

ML CASES

OXFORD MACHINE LEARNING SUMMER SCHOOL

MAY-JUNE 2023





WITH YOU TODAY

Vincent Moens

ML Research Scientist at PyTorch (Meta)

Founder and maintainer of TorchRL (<https://github.com/pytorch/rl>), the RL and decision-making library for PyTorch.



WHAT YOU WILL TAKE OUT OF ML CASE



ENGAGING

Hands-on experience on a timely and challenging problem with relevant applications over multiple weeks

INVEST +15H



FUN

You will get to play around Machine Learning models and apply them to real-world applications in either healthcare or finance.

BE CURIOUS



COLLABORATIVE

You will work closely in groups to learn from each other's experiences.

JOIN A TEAM, DEFINE A STRATEGY, RE-GROUP FREQUENTLY



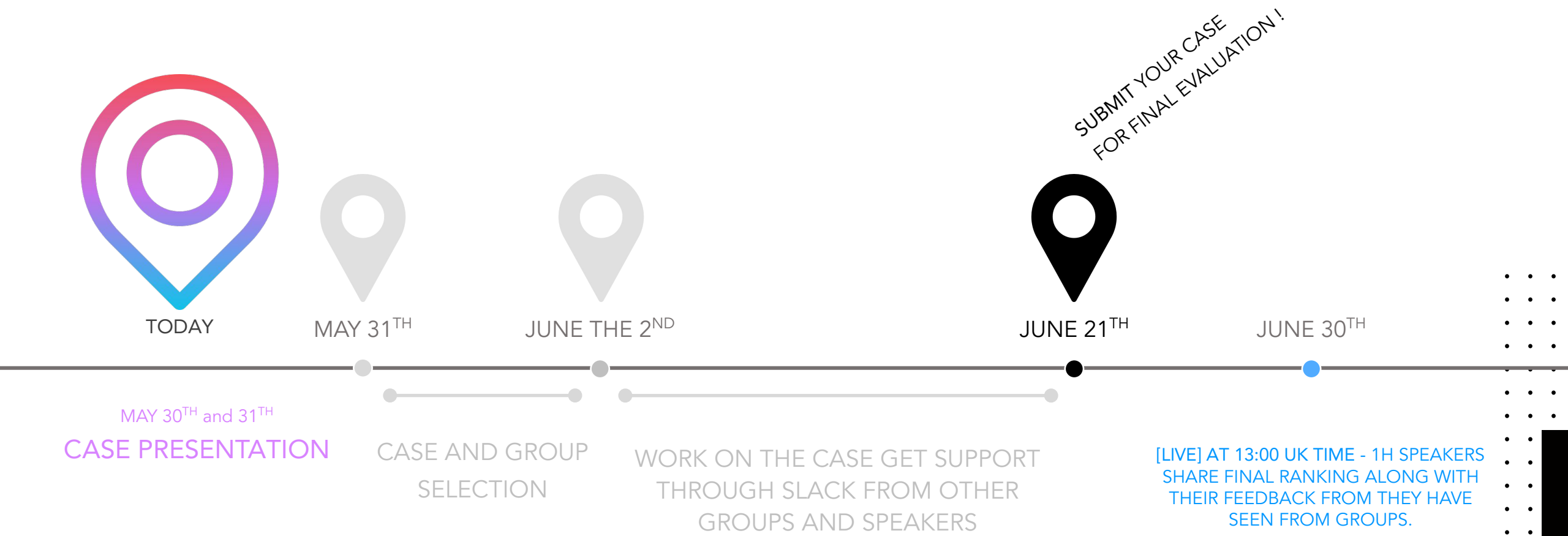
REWARDING

A personal project in your portfolio and if relevant uploaded to github. The best projects will be showcased through the OxML platform

BUILD A DELIVERABLE FOR YOUR PORTFOLIO



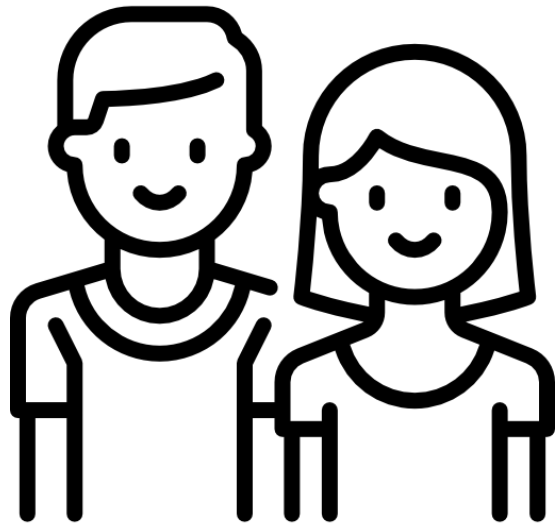
SCHEDULE





The challenge

You're hired by a small scale clinical biology clinic



YOUR CLIENTS...

