ML4Seismic Partners Meeting 2023

A New Seismic Fault Label Uncertainty Dataset: Insights from Expertise, Certainty, and Consistency

Jorge Quesada, Chen Zhou, Mohit Prabhushankar, and Ghassan AlRegib

Center for Energy & Geo Processing CeGP

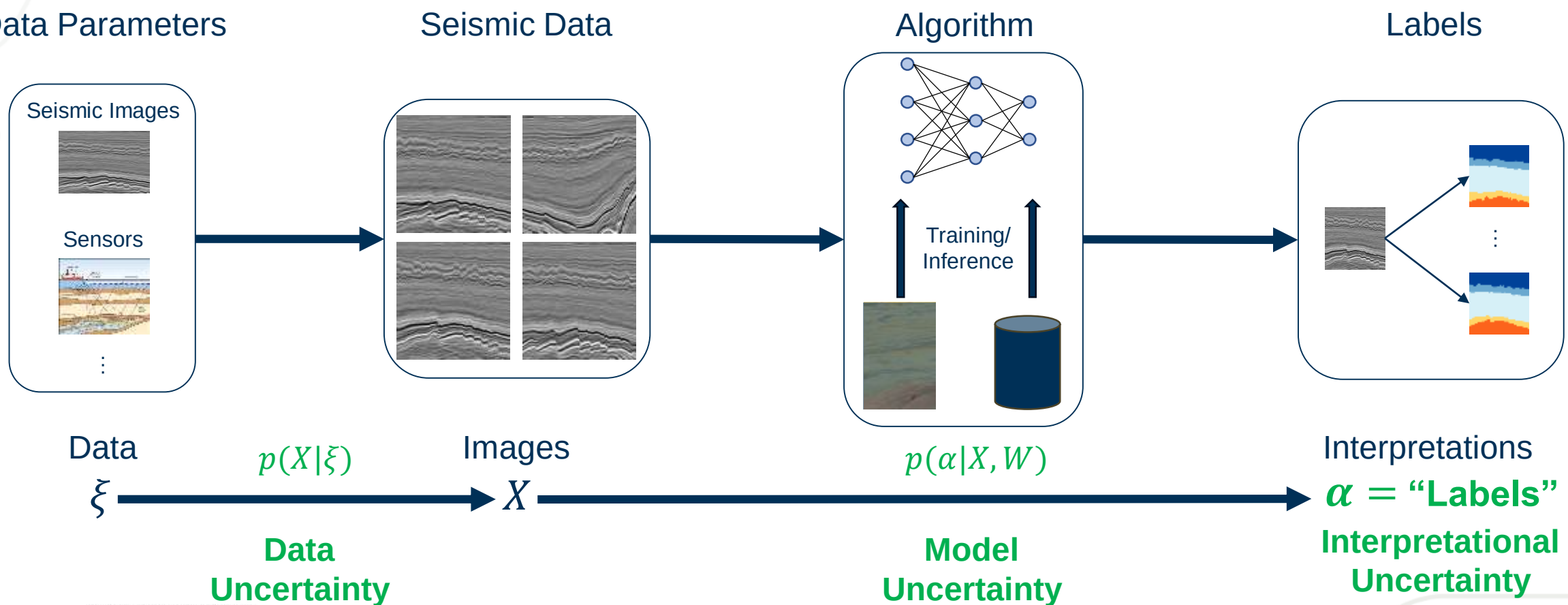
OLIVES @GeorgiaTech

