Hands-On-Session 1: Medical Report Data Analysis – practical

Timetable, Label Studio, Google Colab

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Hands-on Session Organization

The Rules

- Rule 0:
  - You are not alone. I’m here to help.
  - **Don’t panic!**
- Rule 1:
  - Everything (slides, data, links, explanations) can be found here:
  - https://github.com/iml-r/heidelberg-nmt

Timetable

- Theoretical presentation (the one before; 10 minutes)
- Practical presentation (the one now; 10 minutes)
- Annotation block (30-50 minutes)
  - Getting acquainted with Label Studio
    - Our browser-based annotation environment
  - Annotating
  - Discussion
- Programming/analysis block (20-25 minutes)
  - Coding in Colab – annotation evaluation
  - Discussion and specific cases deep dive

Total:

90 minutes tops

Label Studio

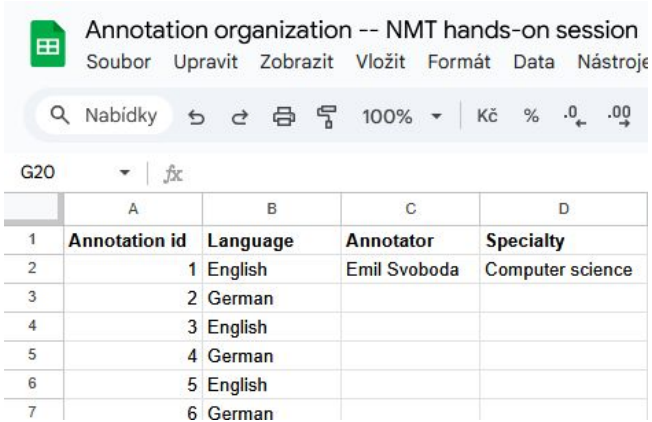
Label Studio

- An open-source annotation environment
- We run it on our hardware – third parties (Google, OpenAI)
- You can join it just by clicking this link (you'll have to register):
  - https://quest.ms.mff.cuni.cz/workshop-heidelberg/
- After you've registered and logged in, you should see something like this:

https://github.com/iml-r/heidelberg-nmt

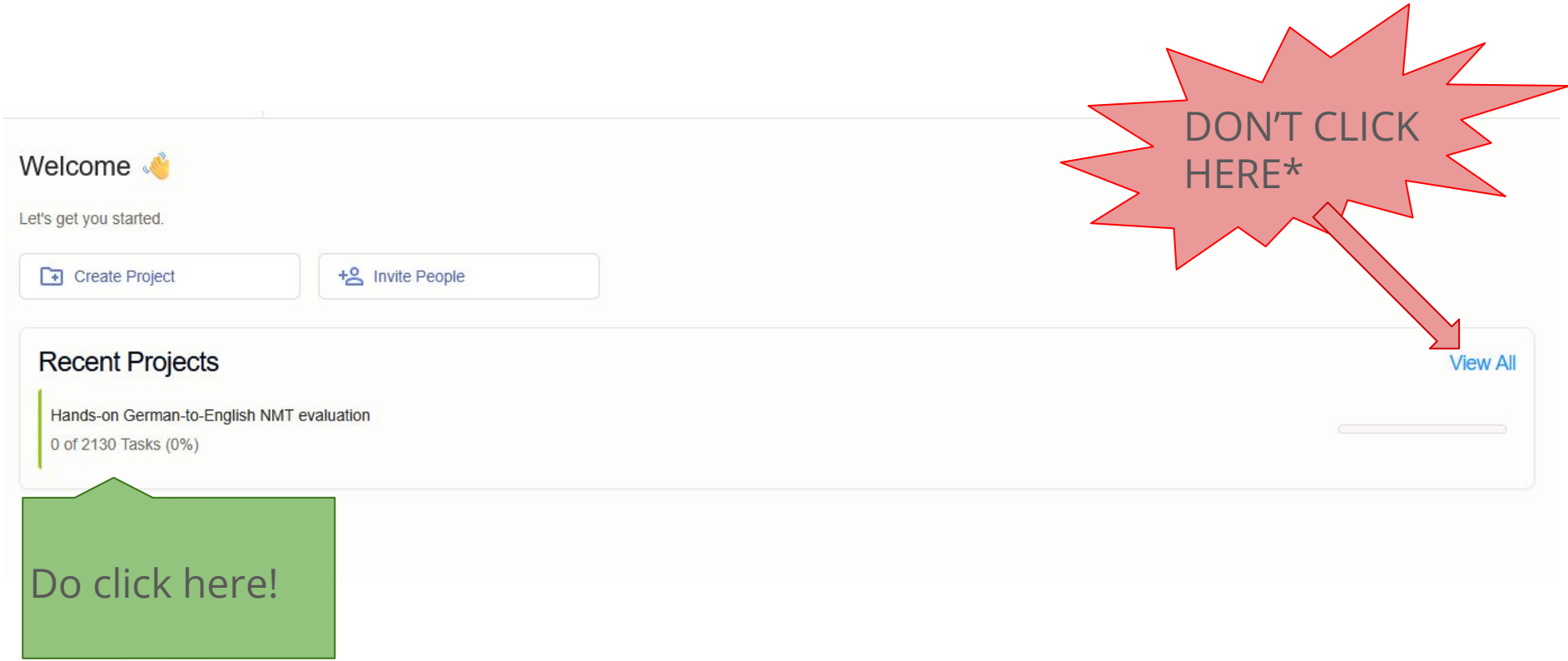
Organization sheet – BEFORE YOU START ANNOTATING

