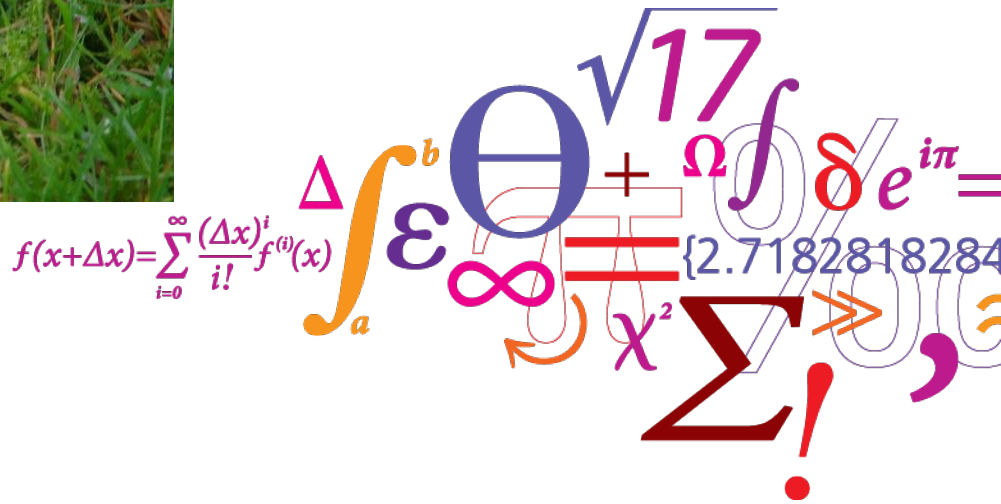
Danish Fungi Challenge



DTU Compute
Department of Applied Mathematics and Computer Science

Danish Fungi Challenge

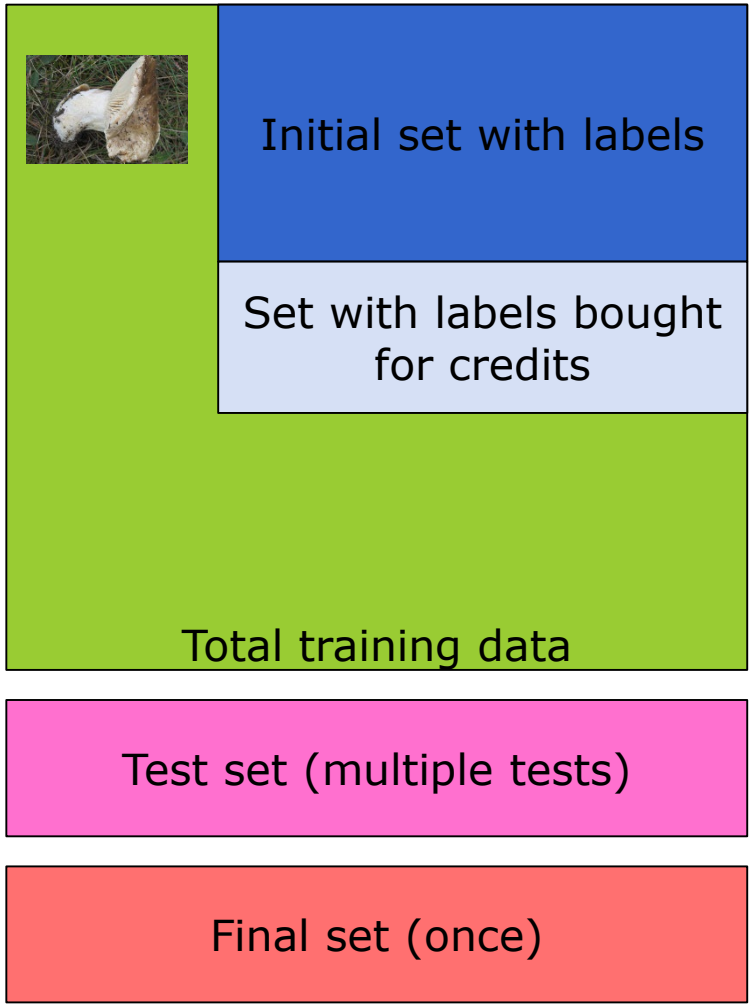
- The goal is to train a network that can identify the species based on a photo of a fungi
- The goal is to achieve the highest score (accuracy, F1) by *intelligently* selecting training data
- We give you a basic classification network
- You should:
  - Train and evaluate this network and the available data
  - Find a strategy for selecting additional data
  - Achieve your best score following your strategy

Why is this relevant?

- Getting reliable ground truth labels is often hard and expensive
- You can pay a domain expert to assign labels
  - It has a cost per image
- We have lots of images
  - *How do we choose which of these images that we should pay the expert to label?*

Data

Similar class distributions in training/test/final sets



Training set

- Total of 13033 images
- 3196 with ground truth labels
- You can buy more ground truth labels from this set

Test set

- Total of 8117 images
- Will be used to test (multiple times) your results during the school
- You will not have access to ground truth labels
- You can get the accuracy/F1-score for your latest submission