Georgia Tech

Introduction

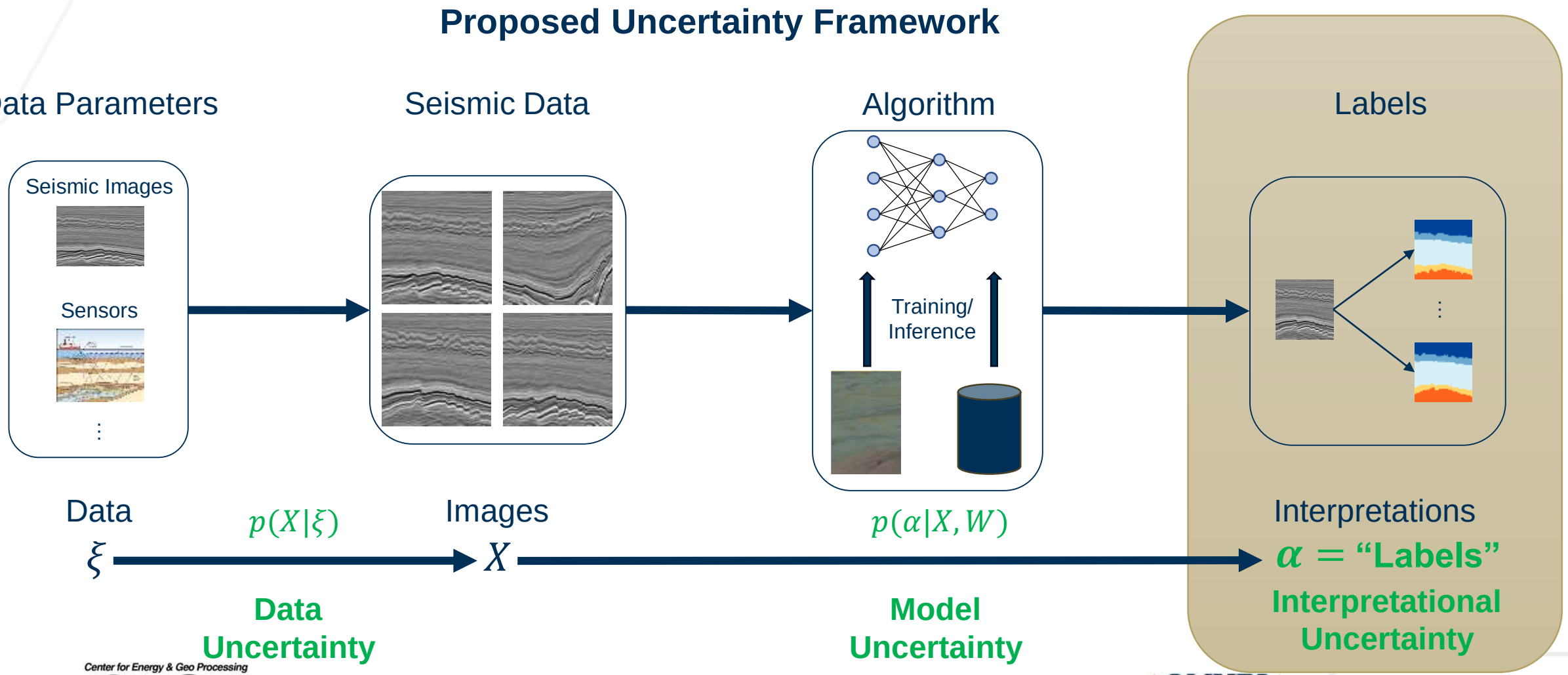
Where does annotator disagreement stand in the interpretation pipeline?

Proposed Uncertainty Framework



Introduction

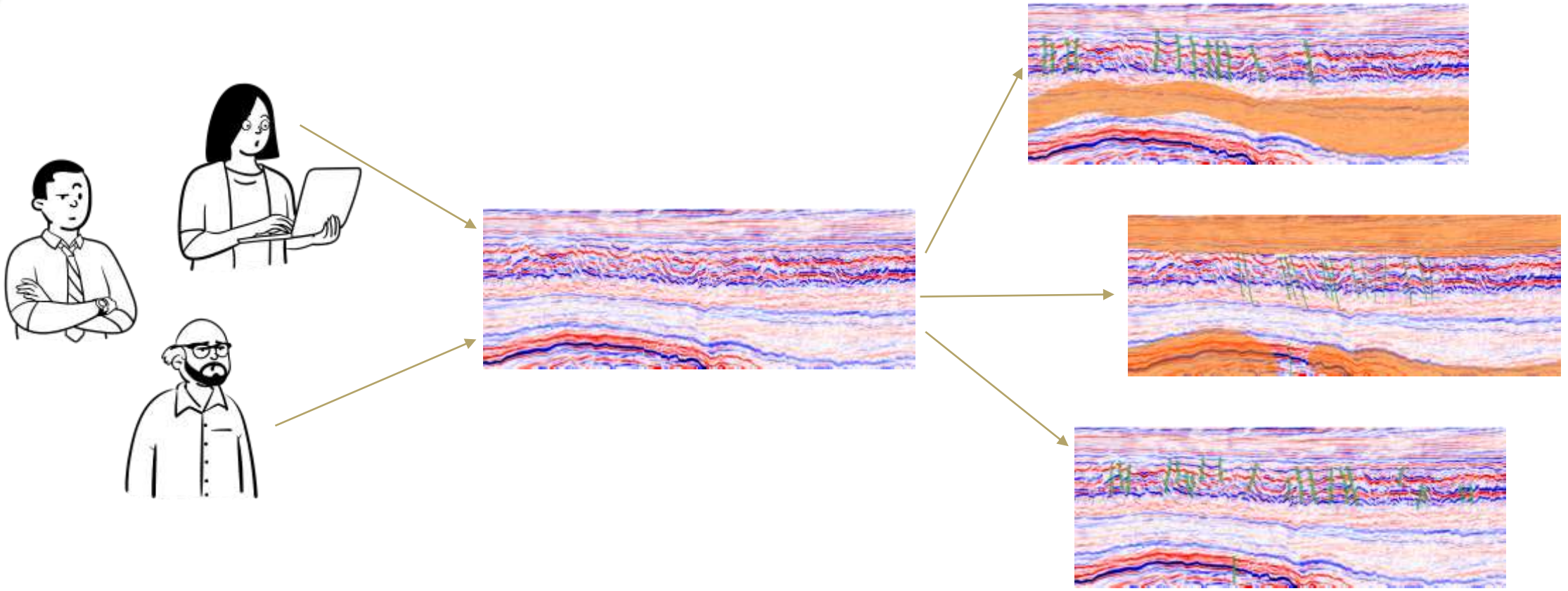
Where does annotator disagreement stand in the interpretation pipeline?



