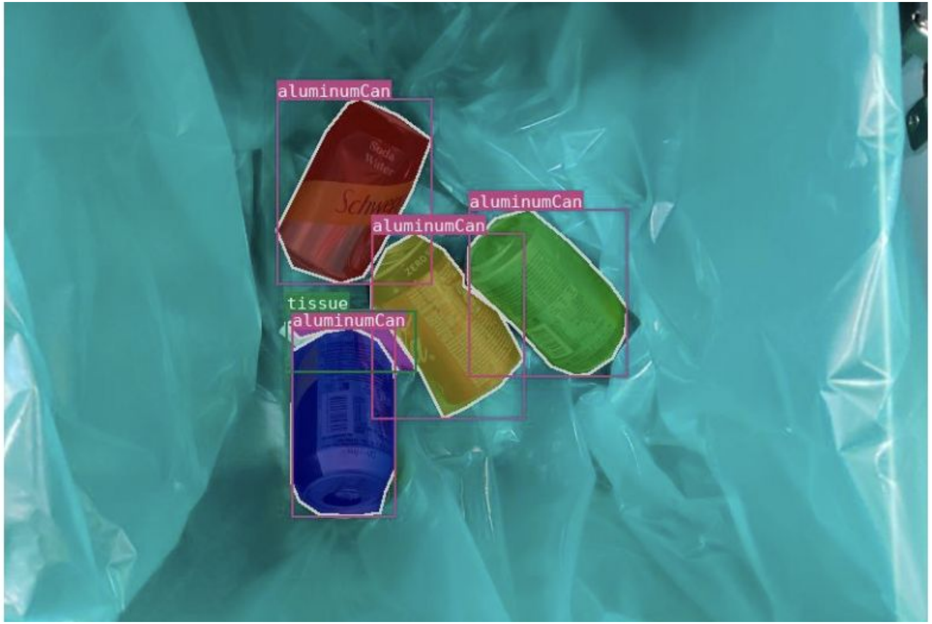
Transform

Reduce and Report (API)

Reduce carbon footprint

Encourage circularity

Automatic reporting



Example1:

In this simple example a picture is taken from a “metal” bin from a customer in Hong Kong, 5 items have been identified.

1) 4 metal cans
2) 1 tissue

The tissue is a contaminant in the “metal” waste stream. The contamination from this photo is estimated to ⅕ = 20%.

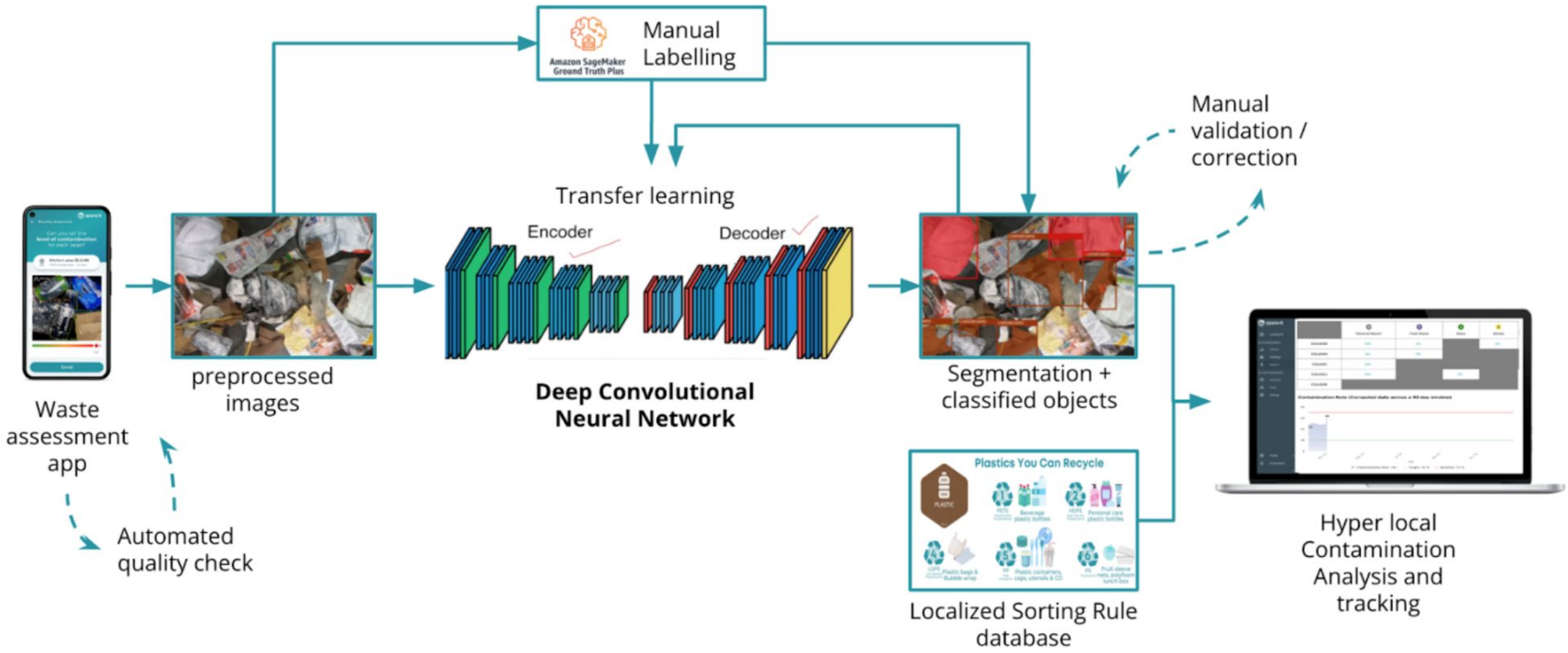
Example2:

