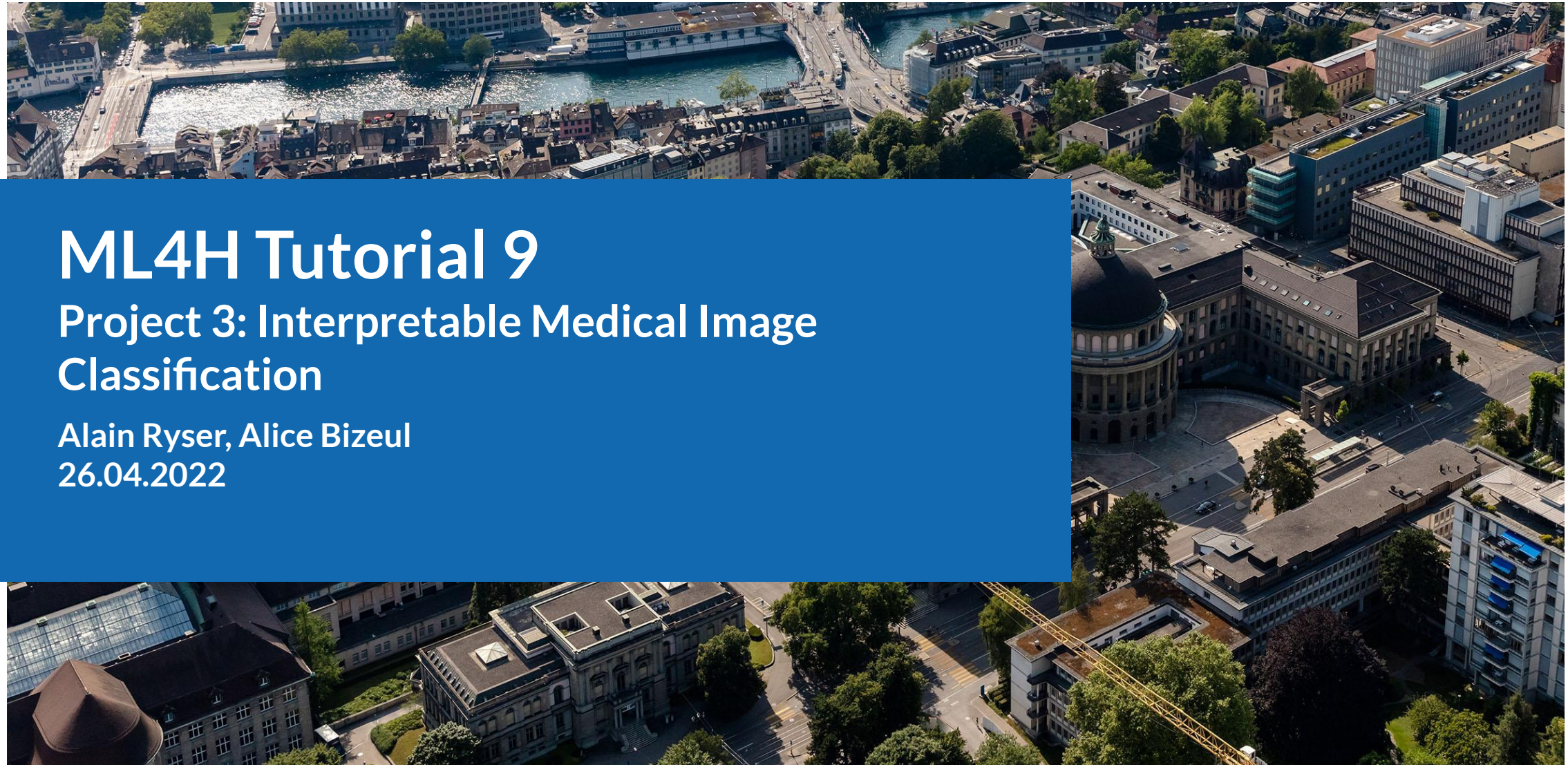
ETH zürich



ML4H Tutorial 9

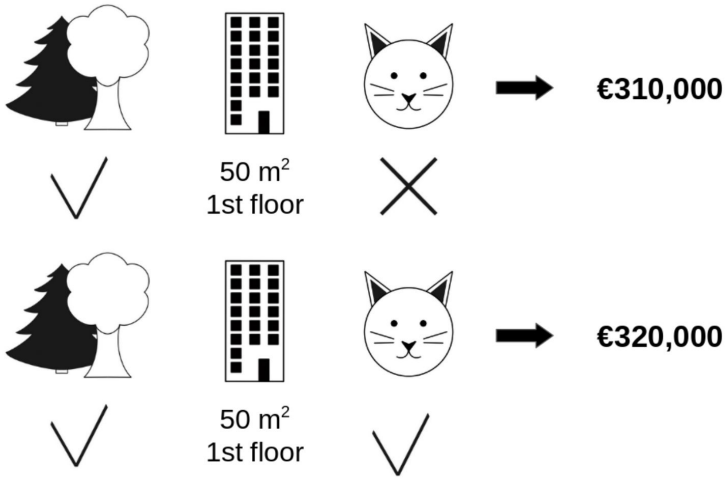
Project 3: Interpretable Medical Image Classification

Alain Ryser, Alice Bizeul

26.04.2022

Shapley Values https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

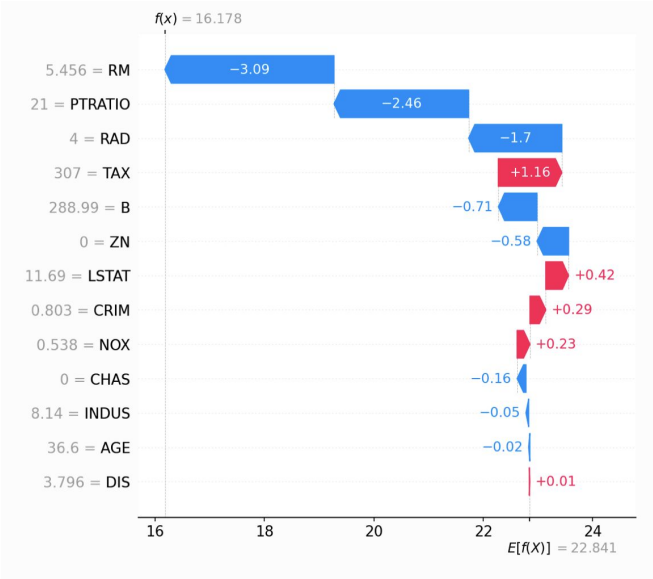
- Feature contribution metric originating from **cooperative game theory**
- Adapted and used for **post-hoc** ML interpretability purposes
