

# **Data Labeling & Annotation**

**CS 203: Software Tools and Techniques for AI**

Prof. Nipun Batra, IIT Gandhinagar

# The Labeling Bottleneck

## The Reality:

- Data is abundant (unlabeled).
- Labels are scarce (expensive).
- 80% of AI project time is Data Prep.

## Why Labeling Matters:

- **Supervised Learning:** Needs ground truth ( $y$ ).
- **Evaluation:** Even unsupervised methods need a test set to verify.
- **Ambiguity:** Labeling forces you to define your problem clearly.

# Types of Annotation

## Computer Vision:

1. **Classification:** "Cat" vs "Dog".
2. **Bounding Box:** Object Detection ( $x, y, w, h$ ).
3. **Polygon:** Segmentation (pixel-precise).
4. **Keypoints:** Pose estimation (skeleton).

## NLP:

1. **Text Classification:** Sentiment, Intent.
2. **NER (Named Entity Recognition):** Highlighting spans (Person, Org).
3. **Seq2Seq:** Translation, Summarization.

# Annotation Interfaces: Label Studio

**Label Studio:** Open-source, flexible, web-based tool.

**Configuration (XML):**

```
<View>
  <Image name="img" value="$image"/>
  <RectangleLabels name="tag" toName="img">
    <Label value="Car" background="red"/>
    <Label value="Person" background="blue"/>
  </RectangleLabels>
</View>
```

**Why usage XML?**

- Allows custom UI layouts.
- Can mix modalities (Image + Text + Audio).

# Quality Control: The Gold Standard

How do we know if labels are "correct"?

## 1. Gold Standard (Ground Truth):

- Experts label a small subset (e.g., 100 items).
- Annotators are tested against this set.

## 2. Consensus (Majority Vote):

- 3 annotators label the same item.
- If 2 say "Cat" and 1 says "Dog", label is "Cat".

# Inter-Annotator Agreement (IAA)

Measure of reliability. Do annotators agree with each other?

Percent Agreement:

$$\text{extAgreement} = \frac{\text{Agreed Items}}{\text{Total Items}}$$

*Problem:* Doesn't account for chance agreement.

Cohen's Kappa ( $\kappa$ ):

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

- $p_o$ : Observed agreement.
- $p_e$ : Expected agreement by chance.

# Cohen's Kappa Example

**Scenario:** 2 Annotators, 100 items (Yes/No).

- Both say Yes: 45
- Both say No: 45
- Disagree: 10
- $p_o = 0.90$

**Chance Calculation:**

- A says Yes 50% of time.
- B says Yes 50% of time.
- Chance they agree on Yes =  $0.5 \times 0.5 = 0.25$ .
- Chance they agree on No = 0.25.

# Managing Labeling Teams

## Workflow:

1. **Guidelines:** Write detailed instructions (e.g., "Does a reflection count as a car?").
2. **Pilot:** Label 50 items, calculate Kappa. Refine guidelines.
3. **Production:** Label large batch.
4. **Review:** Spot check 10% of labels.

## Human-in-the-Loop:

- Use Model to pre-label (Predictions).
- Humans verify/correct (faster than starting from scratch).



# Lab Preview

## Today's Goals:

1. **Install Label Studio.**
2. **Config:** Set up a Sentiment Analysis project.
3. **Label:** Annotate 10 examples.
4. **Export:** Get JSON/CSV data.
5. **Analysis:** Write a Python script to calculate Cohen's Kappa between two "simulated" annotators.

Let's start labeling!