In this simple example a picture is taken from a “plastic” bin from a customer in Hong Kong, 7 items have been identified.

1) 3 other plastics
2) 1 snack chips bag
1) 1 compostable hot cup
2) 1 cardboard
3) 1 beverage carton

The beverage-carton, compostable-hot-cup and cardboard are contaminants in the “plastic” waste stream. The contamination from this photo is estimated to 3/7 = 42.9%.

Taking an example of a computer vision - waste contamination tracking



Data Labeling Life Cycle

Overview of the Life Cycle

Step 1

Define Objectives

Align labeling goals with model outcomes (e.g., classification, object detection, sentiment analysis).

Step 2

Select and Prepare Data

Choose the relevant data types (text, images, videos).

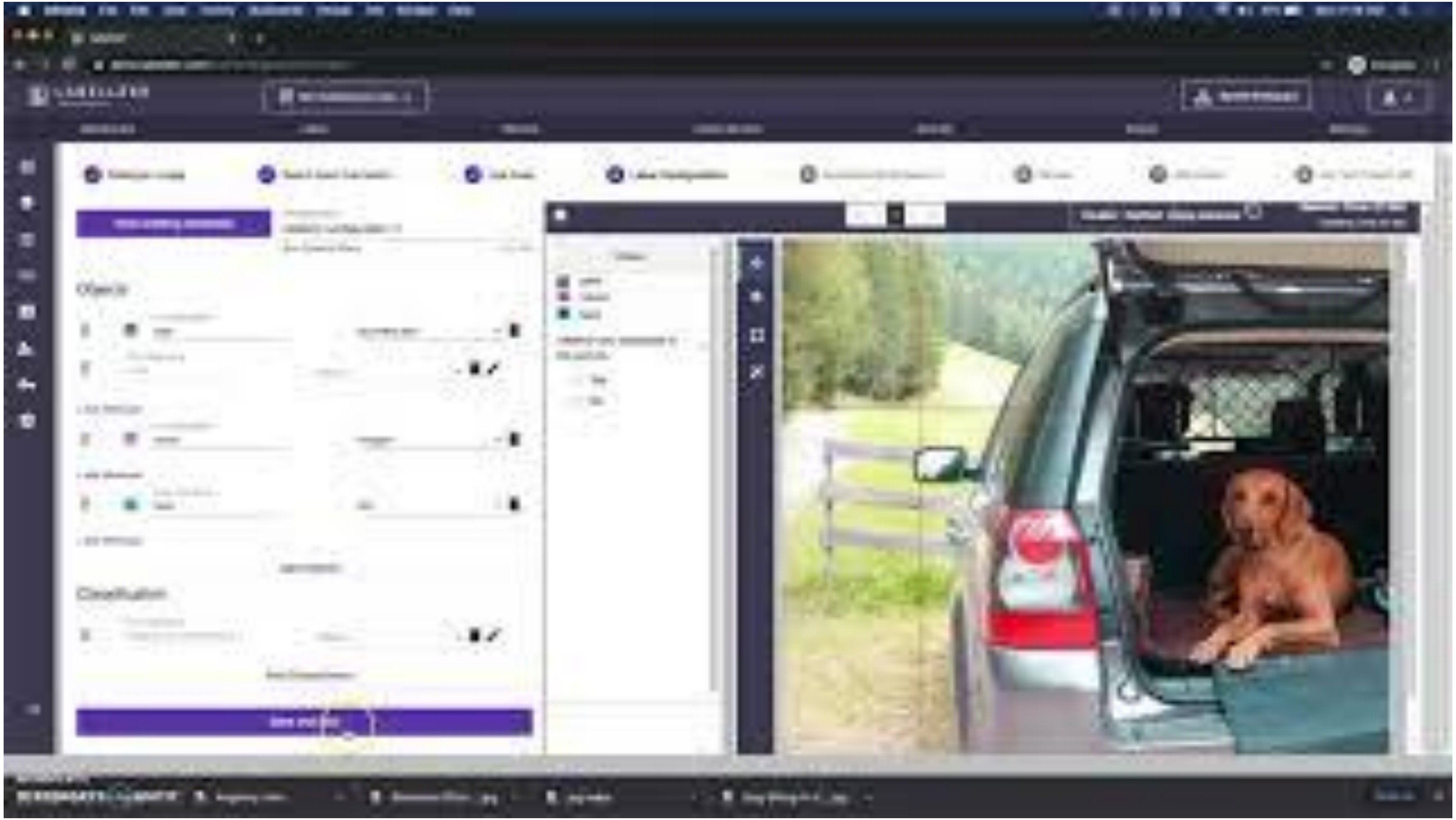
Step 3

Annotation Guidelines

Define clear guidelines for annotators to ensure consistency.

Name	Id	Material Class	Signage Illustration	Sample Pictures (from ...	Visual Description
Paper Towel/Napkins/Tissue/Tissue Paper	24	Compostables			Thin low quality paper More textured than printer paper, may have food stains
Wooden Coffee Stirrer or Utensil or Cho...	25	Compostables			Wooden stick like a popsicle stick or wooden cutlery or wooden chopsticks or
Soiled Cardboard Box	26	Compostables			Cardboard with food or water staining, usually pizza boxes