- Shapley Values are the **average marginal contribution** of each feature to the difference between the given prediction and the average prediction
- Shapley Values provide information on **relative importance** of each input feature for a given prediction v



$$\Phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(N - |S| - 1)!}{N!} \left(v(S \cup \{i\}) - v(S) \right)$$

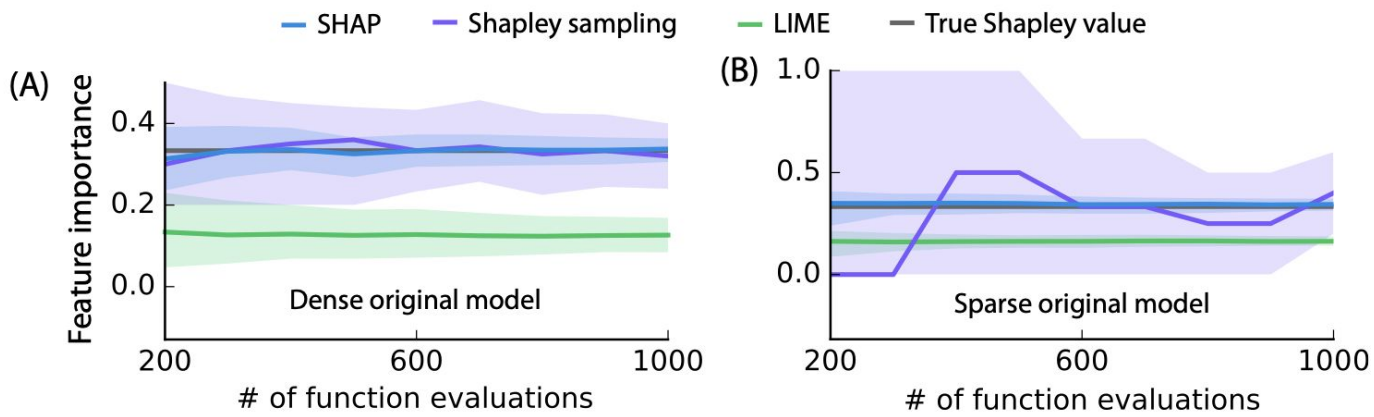
Shapley Values https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

