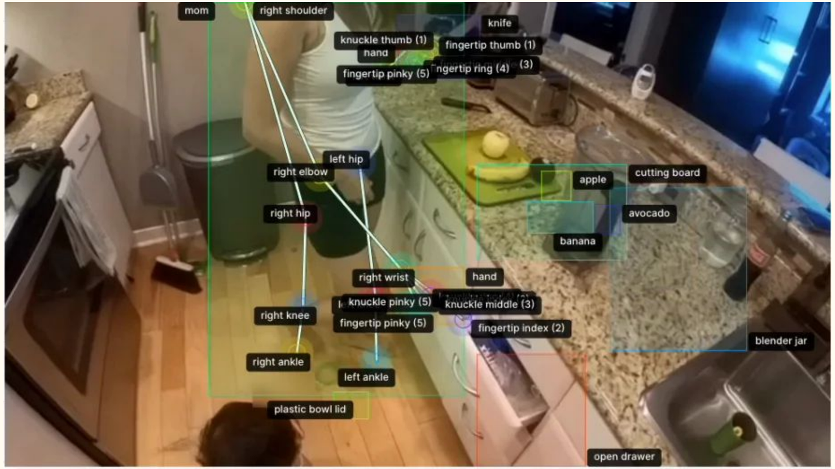
Data Labeling Platforms

Nihaar Charagulla, Anoop Jalla, Akshay Mistry

What is Data Labeling

- **Definition**: The process of tagging raw data, whether it's text, images, or videos—with meaningful labels.
- **Types of data labeled**: Images, text, videos, audio.
- **Importance:**
  - Enables **supervised learning** in AI.
    - High-quality labeled data is required for accurate AI models.
    - AI systems need labeled data to understand patterns and make decisions.
- **Example**: In self-driving cars, we need to identify objects like pedestrians, stop signs, and traffic lights. Without labeled data, machine learning models cannot be trained effectively.



https://scale.com/guides/data-labeling-annotation-guide

Challenges in Data Labeling

1. **Time-consuming & Expensive**
   a. Manual labeling requires large amounts of human effort.
   b. High costs for skilled annotators (e.g., medical professionals for radiology images).
2. **Requires Domain Expertise**
   a. Specialized knowledge is needed for fields like **healthcare, legal documents, and finance**.
   b. Example: Annotating MRI scans requires radiologists, increasing costs.
3. **Data Bias & Ethical Concerns**
   a. Bias in labeling can lead to unfair AI models (e.g., facial recognition misidentifying minorities).
4. **Quality Control & Accuracy Issues**
   a. Crowdsourced labeling may introduce **inconsistent annotations**.
   b. Ambiguous data can lead to incorrect labeling (e.g., sarcasm in text sentiment analysis).
5. **Scaling Challenges**
   a. As AI models require more data, manually labeling millions of samples is **unsustainable**.
   b. AI-assisted labeling helps but is still **not perfect**.

Product Overview: Key Features

- Annotation with audio, text, images
  - Speech-to-text, voice recognition
  - Text tagging, sentiment analysis
  - Bounding box, segmentation
- AI-labeling
- API service to integrate with cloud services and external software
- Efficiently organizing labeled data
- Human review of automated labeling

Product Overview: Comparison

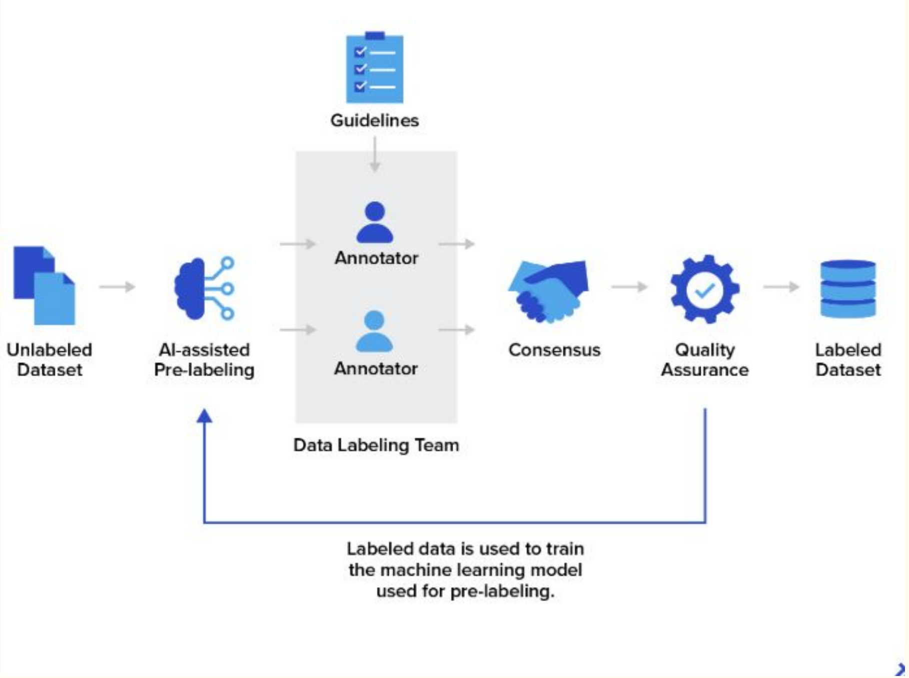
These are some of the most popular data labeling platforms. They are slightly different, though they offer the same core use cases with some alterations in implementation and deployment