Final set

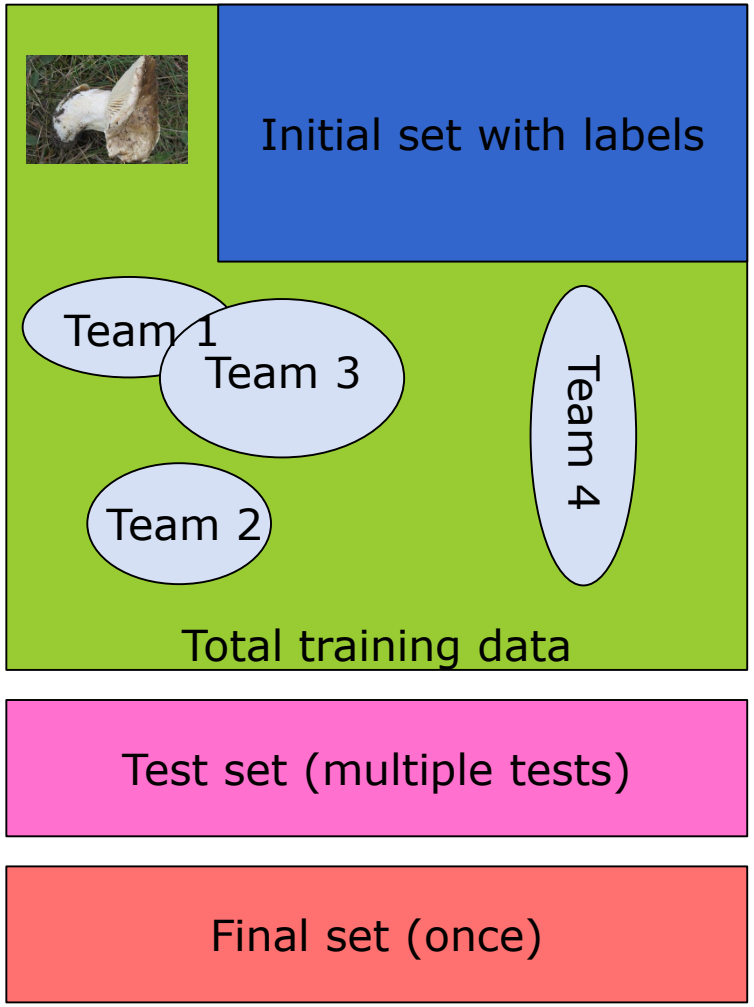
- Total of 8407 images
- For the final (once) test of your algorithm
- You will not have access to ground truth labels

Data – fungi species




- The goal is to classify fungi by their species – in this set called the taxonID.
- There are 183 different species in the set
- Highly unbalanced
- Typical classification accuracies is in the range of 55-75

Conceptual idea



- What subset of the total training data set will give the highest classification scores on the test and final set?
- Each team starts with 5.000 credits
  - Buying the label of one image in the ”total training data” costs 1 credit.

Teams

- We have pre-made 15 teams.
  - Named after animals that actually eats fungi.
- You should add your name and email to a team on the paper in the conference room
- Each team should have a least one computer with a reasonable GPU or access to a GPU cluster
- Data pre-loaded on DTU clusters
- Each team get a USB stick with data and code

What is on the USB stick?

- All the images from all sets
- A Python API
  - Code to get labels and submit results to and from our challenge database server
  - A basic network based on efficient net

Getting starting with the Python code

```
if __name__ == '__main__':
    # Your team and team password
    team = "DancingDeer"
    team_pw = "fungi44"

    # where is the full set of images placed
    image_dir = "C:/data/Danish Fungi/DF20M/"

    # where should log files, temporary files and trained models be placed
    network_dir = "C:/data/Danish Fungi/FungiNetwork/"

    get_participant_credits(team, team_pw)
    # request_random_labels(team, team_pw)
    get_all_data_with_labels(team, team_pw, image_dir, network_dir)
    train_fungi_network(network_dir)
    evaluate_network_on_test_set(team, team_pw, image_dir, network_dir)
    compute_challenge_score(team, team_pw, network_dir)
```

- Copy data and code to your PC
- Get all required Python dependencies
  - Conda / pip or whatever
- Open fungi_classification.py
- Change the team and team_pw
- Update
  - image_dir (where the raw images are)
  - network_dir (currently just an empty dir)
- Try to run the functions
- Do NOT run request_random_labels
  - Use it for inspiration on how to buy labels

Dependencies (some of them)

- Albumentations
- efficientnet-pytorch
  - https://github.com/lukemelas/EfficientNet-PyTorch
- OpenCV
- PyTorch with GPU support
- tqdm
- mysql-connector-python

Running on the DTU Compute GPU clusters

- The data has been pre-loaded here:
- GPU Cluster:
  - /nobackup/fungiimages
- HPC:
  - /dtu-compute/fungiimages

- You should copy the images to a local cluster drive
  - For example /scratch/ for the GPU cluster

https://itswiki.compute.dtu.dk/index.php/GPU_Cluster

Challenge results

- The team results are computed and put on the homepage several times a day
- Based on the test set
- https://human-in-the-loop.compute.dtu.dk/challenge
- You should submit test scores AT LEAST once per day
- You should also submit the final set scores latest Thursday at 16h!

Presentations and results

- The final project presentations and results are on Thursday from 17:10-18:45

- Each team has 4 minutes to present their project with maximum 3 slides
  - How did you design your strategy?
  - Did it work as expected?
  - Etc.

- Finally, the final test results are presented by the organizers

Tips

- Divide your data into a training and validation set (check for overfit etc)
  - You do not need to use the pre-defined number of epochs

- Careful with your credits
  - Make a strategy before spending them

- Run your trained model on the available training data
  - Find a way to measure how certain the network is on the labels on a given image
  - Decide if you want to get labels for the easy or hard cases?

Rules

- We are not checking for cheating and believe in fair play and that you are here to learn

- We do not recommend you to:
  - Use other teams account/password
  - Get images and labels from external sites
  - Hack or modify the fungichallenge database API functions (participant.py)
  - Use other classification networks
    - The goal is to sample intelligently and not to optimize a classification network architecture

Have fun!