- Axiomatic metric:
  - *Efficiency*: sum of values across input features add up to the difference between given and average prediction
- Average prediction is computed on a reference group/background samples, depending on the question we are trying to answer
- Pros:
  - Model agnostic approach, only requires input-output pairs
  - Information about local predictions as well as the global model
- Cons:
  - Amount of input perturbation to consider scales expo. with input size
  - Requires estimates computation in most real-world datasets.



SHapley Additive exPlanations (SHAP)

- Shapley sampling values using **sampling methods** to approximate true Shapley values
- **SHapley Additive exPlanations** (SHAP) combines sampling and other interpretable approaches (i.e. Local Surrogate Models - LIME, DeepLift):
  - Increased **computational efficiency**
  - **Improved approximation** of true Shapley Values
- Various approximation methods :
  - Kernel SHAP: model-agnostic
  - Linear SHAP: independance assumption
  - Deep SHAP: deep networks



SHAP Python library https://shap.readthedocs.io/en/

- provides **automatic estimates** of Shapley values for wide range of ML/DL models types (1)
  - Model Agnostic
  - Linear models
  - Tree based models
  - Neural networks
    - single & multi output
    - including transformers
    - including image classification, image captioning, text generation
- supports Tensorflow, Pytorch, Sklearn models

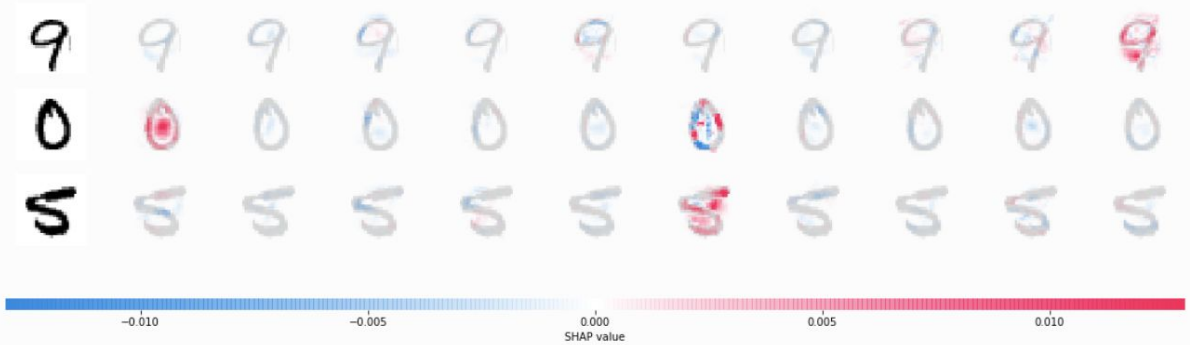
```
# since shuffle=True, this is a random sample of test data
batch = next(iter(test_loader))
images, _ = batch

background = images[:100]
test_images = images[100:103]

e = shap.DeepExplainer(model, background)
shap_values = e.shap_values(test_images)
```

