Data Definition and Baseline

1. Which of these statements do you agree with regarding structured vs. unstructured data problems?

   - It is generally easier for humans to label data and to apply data augmentation on structured data than unstructured data.
   - It is generally easier for humans to label data on unstructured data, and easier to apply data augmentation on structured data.
   - It is generally easier for humans to label data on structured data, and easier to apply data augmentation on unstructured data.
   - It is generally easier for humans to label data and to apply data augmentation on unstructured data than structured data.

   Humans are better able to label unstructured data such as images and audio clips than complex, high-dimensional structured data. As well, it's not always possible to apply data augmentation to structured data.

2. Take speech recognition. Some labelers transcribe with "..." (as in, "Um… today’s weather") whereas others do so with commas ",". Human-level performance (HLP) is measured according to how well one transcriber agrees with another. You work with the team and get everyone to consistently use commas “,”. What effect will this have on HLP?

   - HLP will decrease.
   - HLP will stay the same.
   - HLP will increase.

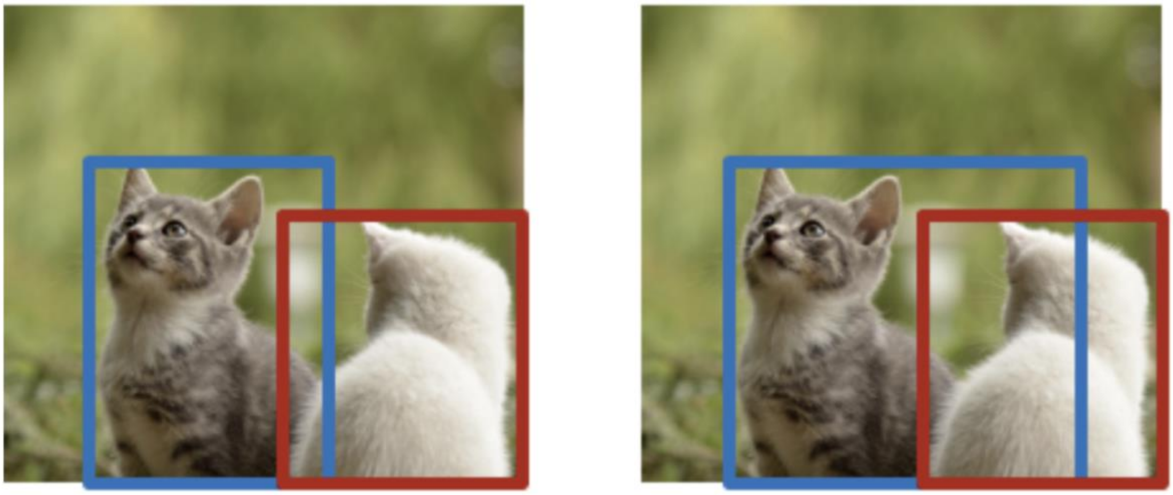
   Since the labels will be more consistent, the labelers will agree with each other more often, raising HLP.

3. Take a phone visual inspection problem. Suppose even a human inspector looking at an image cannot tell if there is a scratch. If however the same inspector were to look at the phone directly (rather than an image of the phone) then they can clearly tell if there is a scratch. Your goal is to build a system that gives accurate inspection decisions for the factory (not publish a paper). What would you do?

   - Carefully measure HLP on this problem (which will be low) to make sure the algorithm can match HLP.
   - Get a big dataset of many training examples, since this is a challenging problem that will require a big dataset to do well on.
   - Try to improve the consistency of the labels, y.
   - Try to improve their imaging (camera/lighting) system to improve the quality or clarity of the input images x.

If even a human looking at the image cannot identify the presence of a scratch, you'll need to improve the optical quality of your camera to improve your system's performance.

4. You are building a system to detect cats. You ask labelers to please “use bounding boxes to indicate the position of cats.” Different labelers label as follows:



What is the most likely cause of this?

- Ambiguous labeling instructions.
- Labelers have not had enough coffee.
- That this should have been posed as a segmentation rather than a detection task.
- Lazy labelers.

That's right! Your hardworking labelers may interpret the ambiguous instructions differently and label the images differently. Improve your instructions, and your labelled data will improve!

5. You are building a visual inspection system. HLP is measured according to how well one inspector agrees with another. Error analysis finds:

Type of defect	**Accuracy**	**HLP**	**% of data**
Scratch	95%	98%	50%
Discoloration	90%	90%	50%

You decide that it might be worth checking for label consistency on both scratch and discoloration defects. If you had to pick one to start with, which would you pick?

- It is more promising to check (and potentially improve) label consistency on scratch defects than discoloration defects, since HLP is higher on scratch defects and thus it’s more reasonable to expect high consistency.

- It is more promising to check (and potentially improve) label consistency on discoloration defects than scratch defects. Since HLP is lower on discoloration, it's possible that there might be ambiguous labelling instructions that is affecting HLP.

HLP is lower for discoloration defects, perhaps there is an opportunity to improve this metric by improving label consistency.

6. To implement the data iteration loop effectively, the key is to take all the time that's needed to construct the right dataset first, so that all development can be done on that dataset without needing to spend time to update the data.

- True
- False

Collecting and labelling data is an iterative process, get into the data iteration loop as quickly as possible.

7. You have a data pipeline for product recommendations that (i) cleans data by removing duplicate entries and spam, (ii) makes predictions. An engineering team improves the system used for step (i). If the trained model for step (ii) remains the same, what can we confidently conclude about the performance of the overall system?

- It will get worse because stage (ii) is now experiencing data/concept drift.
- It will definitely improve since the data is now more clean.
- It will get worse because changing an earlier stage in a data pipeline always results in worse performance of the later stages.
- It's not possible to say - it may perform better or worse.

It's really hard to tell, as it depends on how the data was changed, and how your model behaves.

8. What is the primary goal of building a PoC (proof of concept) system?

- To build a robust deployment system.
- To collect sufficient data to build a robusts system for deployment.
- To check feasibility and help decide if an application is workable and worth deploying.
- To select the most appropriate ML architecture for a task.

A proof of concept system is a simple way to determine whether its worth the time and effort to develop the product.

9. MLOps tools can store meta-data to keep track of data provenance and lineage. What do the terms data provenance and lineage mean?

- Data provenance refers data pipeline, and data lineage refers to the age of the data (i.e., how recently was it collected).
- Data provenance refers the input x, and data lineage refers to the output y.
- Data provenance refers to the sequence of processing steps applied to a dataset, and data lineage refers to where the data comes from.
- Data provenance refers to where the data comes from, and data lineage the sequence of processing steps applied to it.

10. You are working on phone visual inspection, where the task is to use an input image, x, to classify defects, y. You have stored meta-data for your entire ML system, such as which the factory each image came from. Which of the following are reasonable uses of meta-data?

- As an alternative to having to comment your code.
- To suggest tags or to generate insights during error analysis.
- Keeping track of data provenance and lineage.

  Meta-data will contain information about where the data come from and what processing steps were applied to it. This can be helpful when performing error analysis.
- As another input provided to human labelers (in addition to the image x) to boost HLP.