- **!!! WE NEED TO ENSURE THAT EACH TASK IS ANNOTATED BY ONE PERSON AND ONE PERSON EXACTLY !!!**
- So pick a task from the (publicly editable) Google spreadsheet:
  - https://docs.google.com/spreadsheets/d/17H3dxLnWugpMYFxhqGrsz3bBfefHUCPy916EZSaovrA/edit?usp=sharing
- Pick a task by id and language, fill out your name (you can make up one up, but use the same one for each task) and specialty (computer science vs. medicine)



	A	B	C	D
1	**Annotation id**	**Language**	**Annotator**	**Specialty**
2	1	English	Emil Svoboda	Computer science
3	2	German		
4	3	English		
5	4	German		
6	5	English		
7	6	German		

Label Studio

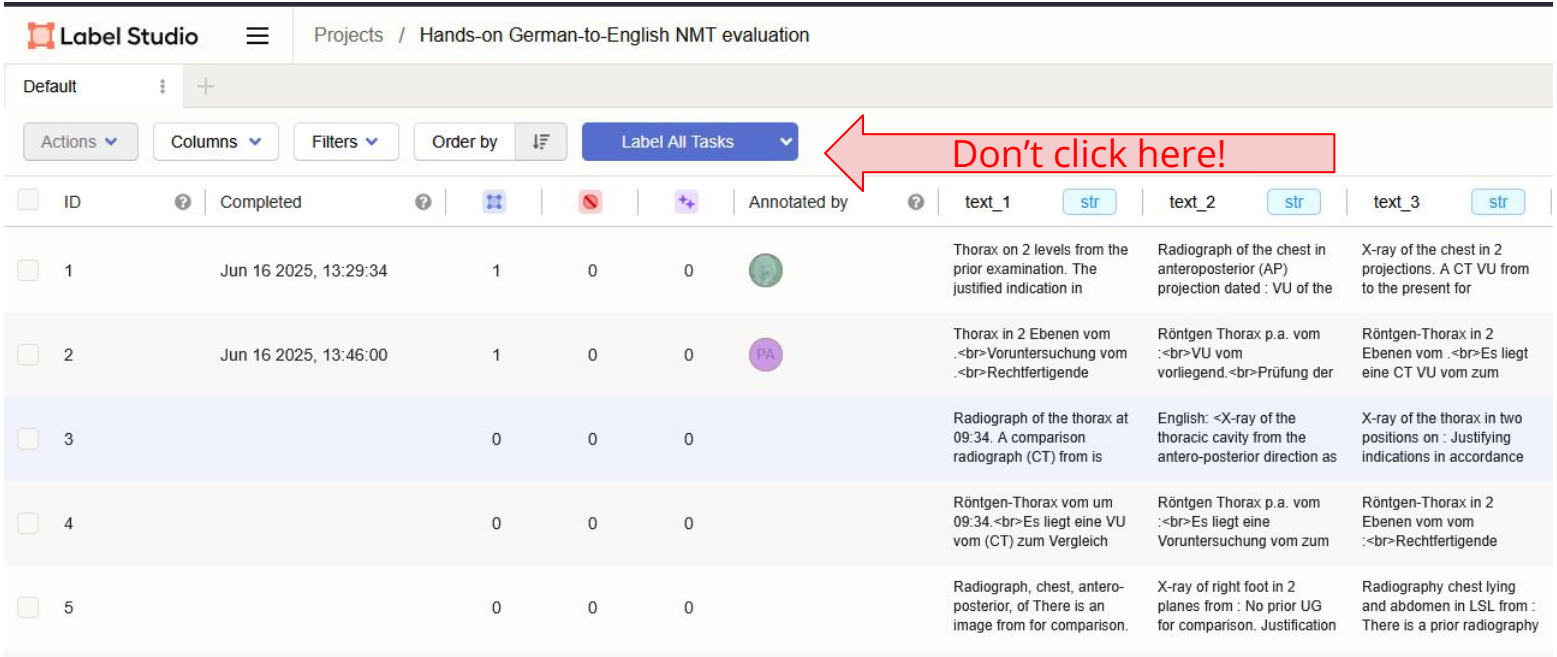


*if you do, just hit F5

https://github.com/iml-r/heidelberg-nmt

Label Studio

Always reserve a task first in the organization sheet – the “Annotated by” only appears after the task has been completed – and DON’T click “Label All Tasks”

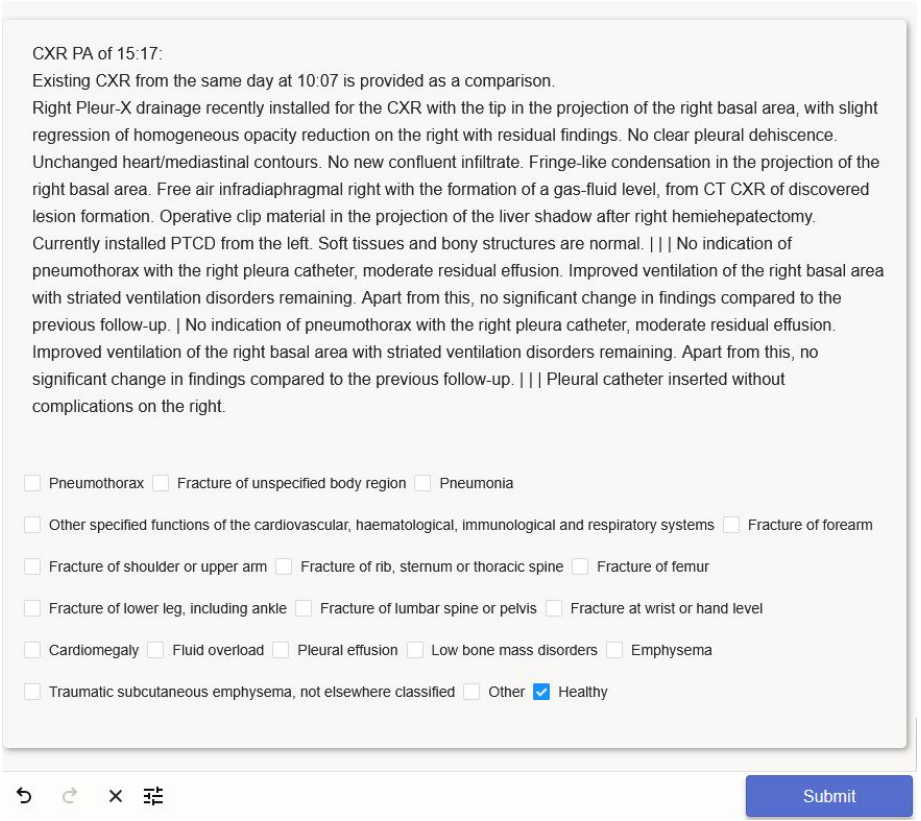


https://github.com/iml-r/heidelberg-nmt

Label Studio

- Then, you'll be presented with a sequence of 8 texts.
- Select 1 (or more) diagnoses for each text.
- Tick only the diagnoses that are *new findings*; e.g. don't annotate *Wrist fracture* if the report is from a follow-up exam.

