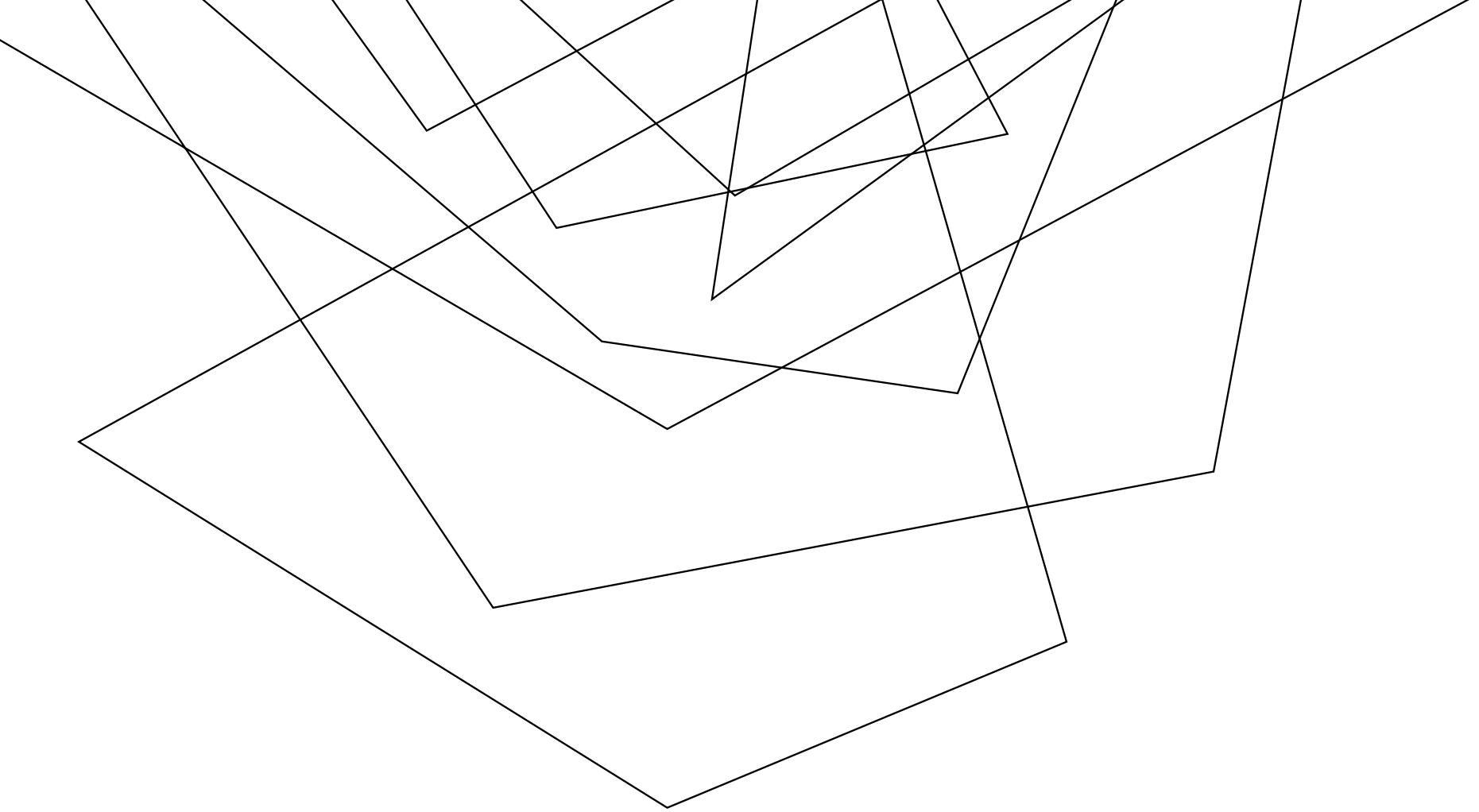


ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

LLM - Large LUNGage Model

Tommaso Giacometti
Niccolò Barbieri

Fake or not? Who knows?

DEEP FAKES? NOT ALWAYS THAT BAD

Why Synthetic Images?