Introduction

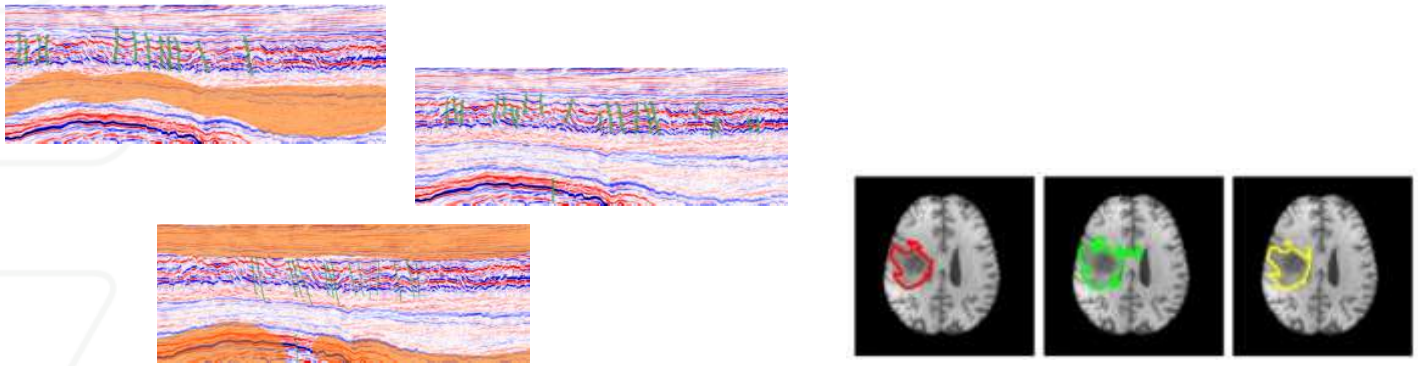
Multiple annotators can disagree during data labeling

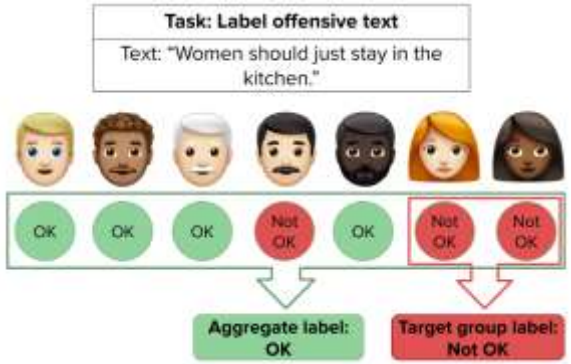
An image can yield different labels due to annotator confidence, expertise, etc



Introduction

When not addressed, annotator disagreement hinders downstream performance





Noisy labels

Degraded model/task performance

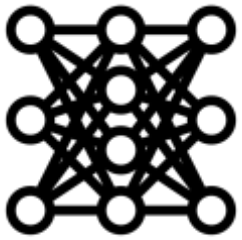
Fleisig, A., Abebe, R., Klein, D. (2023). When the Majority is Wrong. *arXiv preprint arXiv:2305.06626v*
Wang, S., Li, C., Wang, R., … & Zheng, H. (2021). Annotation-efficient deep learning for automatic medical image segmentation. *Nature Communications*

Introduction

How is annotator disagreement generally dealt with?

Voting

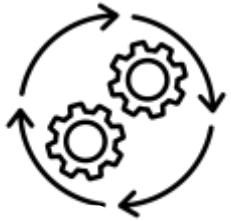
Averaging



Model integration

Cleaning/pruning

Schema modification



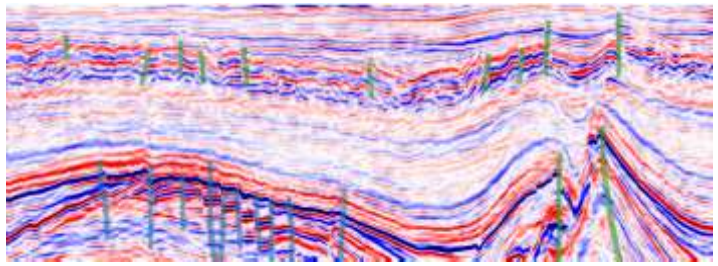
Mostafazadeh Davani, A., Diaz, M., Prabhakaran, V. (2022). Dealing with Disagreements. *Transactions of the Association for Computational Linguistics*

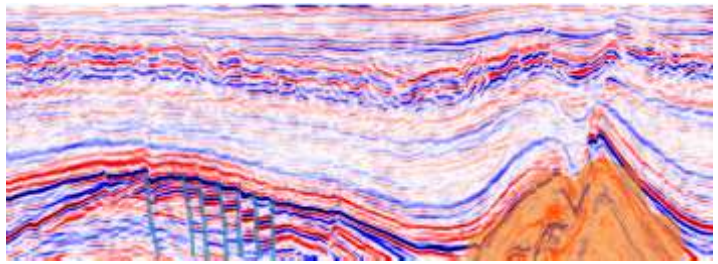
Introduction

Where does annotator disagreement stem from?