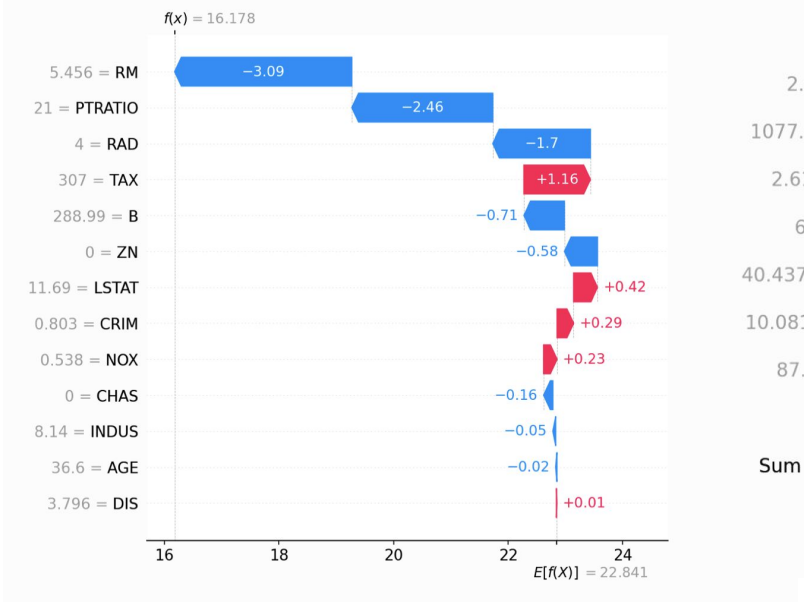
```
shap_numpy = [np.swapaxes(np.swapaxes(s, 1, -1), 1, 2) for s in shap_values]
test_numpy = np.swapaxes(np.swapaxes(test_images.numpy(), 1, -1), 1, 2)
```

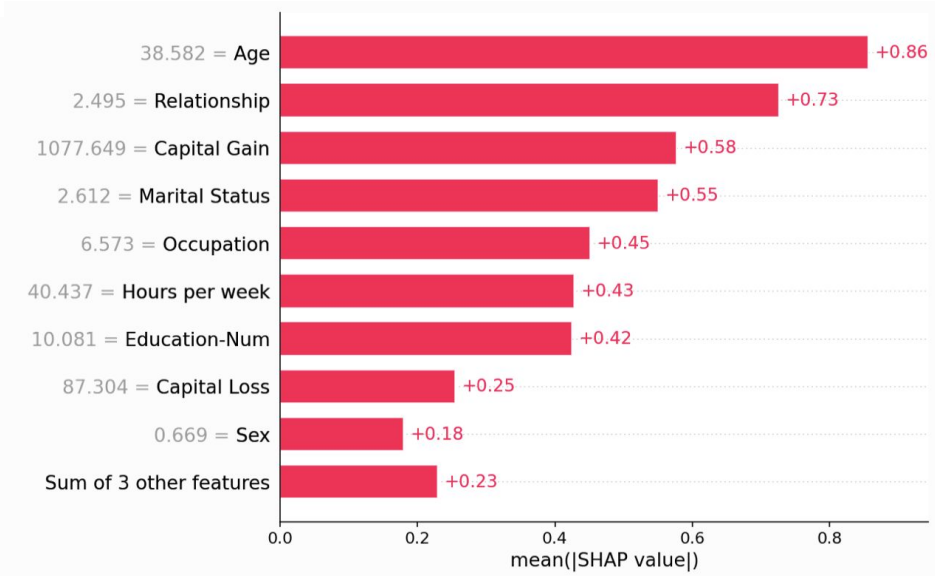
```
# plot the feature attributions
shap.image_plot(shap_numpy, -test_numpy)
```