- Fill gaps when real data is scarce
- Bypass privacy restrictions
- Enable controlled, reproducible experiments

Key Challenges:

- Limited training data for image generators
- Ensuring realism and variability
- Avoiding bias and overfitting
- Guaranteeing reproducibility

Can AI fool AI?

Can images generated by another neural network trick a classifier?

Your Challenge

How good is AI at recognizing AI?

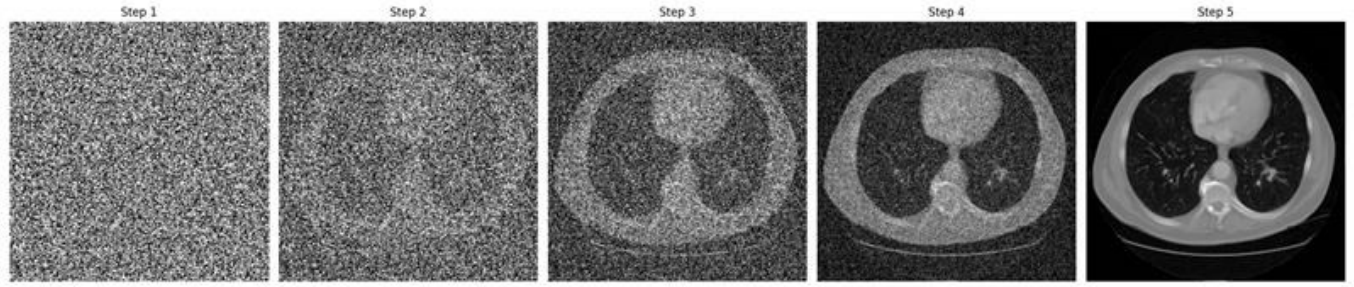
→ It's up to *you* to test and evaluate them today.

HOW CAN WE CREATE GOOD FAKE IMAGES?

Conditional Flow Matching (CFM)

Learns to generate images by **matching the flow between data distributions**, conditioned on specific labels (e.g. **position of the slice**), starting from a **simple and sampleable distribution** (e.g. a Gaussian).

$$x_1 = x_0 + \int_0^1 v_\theta(x_t, t, y)\, dt$$



FAKE OR NOT? WHAT YOU THINK?

FAKE OR NOT? WHAT AI THINK?

You will implement some **supervised learning *classifiers*** able to recognise real and synthetic images!

Let's **split** into 5 teams (around 10 members per teams)

Each team will:

- Choose a team **name**, a **leader** and divide the work
- **Implement** and **train** some classifiers
- **Evaluate** the performance of each classifier (following some metric)
- **Present** the found results (possibly with nice slides)

This is a **challenge**, so there will be a **winning team!**
The evaluations are based on:

- **Suited metrics** (accuracy and precision)
- Subjective **graphical design** of slides and **presentation quality** (by speakers and tutors)

WHAT WILL YOU HAVE?

Training Dataset

To train your models.

Set of images and corresponding lables (1 for real and 0 for synthetic)

Test data

To test your models.