Expertise

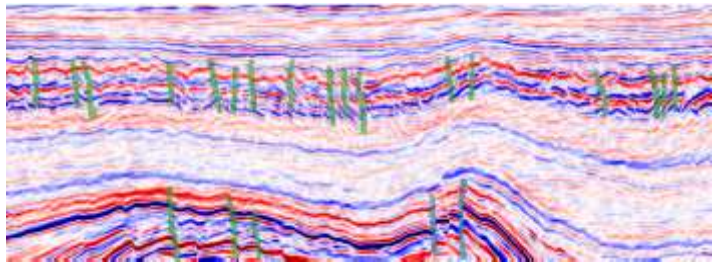
How much of a role does domain knowledge play?

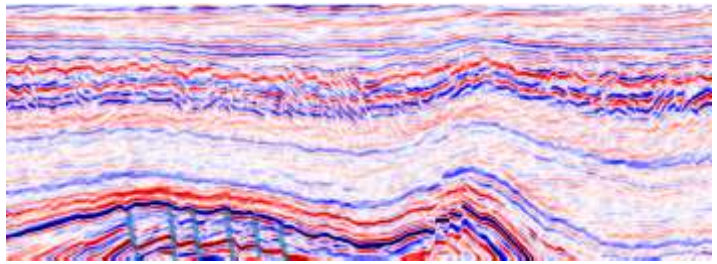




Confidence

How certain is each labeler of their own annotations?