Compostable Plastic Lid	27	Compostables			clear plastic lid, adjacent to a compostable container, usually paired with comp
Food Soiled Paper	30	Compostables			Paper with food or grease staining - this DOES NOT include compostable or tr
Other compostable material	125	Compostables			Compostable bags, biodegradable packing peanuts
Sandwich paper wrapper	126	Compostables			Paper used to wrap sandwich, can be soiled; compostable
Batteries	38	E-waste			batteries
Cables	40	E-waste			Wire taht will tangle, usually connected to electronics, looks like phone chargi
Computers	43	E-waste			computer
Monitors	44	E-waste			monitor
Toner and Ink Cartridges	45	E-waste			black plastic container holding ink or toner from a printer
Miscellaneous Electronics	46	E-waste			anything else electronic
LED Lightbulb	47	E-waste			Lightbulb, but with electronic components. Usually frosted
Meat and Fish	48	Food Waste			any meat
Bones and Shells	49	Food Waste			animal bone
Cheese and Other Fats	50	Food Waste			Cheese or grease
Fruits And Veggies	51	Food Waste			Fruit of any kind
Other Food or Mixed Food	52	Food Waste			

Annotation Guidelines



Demo of manual annotation on Labellerr

Data Labeling Life Cycle

Overview of the Life Cycle

Use a mix of automated QA and manual review to maintain data integrity.