Feature	Labelbox	Amazon SageMaker Ground Truth	Scale AI
AI-Assisted Labeling	Yes	Yes	Yes
Human Review	Yes	Yes	Yes
Cloud Integration	No	Yes	Yes
Open-Source	Yes	No	No

Technical Details

Scale AI

- AI-powered pre-labeling
- Human reviewing to supervise AI-labeling
- API integration to mass-ingest data
- Powerful enough to support autonomous vehicles, medical image recognition, and location mapping
- Security GDPR and HIPAA standards
- Highly scalable



https://www.toptal.com/artificial-intelligence/how-to-label-data

Sample Applications

Recommendations

- **Amazon (E-commerce)**
- **H&M (Retail)**
- **Netflix (Entertainment)**

Autonomous Agents

- **Waymo (Self-driving cars)**
- **PathAI (Healthcare assistants)**
- **Skydio (Drones)**

Case Study: Netflix Recommendation System

- Label data to classify movies/show by genre, themes, and user preference
- Track user behavior to optimally suggest content
- Tagging data based on user interactions to perpetually train models on user preference
  - Clicks, watch time, engagement

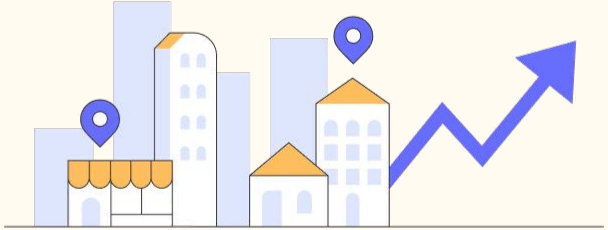
Impacts

- Increases engagement to increase viewership and subscriptions
- Personalized content to provide accurate suggestions

- Bias in data labeling per user
- User unable to escape old preferences

Market Analysis

- **Key players**:
  - **Amazon SageMaker Ground Truth**: Market leader due to AWS integration.
  - **Scale AI**: Preferred by autonomous vehicle and defense companies.
  - **Labelbox**: Popular for flexibility and user-friendly annotation tools.
  - **SuperAnnotate**: Gaining traction with automation and AI-assisted labeling.
- **Annual Revenues**
  - **Market size**: Expected to exceed **$5 billion by 2027**.
  - **Key revenue drivers**: Growth in AI adoption, increasing need for labeled data.
- **Marketing Strategies**
  - **Enterprise focus**: Selling to companies in AI-heavy industries (e.g., Tesla, OpenAI, healthcare).
  - **Subscription Models**:
    - Labelbox: Free tier for small teams, enterprise plans for large-scale usage.
    - Scale AI: Custom pricing for high-volume labeling.



Future Trends

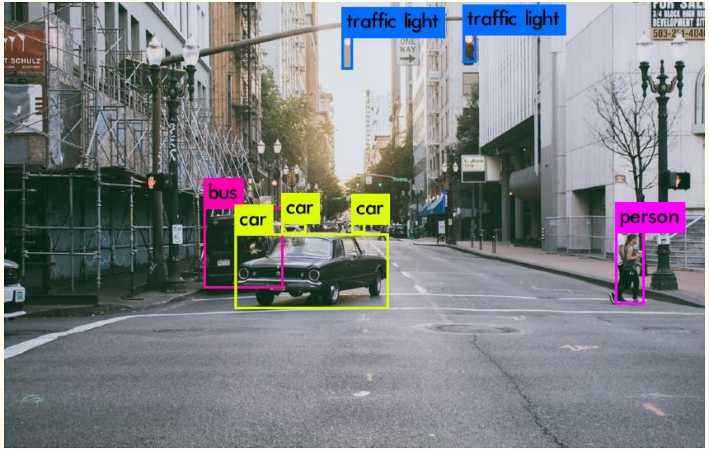
The future of data labeling is rapidly evolving, with AI taking on more of the heavy lifting. Instead of humans manually labeling massive datasets, new initiatives are being developed to boost efficiency, privacy, and fairness

Reducing Manual Labeling

- AI models increasingly handle annotation tasks.
- Active learning loops improve labeling over time, boosting efficiency and reducing the need for human input.
- Human-in-the-loop still needed for complex edge cases.

