Data Annotation Glossary

A

- **Active Learning:** A machine learning technique where an algorithm selects the most uncertain data points for annotation, reducing labeling time. As an annotator, you might prioritize these samples.
- **Annotation:** The process of labeling or tagging data (e.g., images, text, audio) to make it usable for training AI models. Example: Drawing boxes around objects in an image.

B

- **Bounding Box:** A rectangular shape drawn around an object in an image to label its location and category (e.g., "car"). Common in object detection tasks.
- **Bias (in Annotation):** Systematic errors in labeling that can skew AI model outcomes, often due to annotator subjectivity (e.g., labeling "cloudy" vs. "partly cloudy").

C

- **Classification:** Assigning a category or label to data. Example: Labeling an email as "spam" or "not spam."
- **Computer Vision:** A field of AI that enables machines to interpret visual data (e.g., images, videos). Annotators often label images for computer vision tasks like facial recognition.

D

- **Dataset:** A collection of data (e.g., images, text) used to train or test an AI model. Your annotations create the "ground truth" for these datasets.
- **Domain Knowledge:** Expertise in a specific field (e.g., medical imaging, automotive). Annotators with domain knowledge can label specialized data more accurately.

E

- **Entity Recognition:** Identifying and labeling specific entities in text, such as names, dates, or locations. Also known as Named Entity Recognition (NER) in NLP.
- **Evaluation Metrics:** Measures to assess annotation quality, like inter-annotator agreement (how much annotators agree on labels).

F

- **Feature Extraction:** Identifying key elements in data (e.g., edges in an image) that an AI model uses to learn. Annotations help define these features.
- **Freelance Annotation:** Working independently on platforms like Upwork or Appen, often with flexible hours but variable income.

G

- **Ground Truth:** The correct labels for a dataset, provided by human annotators, which AI models use to learn and evaluate performance.

I

- **Image Annotation:** Labeling images with tags, boxes, or masks to train computer vision models. Types include bounding boxes, semantic segmentation, and keypoints.
- **Inter-Annotator Agreement:** A measure of how often annotators agree on the same label for a given data point, used to ensure consistency.

K

- **Keypoints:** Specific points on an image (e.g., joints in a human body) labeled to track movement or structure, often used in pose estimation.

L

- **Labeling Tool:** Software used for annotation, such as LabelImg (for images), Prodigy (for text), or CVAT (for video). Familiarity with tools is key for annotators.
- **Label Noise:** Errors or inconsistencies in annotations that can reduce AI model accuracy (e.g., mislabeling a dog as a cat).

M

- **Machine Learning (ML):** A subset of AI where models learn from data to make predictions or decisions. Your annotations provide the training data for ML models.
- **Metadata:** Additional information about data (e.g., timestamp, source). Annotators may add metadata to enrich datasets.

N

- **Natural Language Processing (NLP)**: A field of AI focused on understanding and generating human language. Annotators label text for NLP tasks like sentiment analysis or entity recognition.
- NER (Named Entity Recognition): See Entity Recognition.

O

- **Object Detection:** Identifying and locating objects in images or videos, often using bounding boxes or polygons. A common task for annotators in computer vision.

P

- **Polygon Annotation:** Drawing irregular shapes around objects in an image for precise labeling, often used in semantic segmentation.
- **Pre-Labeling:** Using AI to suggest labels before human annotators review them, speeding up the process but requiring validation.

Q

- **Quality Control:** Processes to ensure annotation accuracy, such as cross-checking labels or using inter-annotator agreement metrics.

R

- **Reinforcement Learning from Human Feedback (RLHF):** A machine learning approach that uses human input to enhance AI systems. By integrating human feedback, it aligns AI behavior with human values.

S

- **Semantic Segmentation:** Labeling each pixel in an image with a category (e.g., “road,” “car”), creating detailed masks for computer vision models.
- **Sentiment Analysis:** Labeling text to determine its emotional tone (e.g., positive, negative, neutral), often used in NLP for social media analysis.
- **Synthetic Data:** Artificially generated data (e.g., via Unity) to train AI models, sometimes reducing the need for human annotation.

T

- **Text Annotation:** Labeling text data for NLP tasks, such as tagging entities, classifying sentiment, or identifying intent (e.g., “book a flight”).
- **Time-Series Annotation:** Labeling sequential data (e.g., sensor readings) to identify patterns, used in applications like IoT or finance.

V

- **Video Annotation:** Labeling frames in a video to track objects over time, often combining bounding boxes and tracking for tasks like surveillance or autonomous driving.