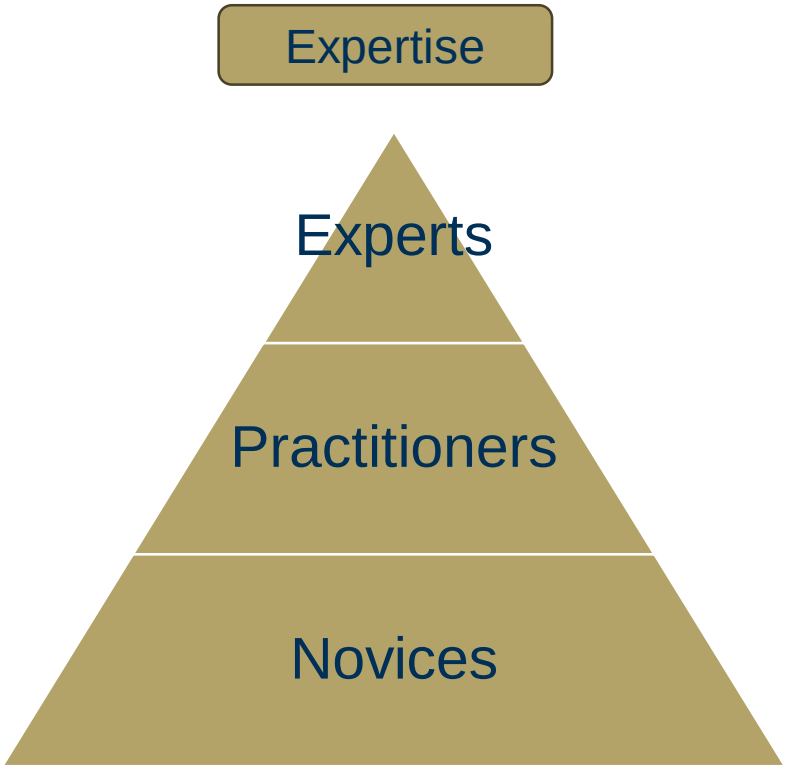
Center for Energy & Geo Processing CeGP

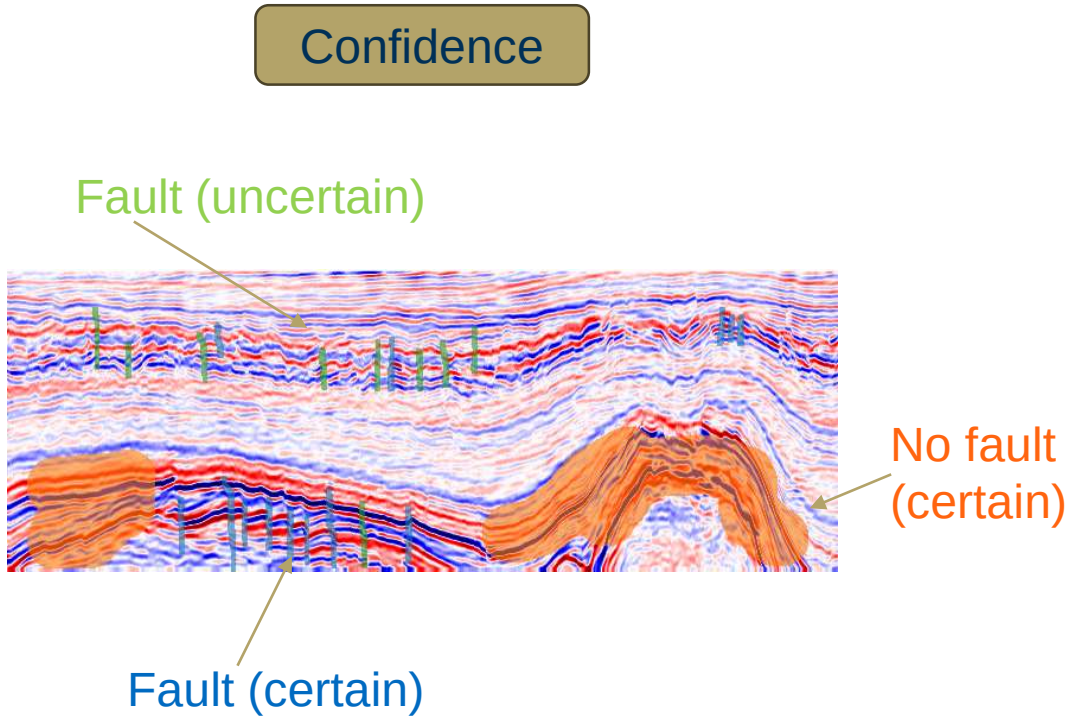
OLIVES @Georgia Tech

Georgia Tech

Introduction
Novelty of Our Work

A dataset comprising labels across multiple levels of ***expertise*** and ***confidence***





Center for Energy & Geo Processing CeGP

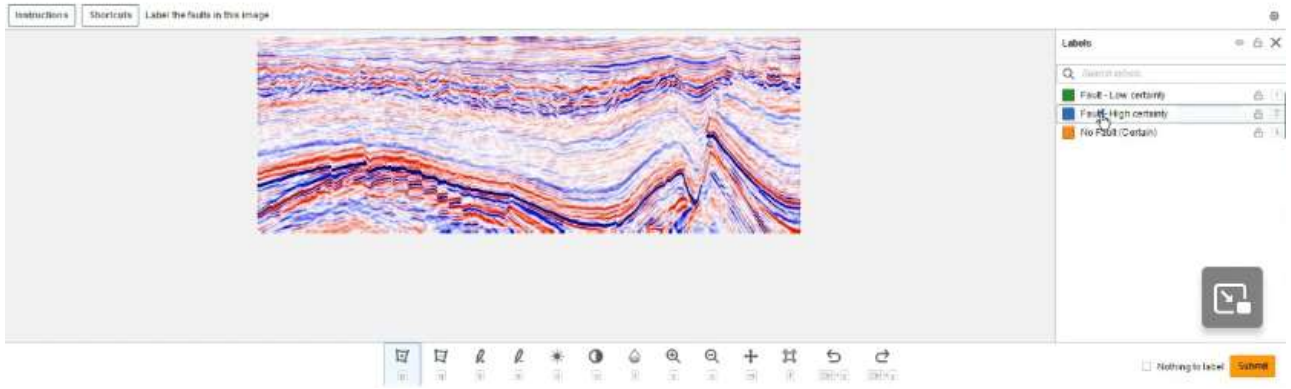
OLIVES

Georgia Tech

Summary of our dataset

We leverage Amazon Mechanical Turk for the labeling process

- 400 images, divided into 20 batches
- For each batch, 2 images are repeated 3 times for quality assessment
- 2 bonuses:
  - Number of images, promotes full dataset completion
  - Consistency, promotes thorough labeling



Current contribution:

- 1 expert, 8 practitioners

Experimental design - Platform

What is Amazon Mechanical Turk (MTurk)?

MTurk allows for a distributed outsourcing of virtual tasks

- Crowdsourcing marketplace
- Distributed workforce can perform tasks virtually
- Data validation and research, survey participation, content moderation, etc

Center for Energy & Geo Processing CeGP

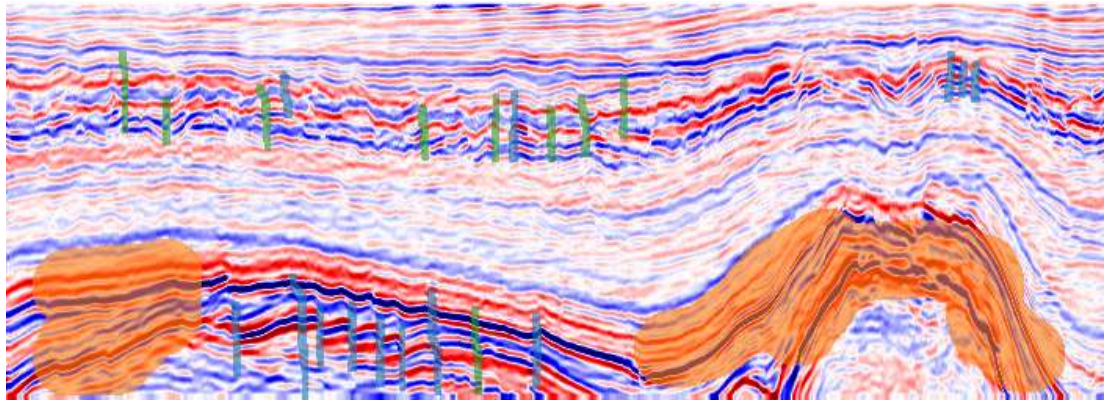
OLIVES @Georgia Tech

Georgia Tech

Experimental design - Labels

Three label categories enable flexible annotations

- We present 1 image per HIT, and consider 3 fault label categories for different certainty levels:
  - Fault (certain)
  - Fault (Uncertain)
  - No fault (Certain)