Synthetic Data

- AI generates realistic, labeled training data.
- Useful for rare event modeling (e.g., autonomous vehicle crashes, fraud detection).
- Synthetic data helps AI models generalize better and reduces reliance on massive, manually labeled datasets.



https://ied.eu/blog/data-labeling-for-ai-5-important-considerations-to-accelerate-quality/

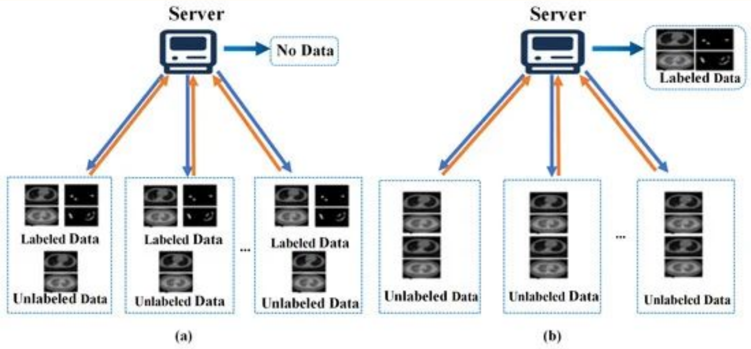
Future Trends

Federated Learning

- Model training occurs on distributed devices without sharing raw data.
    a. Each device shares local model parameters with a server, the server uses the parameters to learn a global model, and the server then sends updated parameters back to the devices.
- Used in privacy-sensitive industries (e.g., healthcare, finance).
- Reduces risks of data breaches and regulatory concerns.
- Allows labeling models to be personalized based on user-specific data while keeping personal information private.

Fairer AI Models

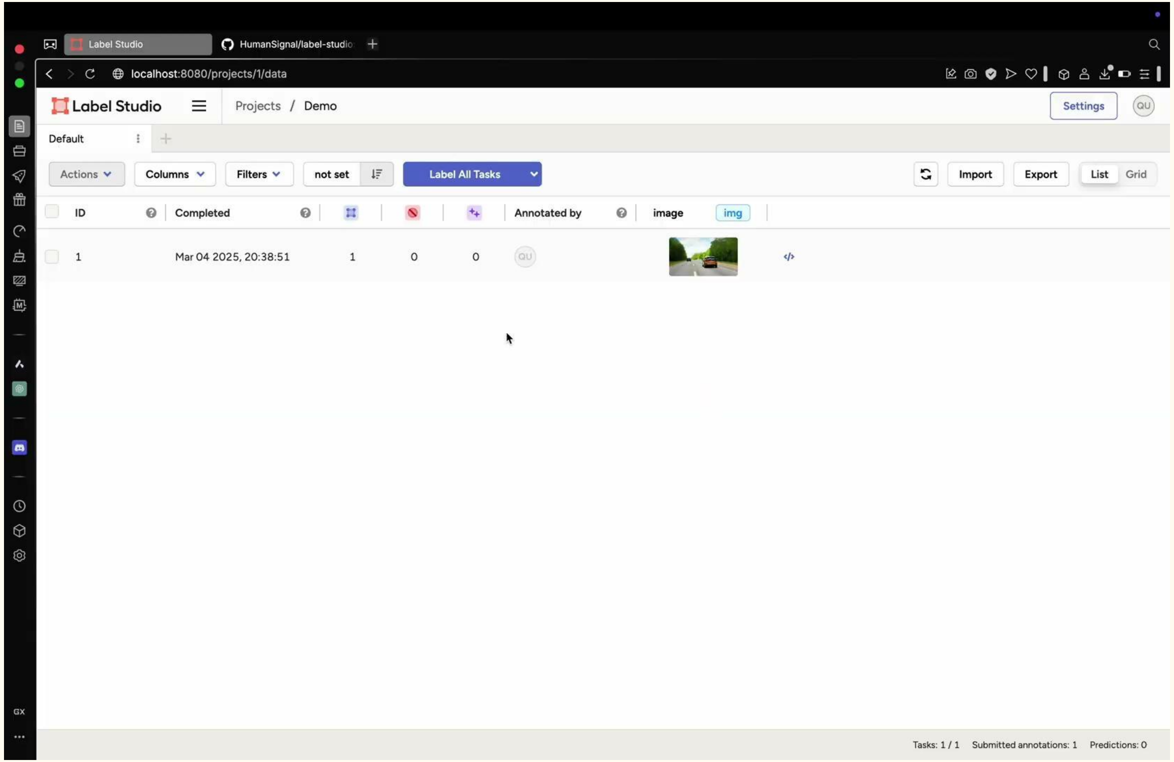
- AI models are learning to detect and mitigate bias.
- Synthetic data can help balance underrepresented groups in datasets.
- Federated learning ensures diverse training data without violating privacy.



Two scenarios in federated semi-supervised learning.
(a) Labels-at-Client Scenario (b) Labels-at-Server Scenario

https://www.mdpi.com/2079-9292/12/7/1687

Live Demo



ID	Completed				Annotated by	image img	
1	Mar 04 2025, 20:38:51	1	0	0	QU		</>

Tasks: 1 / 1 Submitted annotations: 1 Predictions: 0

http://localhost:8080/projects/1/data?tab=1

https://github.com/HumanSignal/label-studio