Looking for a quick, cheap solution to automatically triage histopathological samples as high or low priority.

YOUR COMPANY...

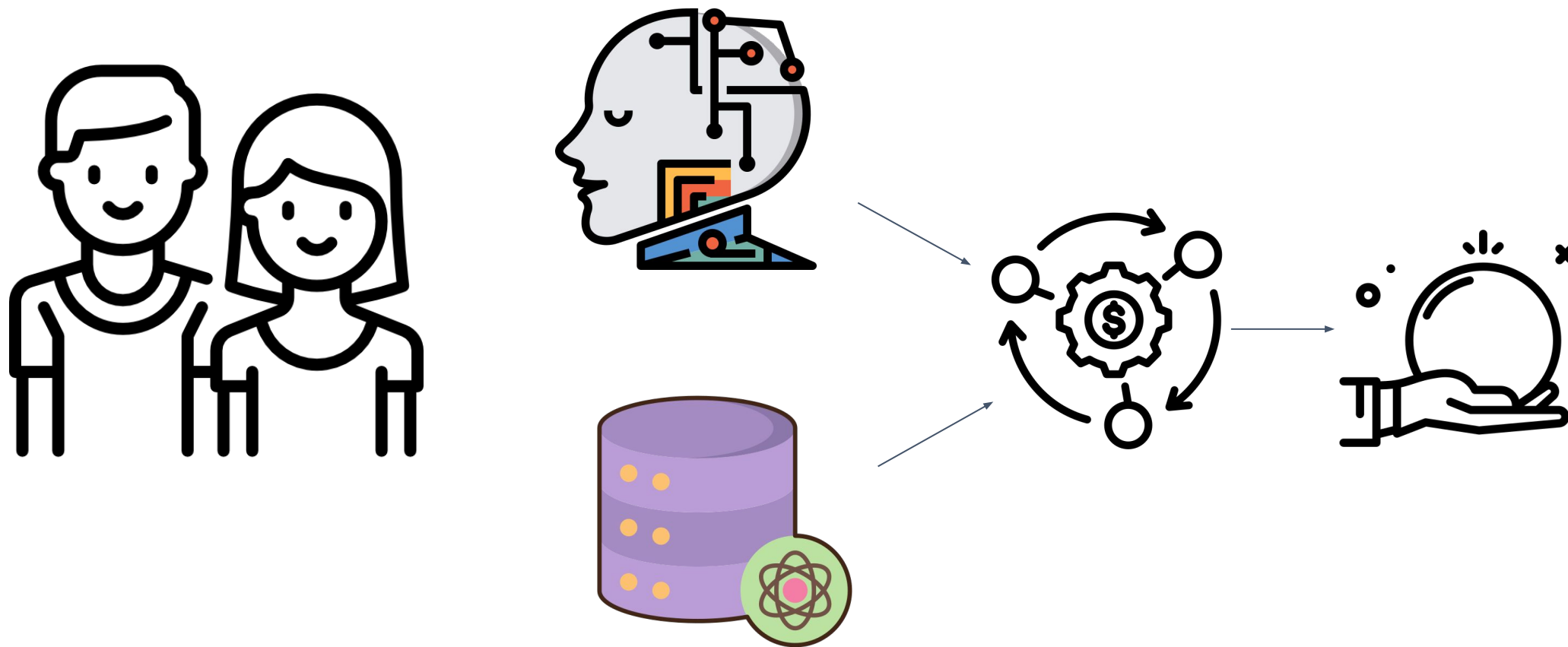
You provide efficient, lightweight off-the-shelf ML solutions for clients looking at automating their workflow.