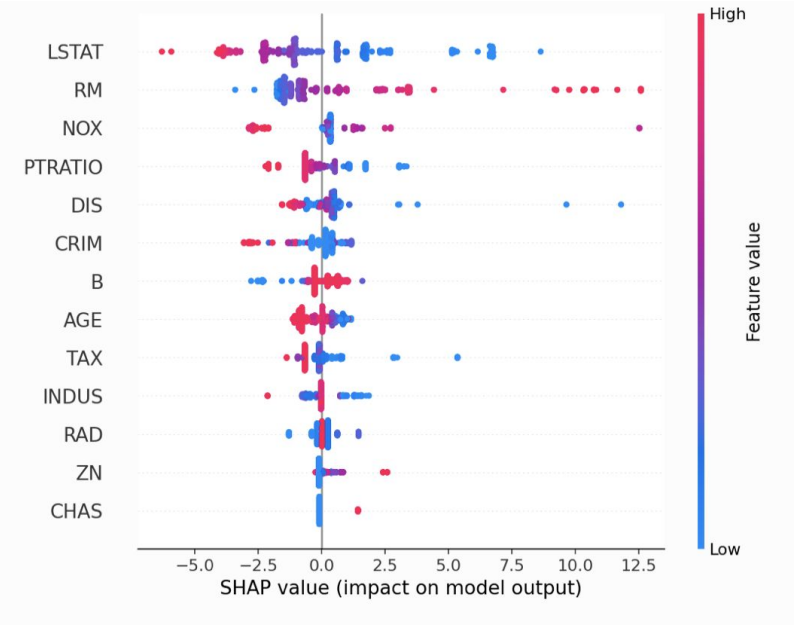


SHAP Python library https://shap.readthedocs.io/en/

- provides **automatic estimates** of Shapley values for wide range of ML/DL models types (1)
- provides **visualisations** of Shapley values for interpretability purposes (2)





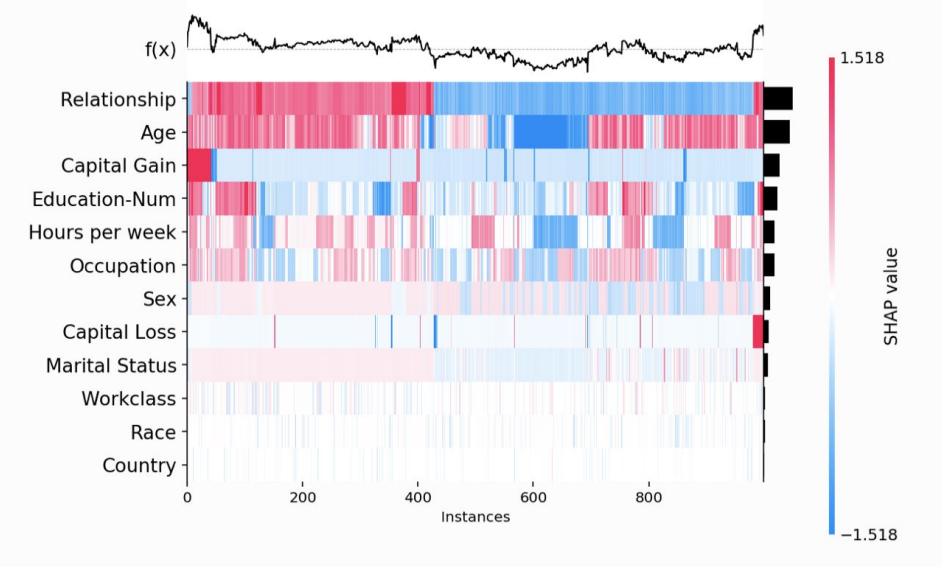


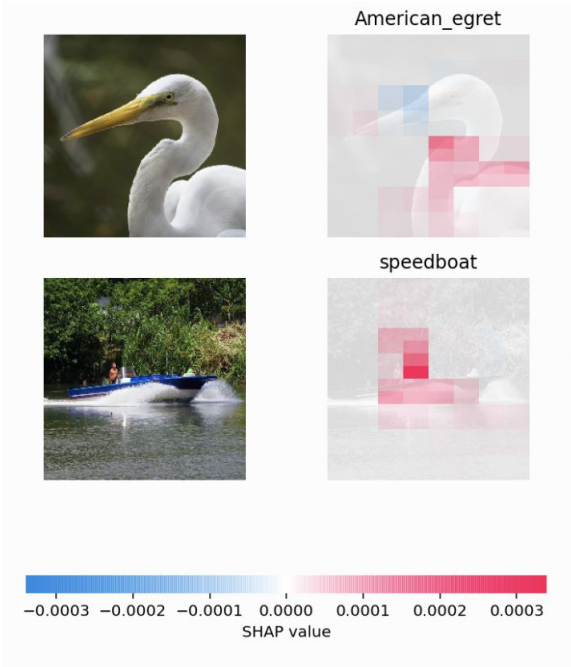
(1) https://shap.readthedocs.io/en/latest/api.html#explainers

