DATA 3464: Fundamentals of Data Processing

Data Labelling and Augmentation

Charlotte Curtis

April 2, 2026

Topic overview

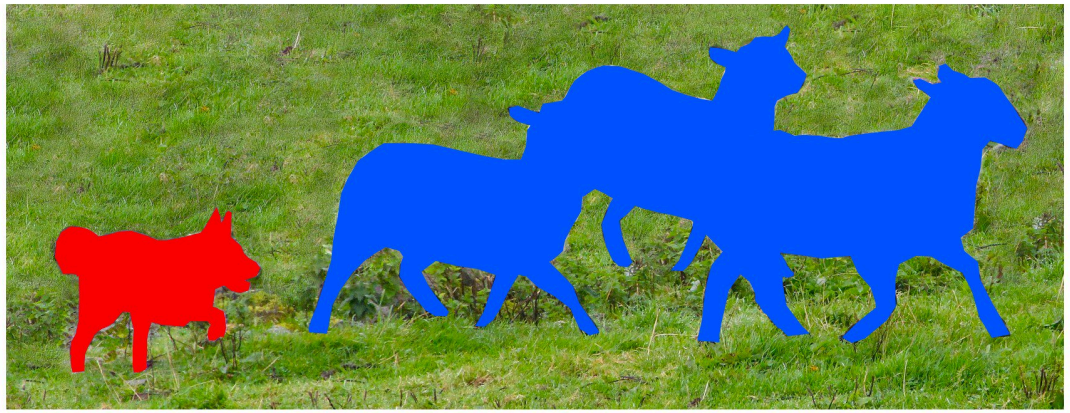
- Tools for fancy labelling
- Annotation conventions
- Augmenting data

Resources used:

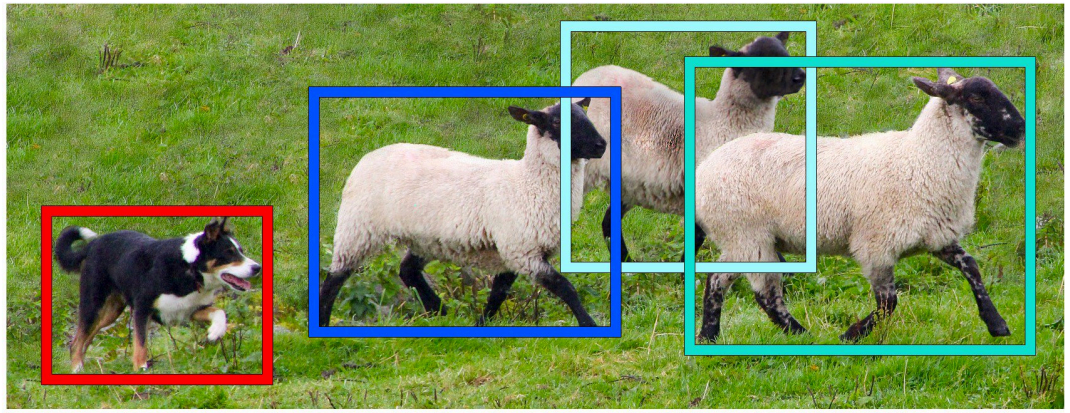
- Label Studio
- Labelformat docs
- Coco Dataset Description

Computer vision tasks

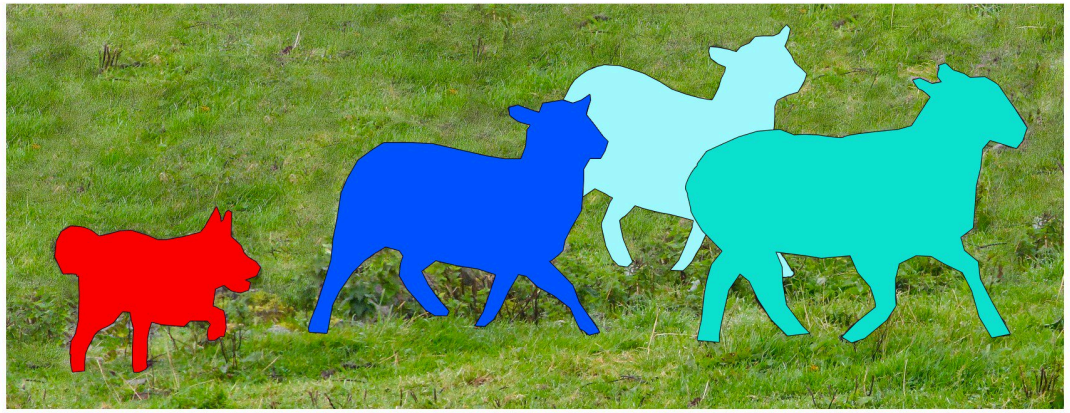
Image Recognition



Semantic Segmentation



Object Detection



Instance Segmentation

Image from https://manipulation.csail.mit.edu/segmentation.html

How are annotations stored?

Usually in plain text!