YOUR PRODUCT

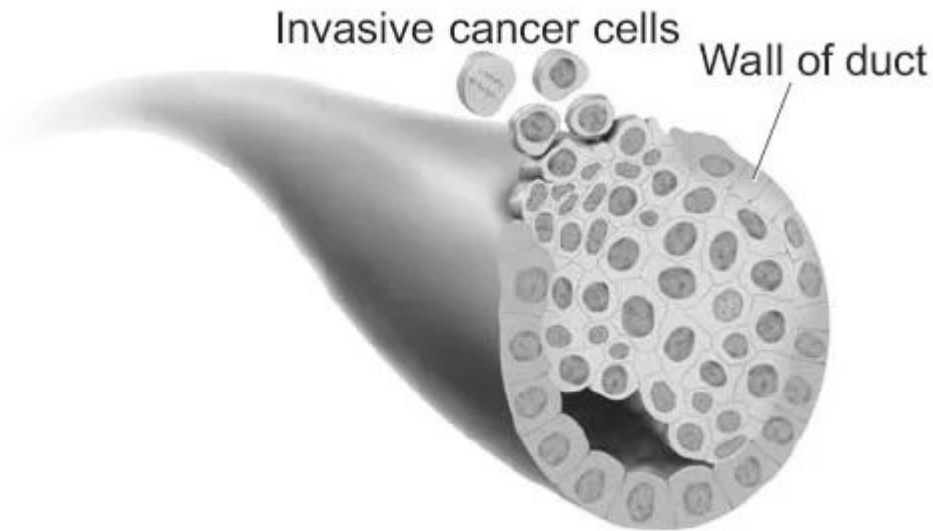
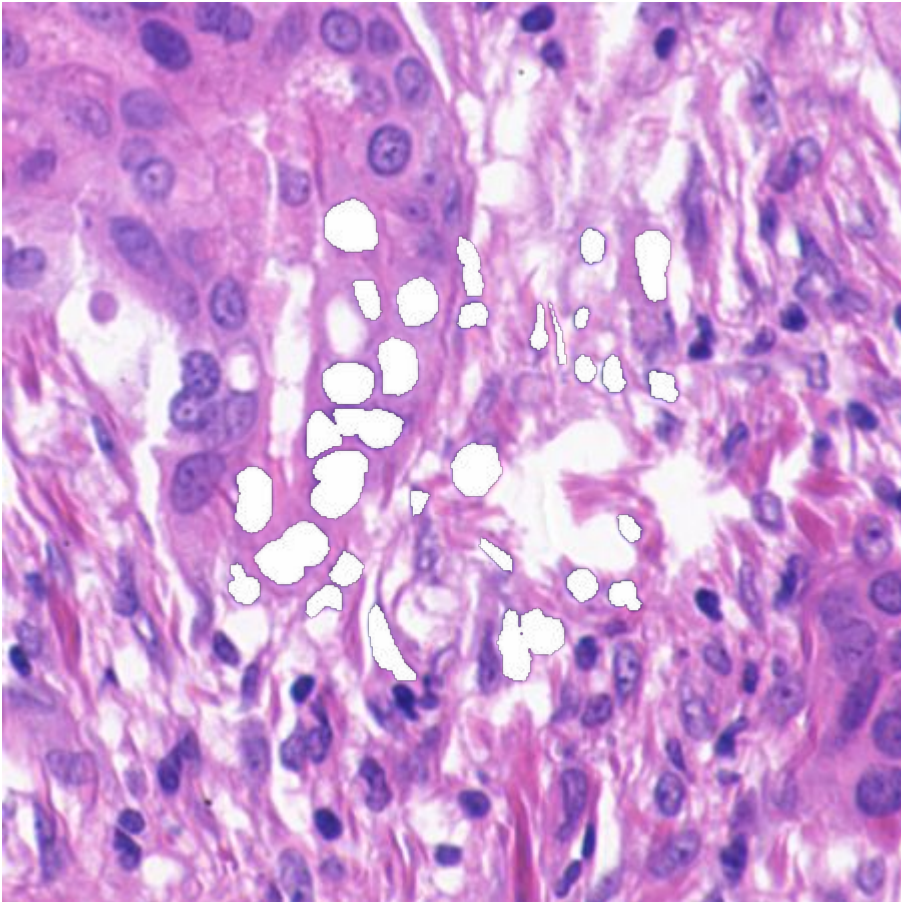
You bring a collection of pretrained ML architectures ready to be put in production as well as relevant tools to quickly assess their performance. Your strength is that you are able to work with small training datasets.

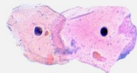


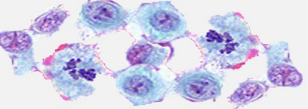

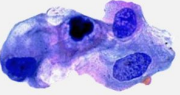
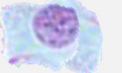

YOUR CHALLENGE

Give an easy-to-use PoC of a carcinoma classifier



#1 The dataset



Normal	Cancer	
		Large, variably shaped nuclei
		Many dividing cells; Disorganized arrangement
		Variation in size and shape
		Loss of normal features



#1 The dataset

A total of 186 histopathological slides from breast biopsies.