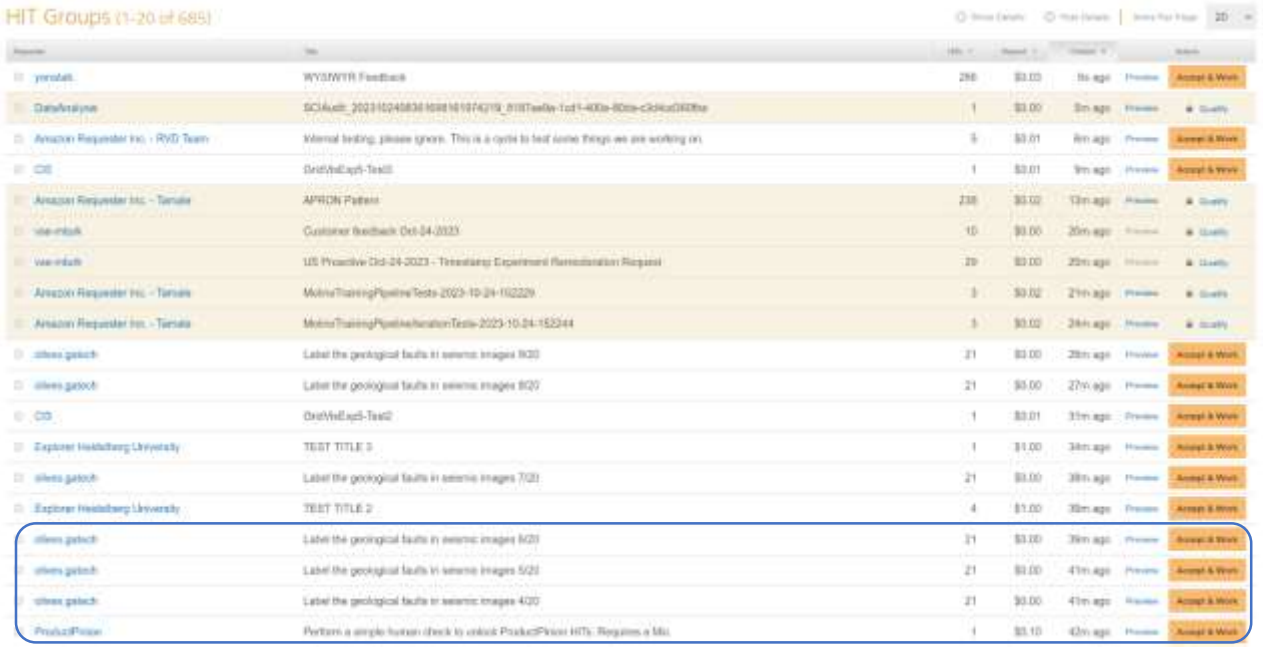


- Other schemes considered:
  - Multiple imgs per HIT
  - Sliders and text input boxes for certainty

Experimental design - Batches

Task is divided in 20 batches for easy navigation

- We divide the dataset into 20 batches, with 20 unique images each
- Each batch also contains 3 copies of 2 redundant images for quality assessment, totaling 6 QA images per batch
- Total batch size --> 24 imgs



Experimental design - Instructions

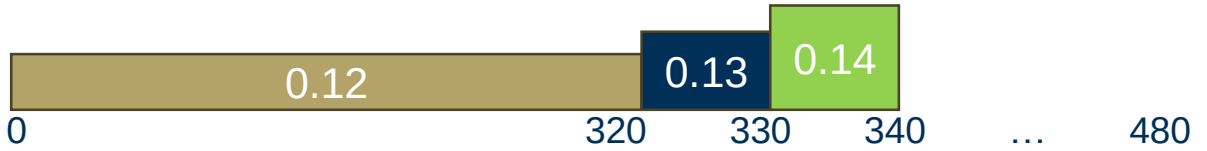
Concise guidelines aim to allow novices to annotate effectively

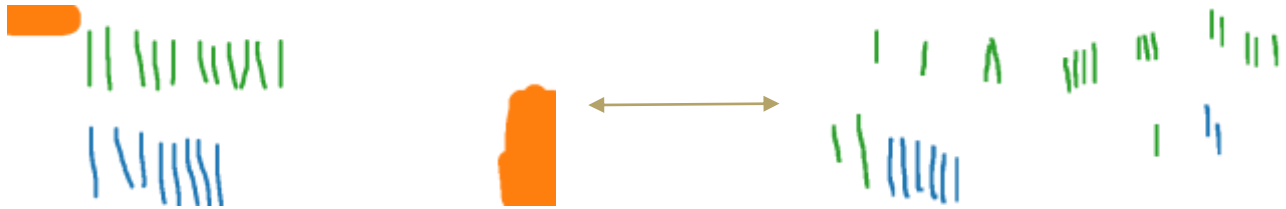
- We provide an instructional video in the task website and inside the layout, with the following details:
  - Fault definition
  - Sample image and label meaning
  - Platform usage
  - Payment scheme

Experimental design - Payment

Concise guidelines aim to allow novices to annotate effectively

- Base pay:
  - Reviewed existing mturk tasks and literature to arrive at initial pay rate
  - Adjusted pay rate for average time after internal task completion
- Number of image bonus:
  - Prorated bonus applied to final section of dataset to motivate completion
- Consistency bonus
  - Internal self-agreement metric for quality assessment