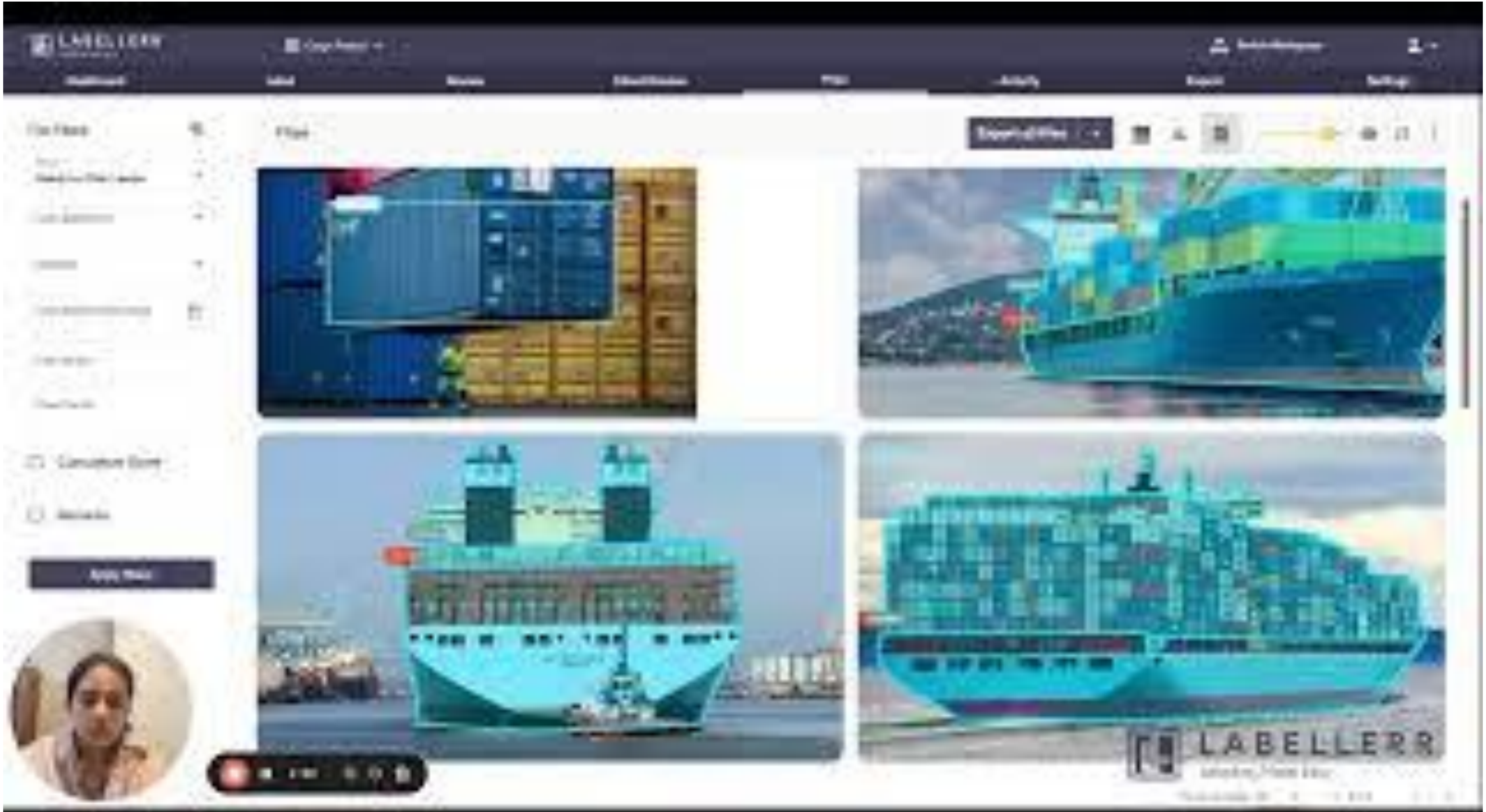
Quality Assurance

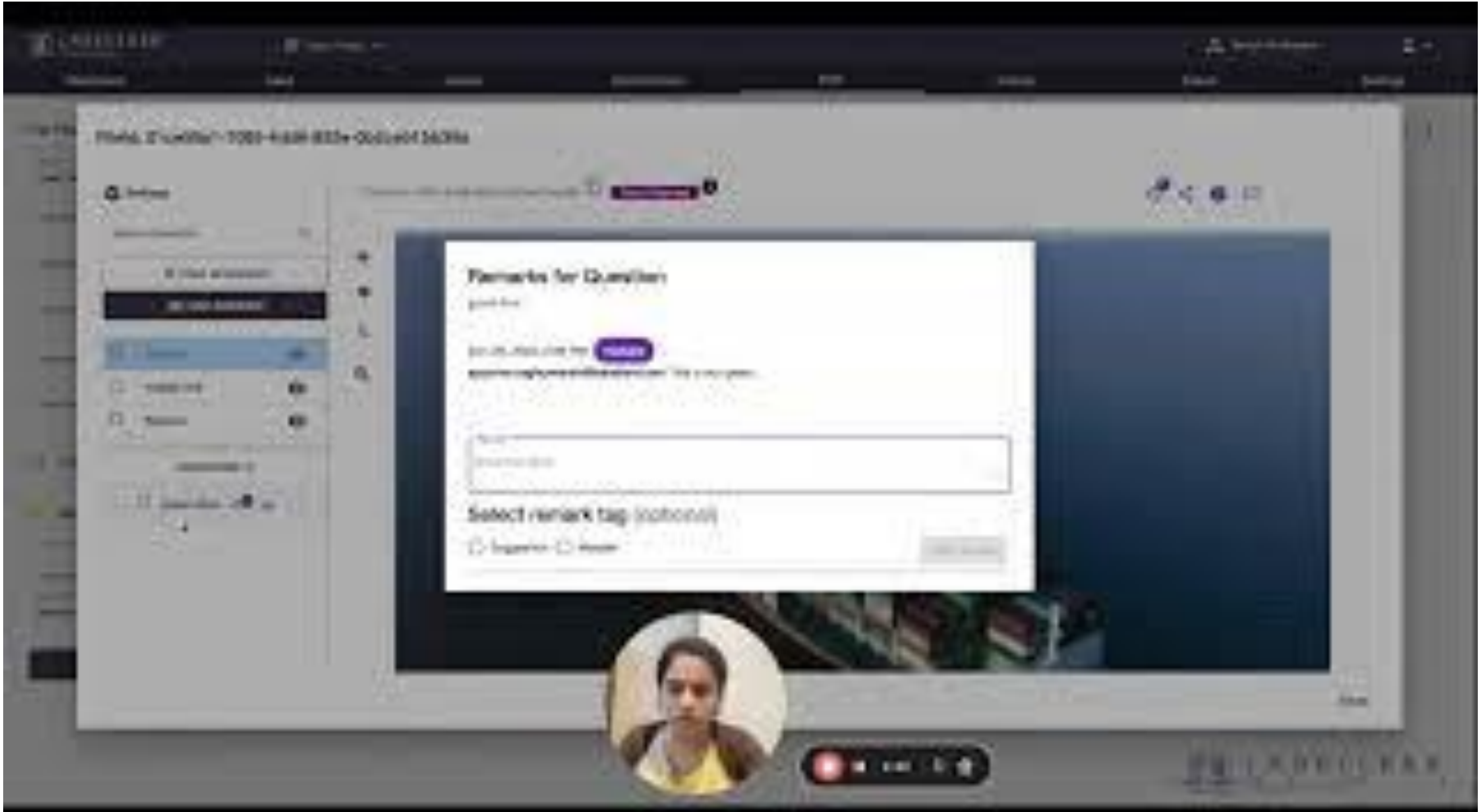
Step 4

Step 5

Feedback Loop

Implement feedback mechanisms for continuous improvement of annotations.


LABELLERR



Remarks for Question
Select remark tag (optional)

Common Pitfalls in data Labeling

1. **"Labeling is just about drawing boxes or polygons."**
   Labeling requires nuanced understanding beyond simple shapes; accurate annotations depend on context, clarity, and adherence to specific guidelines.
2. **"Basic guidelines in a document are enough for high-quality annotations."**
   While initial guidelines are essential, ongoing clarification, regular updates, and interactive feedback are key to consistent quality across complex projects.
3. **"Training annotators once ensures perfect quality."**
   Continuous training and periodic quality checks are necessary, especially as task complexity and annotator expertise vary, impacting annotation quality.
4. **"Volume spikes can be handled instantly with increased speed."**
   Scaling up annotations quickly can compromise quality, as larger volumes require more robust management, including quality control and oversight mechanisms.
5. **"Once accuracy is achieved, annotation can run on autopilot."**
   Consistent supervision and iterative quality checks are needed to maintain standards, as even well-performing pipelines may need adjustments over time.
6. **"Data should be clean by default."**
   Raw data frequently includes noise, ambiguities, and unexpected complexities that need to be addressed through pre-processing and clarification for effective labeling.
7. **"Data security and compliance are secondary concerns."**
   Maintaining data privacy and security is critical in every annotation project, as sensitive data must adhere to compliance standards to protect individuals and organizations.
8. **"Data is unbiased and won’t need distribution adjustments."**
   Data bias is often unintentional and can evolve; continuously assessing and adjusting for fair representation and balanced distribution is essential for robust model performance.