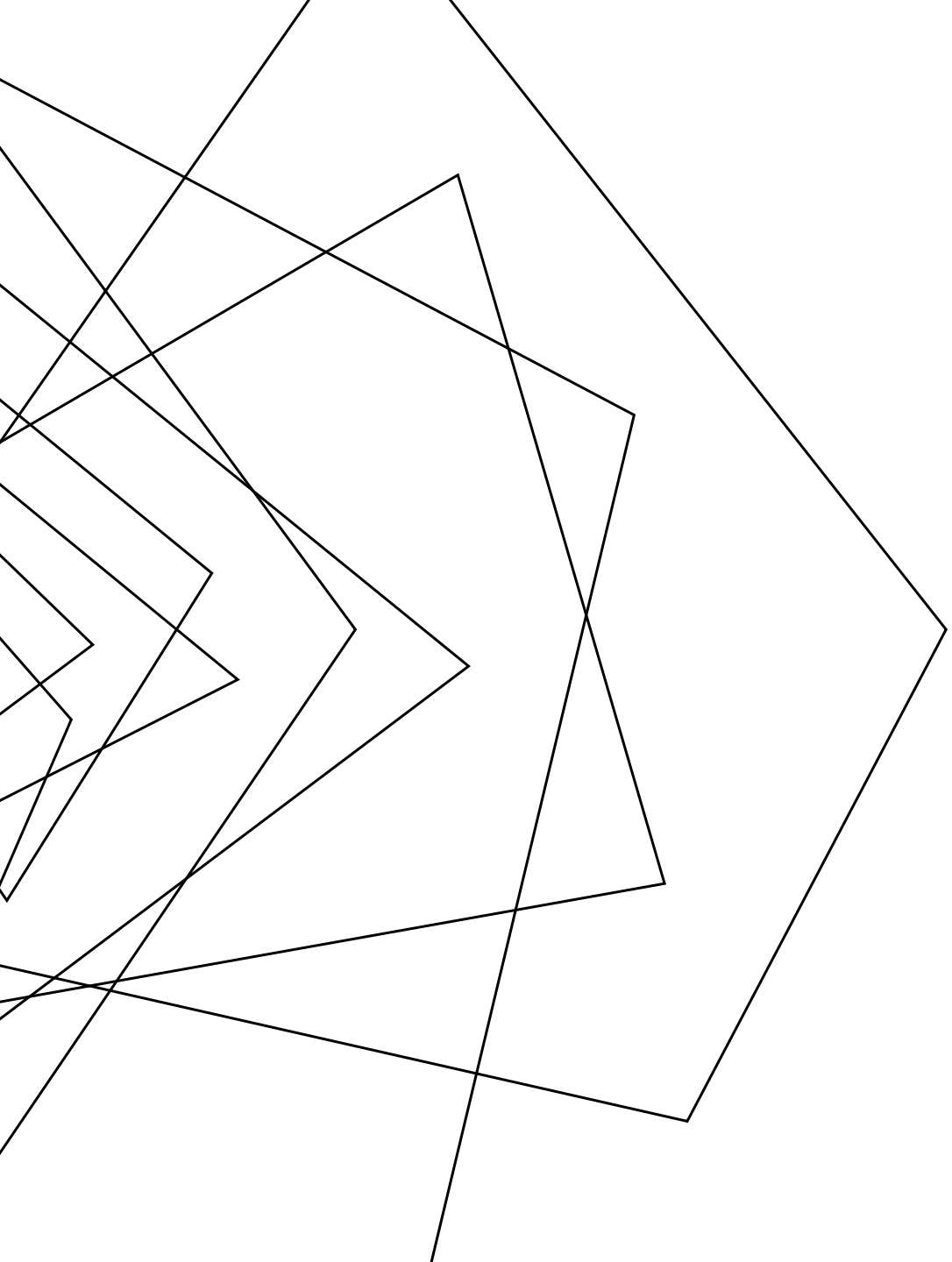
Set of images without corresponding labels (you will recive the labels at the end of the laboratory)

Jupyter Notebook

Trace of the script in Python to help you in the data downloading and managing.

Tutor

Don’t hesitate to ask if you have any questions or issues.



SUGGESTED MODELS FOR SUPERVISED LEARNING:

Machine learning (sklearn):

- Logistic Regression
- Random Forest

XGBoost

Deep Learning (pytorch, tensorflow):

- MLP
- CNN

Advanced:

feel free to choose whatever you want

YOUR PLAYGROUND:

Download the notebook and **load** it on **Google Colab**:

Look at README.md on **github** repo:
TommyGiak/DIFA-Summer-School-Lab---Large-LUNGage-Model

Or https://drive.google.com/file/d/1zo_-mYM4FZw0tPQsT3DI_y60UyRfiEU8

In the **notebook**, you will find instructions for **downloading** and **managing** the **data**.

YOUR RESULTS:

Evaluation metrics

- Accuracy
- Precision score

Presentation

- Model used
- Metrics description and results

Be aware of:

- Scarce data -> *not so big data*
- Overfitting

FAKE OR NOT: HOW GOOD WHO'S BEST?

Let's **test** your **model** with a different set.
Big prizes for the **winner**!

File ID:
1KO_RrxkcYhlATHRFnrwi8ZNLLwSySIJe

THANKS FOR YOUR ATTENTION