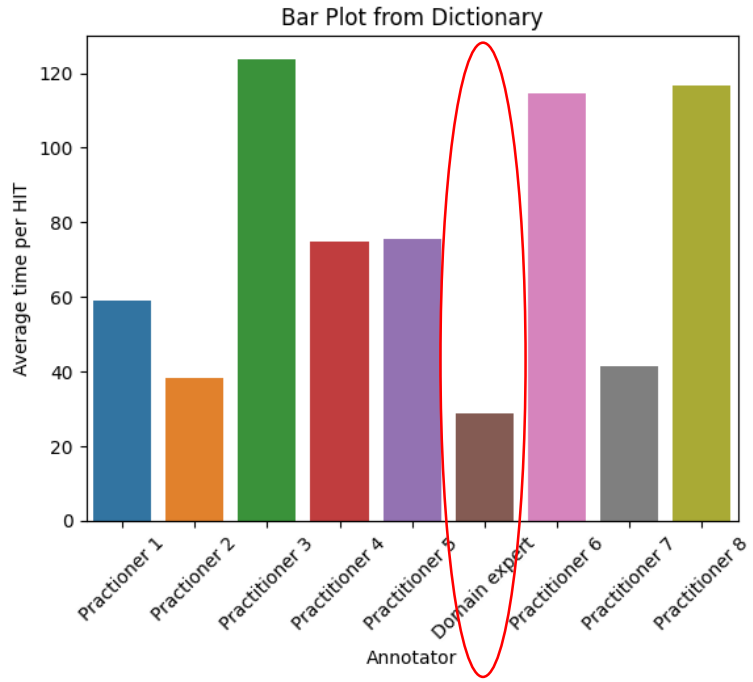


Insights from current annotators - Speed

Domain expert is generally faster than intermediate users

Expert is significantly faster than intermediate users

- Average labeling time: 75s
- Expert average time: 30s

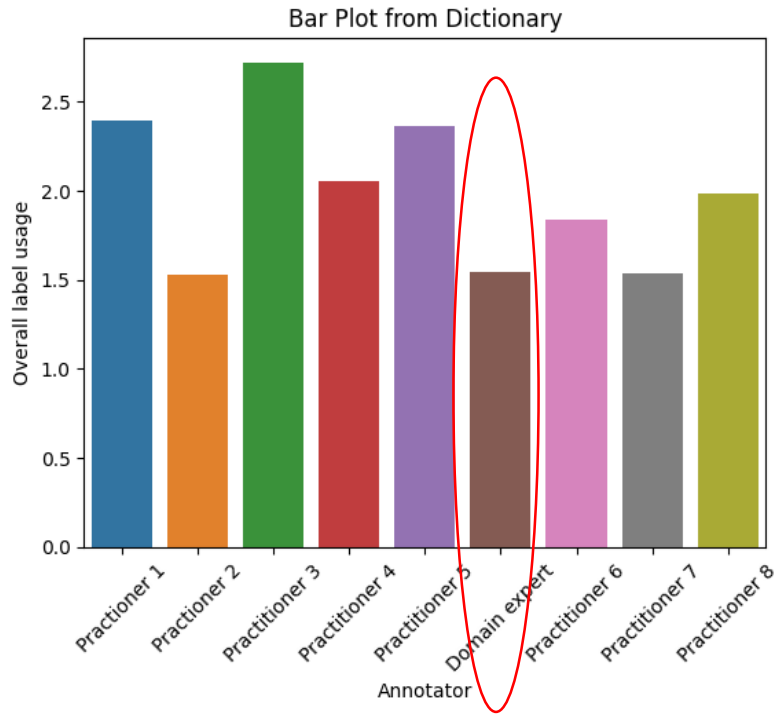


Insights from current annotators - Label usage

Annotators use less labels over time

Most people use only two labels or less (on average)

- Most people use only two labels on average
- Expert uses generally less labels than practitioners