Label Studio

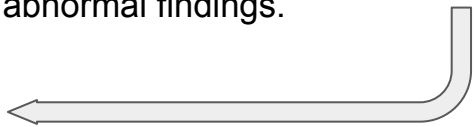


Always click submit after you've annotated your 8 reports

https://github.com/iml-r/heidelberg-nmt

Label Studio – label set

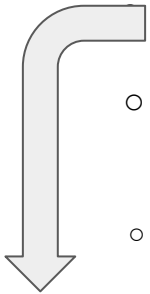
Condition	Description
Pneumothorax	Air leaks into the space around the lungs, possibly collapsing part or all of a lung.
Fracture of unspecified body region	A bone break is detected, but the location is not listed in the other labels (e.g. facial fracture)
Cardiomegaly	Enlargement of the heart, often linked to heart disease or chronic high blood pressure.
Fluid overload	Excess fluid in the body, often affecting the lungs, due to kidney or heart failure.
Pleural effusion	Fluid buildup in the space between the lungs and chest wall.
Low bone mass disorders	Conditions like osteoporosis that weaken bones, increasing fracture risk.
Emphysema	Chronic lung damage where air sacs are destroyed, making it hard to breathe.
Traumatic subcutaneous emphysema	Air escapes into tissue under the skin (as opposed to pneumothorax) after trauma, causing swelling and crackling sensation. (DISTINCT FROM EMPHYSEMA)
Other	Diseases and injuries not categorized above.
Healthy	No abnormal findings.



Google Colab

Google Colab

- Once we are done annotating, I will download the data and run in through a script that strips away the report texts (we can't let Google see them)
- I will then upload the anonymized data onto GitHub.
- The Colab will automatically download the data from there, no effort on your part required.
- To get the Colab notebook, go here:



https://colab.research.google.com/drive/1FGinFepX67gxicsL9O_V8S680CX1BhtD?usp=sharing

  - The notebook has reader-only sharing; you can edit but not save changes, so don't worry about ruining anything
  - If you want to copy it to experiment later, you can just click on 

Google Colab

- If you’re not familiar with Colab:
- It’s a service by Google which allows you to run Python code blocks alongside text and visualizations
    - It also lets you use some very fancy hardware, but we won’t be needing that today
- The Colab we will be using today has instructions written by me and designated empty spaces in some of the code blocks for you to solve the exercises

Final note

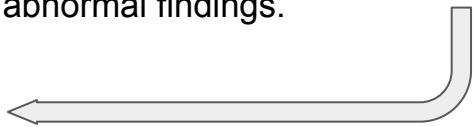
- If you run into any trouble with Label Studio, Colab, can’t find a link, or don’t understand something, just remember the rules:

Final note

- Rule 0:
  - **You are not alone. I'm here to help.**
  - **Don't panic!**
- Rule 1:
  - Everything (slides, data, links, explanations) can be found here:
    - https://github.com/iml-r/heidelberg-nmt

Label Studio – label set

Condition	Description
Pneumothorax	Air leaks into the space around the lungs, possibly collapsing part or all of a lung.
Fracture of unspecified body region	A bone break is detected, but the location is not listed in the other labels (e.g. facial fracture)
Cardiomegaly	Enlargement of the heart, often linked to heart disease or chronic high blood pressure.
Fluid overload	Excess fluid in the body, often affecting the lungs, due to kidney or heart failure.
Pleural effusion	Fluid buildup in the space between the lungs and chest wall.
Low bone mass disorders	Conditions like osteoporosis that weaken bones, increasing fracture risk.
Emphysema	Chronic lung damage where air sacs are destroyed, making it hard to breathe.
Traumatic subcutaneous emphysema	Air escapes into tissue under the skin (as opposed to pneumothorax) after trauma, causing swelling and crackling sensation. (DISTINCT FROM EMPHYSEMA)
Other	Diseases and injuries not categorized above.
Healthy	No abnormal findings.



https://github.com/iml-r/heidelberg-nmt