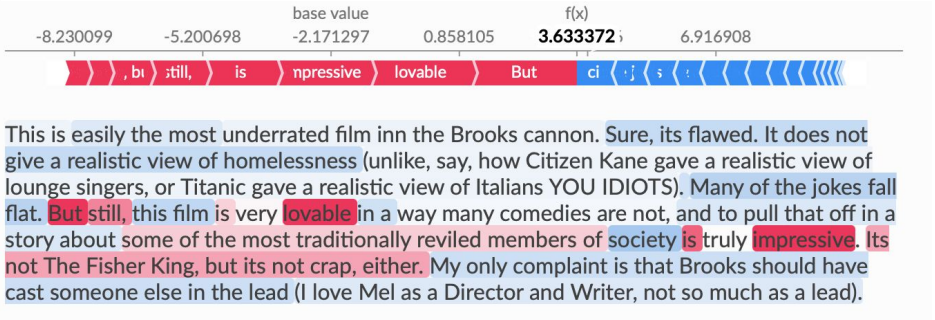
(2) https://shap.readthedocs.io/en/latest/api.html#plots

SHAP Python library https://shap.readthedocs.io/en/

- provides **automatic estimates** of Shapley values for wide range of ML/DL models types (1)
- provides **visualisations** of Shapley values for interpretability purposes (2)







(1) https://shap.readthedocs.io/en/latest/api.html#explainers

(2) https://shap.readthedocs.io/en/latest/api.html#plots

Project Tasks

1. Task 1: Implement Random Forest baseline using provided dataset of PyRadiomics features
2. Task 2**:** Implement CNN baseline and provide SHAP values interpretation of results using the provided MRI dataset
3. Task 3: Implement at least two additional interpretable classification methods on either dataset and one additional post-hoc interpretation method
4. [OPTIONAL] Improve model accuracy using approaches of your choice (e.g transfer learning, data augmentations, …)
5. Compare results and motivate the choice of your final classifier based on performance-interpretability trade-off

- Code Template and Dataset can be accessed **here**.

Important: Elaborate on the interpretability aspect of the implemented classifier/post-hoc method in the report for each subtask!

Project Deliverables

- Solve all tasks.
- **Report** of max. 4 pages, 11pt (+ 1 page for references + 1 page of appendix if needed).
- **Well-commented code/jupyter notebooks** with conda environment and README.
- Do not hardcode any results! We will run your code.
- Ensure sequential execution and **reproducibility**.
- **Do not copy solutions from previous projects!** We are aware of all existing solutions on github and kaggle. We run code similarity checks and check for plagiarism in the reports from previous years solutions. Any plagiarism will result in a 0 grade for all projects.
- **Deadline**: 17.05.2022

Project Grade

- To grade the project we will focus (on equal parts) on:
  - the content, organisation, clarity, quality and writing of the **final report**.
  - the quality of the **implementation** (reproducibility and clarity).
  - the **performance*** of the methods used to solve the tasks, and the **justification** of the choices.
- The **prerequisites** to get the maximum grade are:
  - write a clear and **good report.**
  - submit a **clean code** with **easy instructions** on how to reproduce each result of the report.
  - solve every task with **well-justified** methods and choices.
  - bonus: implement **creative** models for one/the optional task

*We will consider resource constraints. Aside from correct baseline implementation, the aim is not to get the best performance but to explore and discuss relevant methods.

Reading List

Shapley Values

- https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
- https://apps.dtic.mil/sti/pdfs/AD0604084.pdf

GradCAM

- https://arxiv.org/pdf/1610.02391.pdf

Integrated Gradients

- http://proceedings.mlr.press/v70/sundararajan17a/sundararajan17a.pdf

Local Surrogate Models (LIME)

- https://arxiv.org/pdf/1602.04938.pdf?ref=morioh.com

DeepLIFT

- https://arxiv.org/pdf/1704.02685.pdf

Questions?

Also on Moodle (preferred: your classmates probably have similar questions!)
or by email at alain.ryser@inf.ethz.ch , alice.bizeul@inf.ethz.ch .