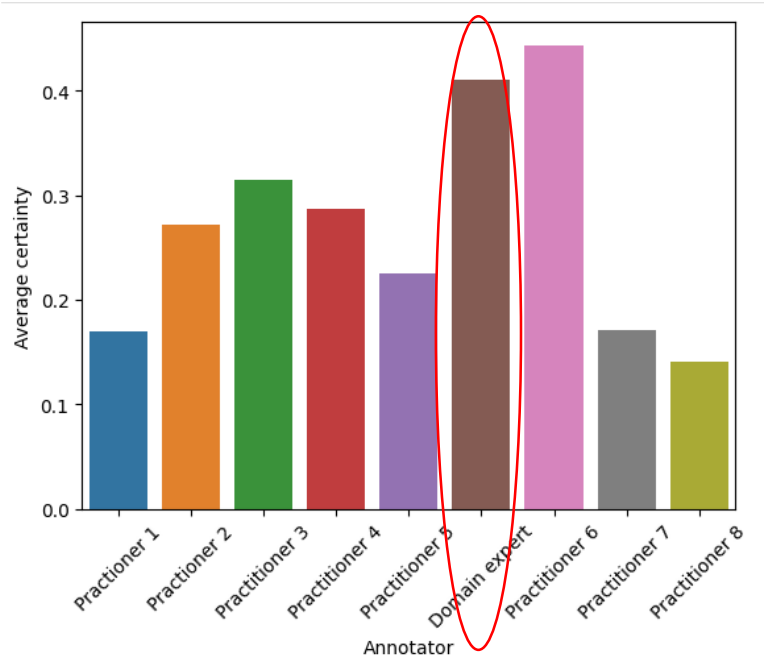


Insights from current annotators - Certainty

Confidence oscillates throughout task

Annotators with the most exposure make more confident labeling

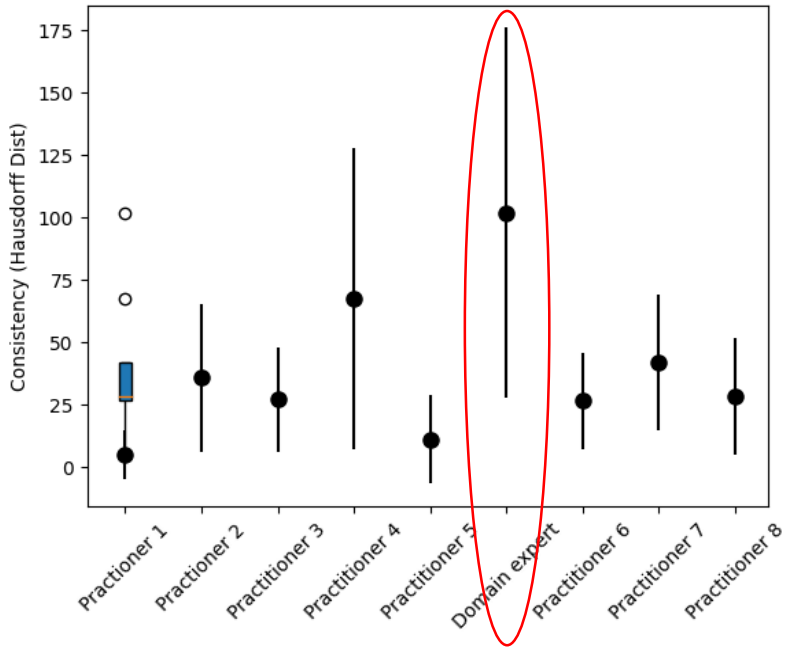
- Confidence taken only for the 2 fault labels
- Most experienced annotators label more confidently



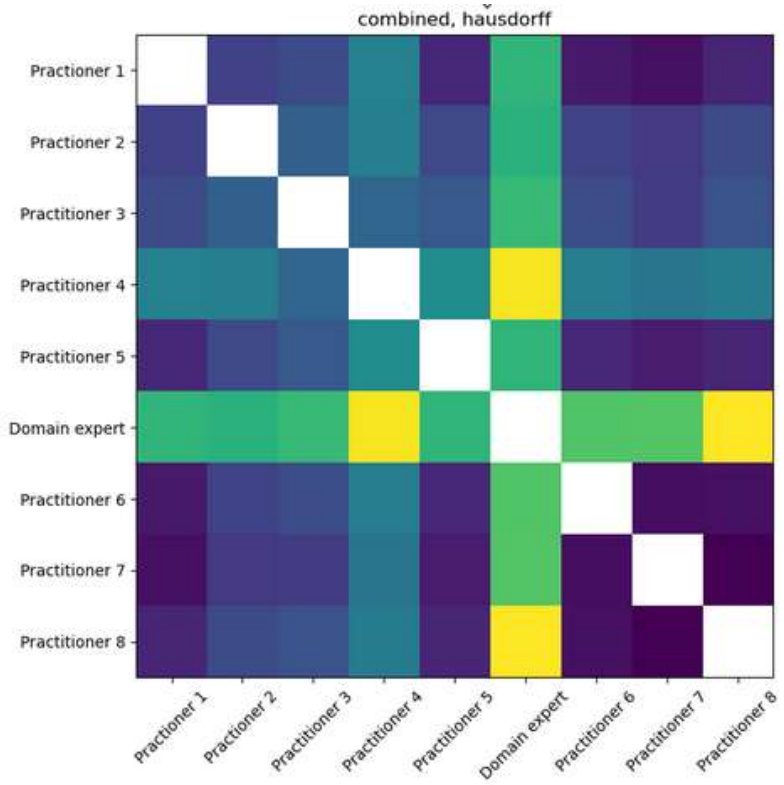
Agreement measurements

Intra and inter annotator agreement varies significantly

Expert is very distanced from other annotators as well as himself



Intra-annotator disagreement



Inter-annotator disagreement

Conclusion

Cross-expertise and multi-confidence dataset

- Uncertainty insights can help machine learning community build better models and methodologies to account for annotator disagreement
- A better understanding of the expertise gap can lead to more efficient fault labeling pipelines, reduced expert workload, and better fault detection models
- Our experimental design also constitutes a valuable resource for the seismic community to harness crowdsourcing platforms for efficient data labeling and annotation

For more OLIVES content, please visit:

GitHub

Publications