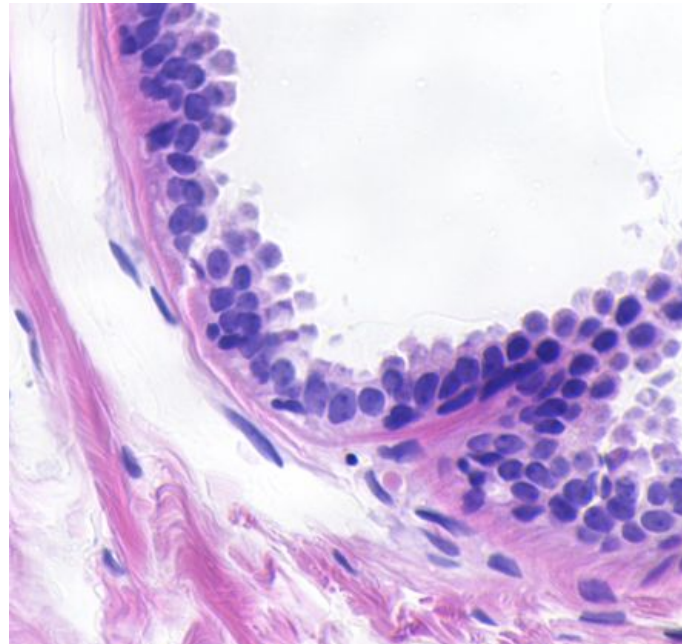
Can be split in three groups:

- Carcinoma \ominus : no tumor cells
- Carcinoma \oplus , benign: benign tumor cells
- Carcinoma \oplus , malignant: malignant tumor cells

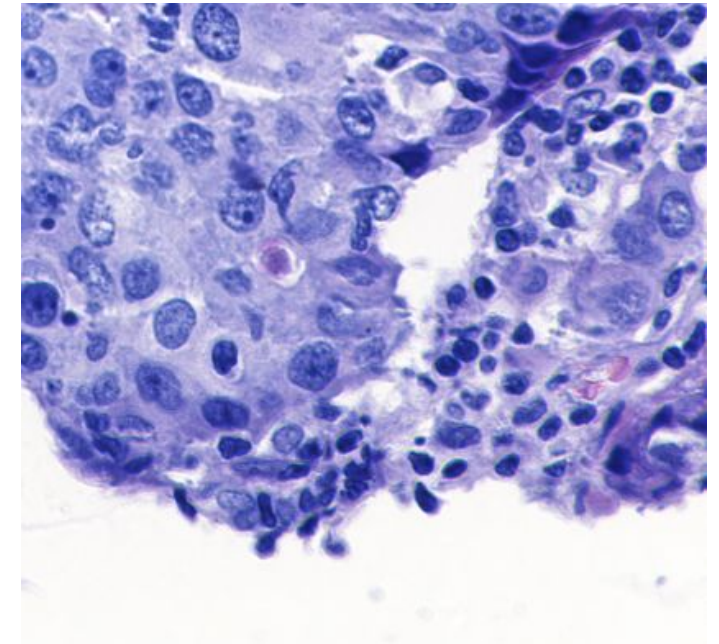
The task is to classify them accordingly.

Histopathology 101:

- Collect sample
- Fixation (formaldehyde >24h)
- Processing (dehydration...)
- Embedding (paraffin)
- Sectioning
- Staining (HES)
- Coverslipping



benign



malignant



#2 RESOURCES and TIPS

You can approach this problem in several ways:

- Using a pretrained model and fine-tune: pick up a pretrained model and fine-tune it on the training dataset. You can usually replace the last layer by a module of your choice (usually a linear layer) that you will train while keeping the rest of the model frozen.

Resources:

<https://pytorch.org/hub/>

<https://huggingface.co/>

<https://pytorch.org/vision/stable/models.html>

- Zero-shot learning: Use a trained model that has not seen the relevant classes but is capable of inferring based on a suggestion of classes.

<https://huggingface.co/tasks/zero-shot-classification>

- Gaussian process or similar (NP etc): Allows you to make Bayesian inference based on the data available. Given data X_{train} and classes Y_{train} , you can use a GP (not a GP :p) to infer $Y_{\text{valid}} | X_{\text{valid}}, X_{\text{train}}, Y_{\text{train}}$.

You will still need a pretrained model to embed your images.

You can also use more “classical” ML solutions: SVMs or XGBoost

<https://botorch.org/>





#3 The competition



kaggle

<https://www.kaggle.com/competitions/oxml-carinoma-classification>

Finding a team: <https://www.kaggle.com/competitions/oxml-carinoma-classification/discussion/414087>