- Classification: subdirectories, csv files
- Bounding boxes: text files, e.g. COCO, VOC
- Segmentations: Text and/or PNG, e.g. COCO

> *Not a lot of formal process here - someone builds something for their purposes, others find it useful, variations abound.*

Audio annotations

- Classification: file naming/subdirectories, csvs
- Transcription: ASR (json files)
- Timestamps to label segments (csv, json)

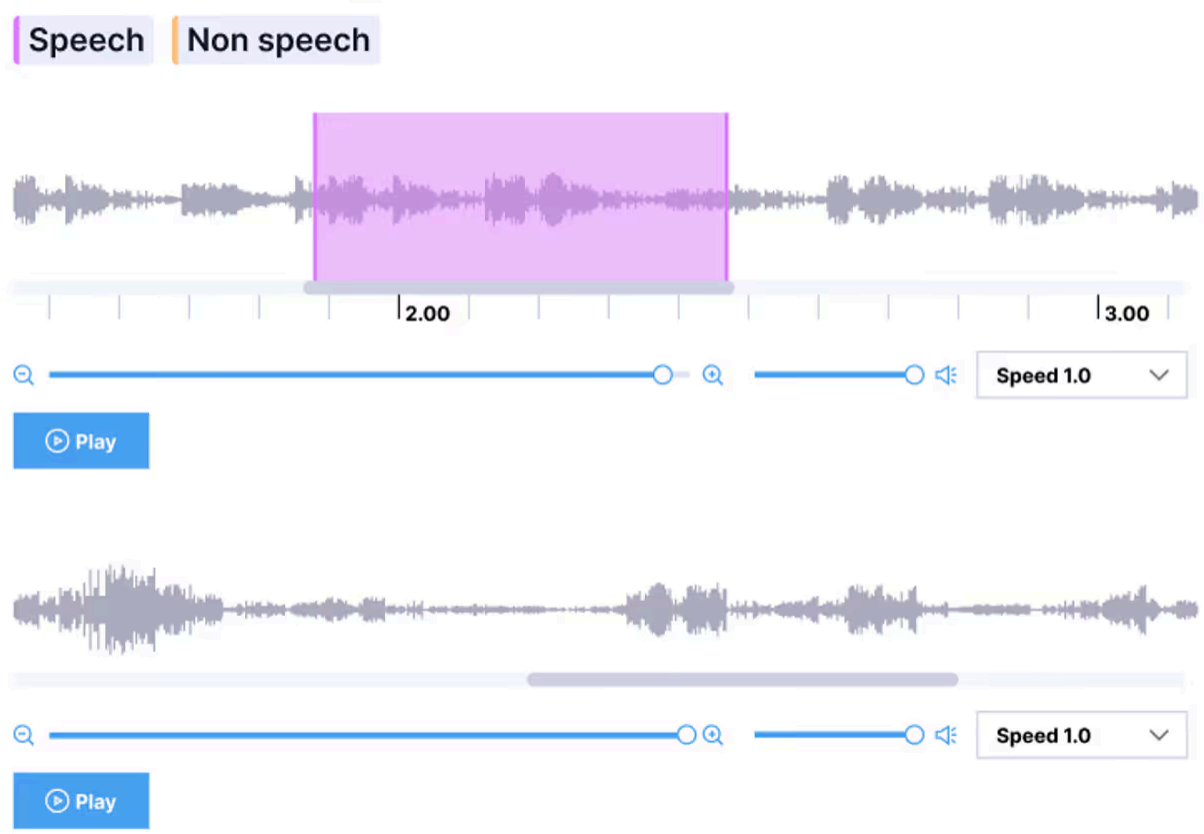


Image from https://labelstud.io

Text annotation and beyond

- Document classification or sentiment analysis
- Named entity recognition, e.g. CoNLL-2003
- Frame-by-frame video labelling

> *Ultimately there are a ton of data and label formats, you may well need to write a parser to interpret whatever you get*

Data Augmentation

The last topic!

When making predictions, consider...

- Your results are only as good as your data
- If it seems too good to be true, it probably is
- Always consider the various sources of data leakage
- If possible, get new samples for testing
- Be skeptical!

> *Example: a super basic MNIST image classifier*

Adding robustness

- It's not always possible to get more data, but we can **augment** what we have
- Images are of the same object even with:
  - Geometric transforms -- flips, rotations, scaling
  - Point operations -- brighter/darker/colour shift
  - Filtering -- Noise, JPEG compression, blurring
- Computer vision libraries like torchvision can turn 1 training image into 50

> *Make sure that the augmentation makes sense for your context*

Augmentation continued

- Typically only used for training data
- Initially, augmentation may make the performance metrics **worse**
- Not just for images! Audio signals can have noise added, pitch modulation, tempo changes, compression, resampling, etc
- Tabular data is trickier, but not impossible, particularly with generative AI

Main takeaways

- Data labels are in a mishmash of formats, and you will likely have to write some kind of parser at some point in your career
- Labelling data is a painful manual process, somewhat assisted by AI tools
- More data = more generalizability, sometimes it helps to invent some
- Above all, **be critical** of your results! Garbage in = garbage out.

The rest of the course

- Assignment 3: due Friday (ish, with the usual weekend leniency)
- Monday: Easter, no lab
- Next week: Assignment presentations + review
- Monday, April 13: come to the lab and stuff your H drive before the exam

> *Home stretch!*