Different Labeling Types in Images

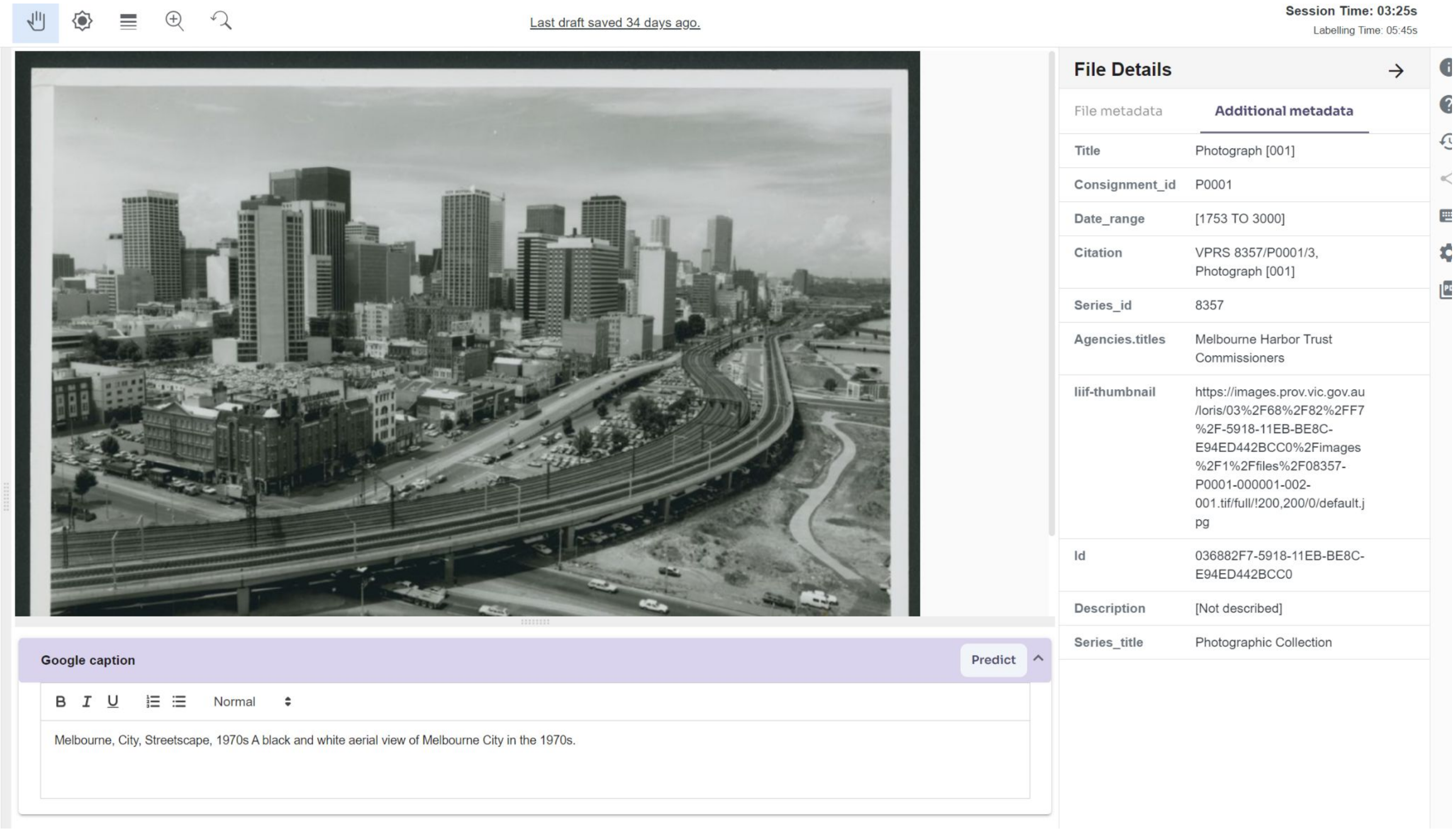
Model Assisted Labeling

1. Off the shelf foundation models for kick start such as Gemini, Llama
2. Fine Tuning models
3. Active learning workflow for data selection
4. SAM and SAM2 for generic polygon annotation speed up

Other approaches to generate training data

Synthetic data generation


Settings
SUBMIT



Other Use cases - GenAI based annotation for describing archived photographs

Automated labeling and QC with upto 90%

Synthetic datasets

Refurbished datasets marketplace

Currently - 80% outsourced

Future - 80% in house

Thanks for Being Here!

Appreciate your queries

Linkedin
linkedin.com/puneetjindalisb

Twitter
twitter.com/puneetjindalisb

Location
651 N Broad St, Suite 201, Middletown, New Castle, DE, 19709, US