# Hydrosat Machine Learning Scientist Challenge

You are tasked with addressing the problem of **land use and land cover (LULC) classification** in high-resolution aerial imagery.

## 1. Dataset:

You will be working with the **FLAIR #1** dataset from the FLAIR #1: semantic segmentation and domain adaptation challenge, organized by IGN (the French National Institute of Geographical and Forest Information). This dataset focuses on large-scale land cover mapping through semantic segmentation of aerial imagery.

You can access the dataset, along with a description, baseline models, and other relevant materials in the same link provided above, so make sure you go through them. Although the same link includes materials for the FLAIR #2 challenge, that is outside the scope of this task and your focus should remain on FLAIR #1

The goal for this task is to develop a **neural network-based solution** for LULC mapping using high-resolution aerial images.

While you can explore the imagery and labels in GIS software like QGIS, this is not a requirement for completing the challenge. The focus should remain on solving the problem from a **computer vision** perspective, rather than emphasizing the geospatial aspect of the imagery.

Because of computational constraints, you may want to work only with a subset of this data. Depending on the computational resources you have, it's up to you to decide how much of the imagery you want to use.

## 2. Tasks:

Once you have the dataset, you are expected to perform the following steps:

1. **Explore the dataset**: Familiarize yourself with the data to define your approach and methodology.
2. **Preprocess/transform the data**: Convert the data into a format that can be consumed by your neural network model.
3. **Develop a neural network-based solution**: Implement a model for **semantic segmentation** of aerial images.
4. **Train and evaluate**: Train your model(s) and evaluate its performance, considering your computational limits and data subset.

5. **Analyze and report**: Evaluate your results and prepare a comprehensive report.

You are free to choose how to approach the problem, what method(s)/architecture(s) to implement, and what metrics to choose for evaluation. Also, framework/tools/libraries are up to you to decide on. The only restrictions are not to use the FastAI framework, statistical machine learning methods, or 1D-based models.

## 3. Deliverables:

**1. Report:**

Please submit a concise summary report with a **maximum of 4 pages** containing:

- **Dataset findings**: Your observations about the dataset, potentially including plots and visualizations;
- **Solution explanation**: Explain your approach and outline the rationale behind your methodology and implementational choices. Also include a short (max. ½ page) overview of the key concepts of the chosen neural network.
- **Training and evaluation notes**: Highlight any key points or observations regarding the training and evaluation procedures.
- **Results analysis**: Provide visualizations and a discussion of your results;
- **Pros and cons**: Discuss the strengths and weaknesses of your solution;
- **Future improvements**: Suggest what you would do next if given more time or resources;

The report should be in the format of a Google doc or a PDF file, and should tell us how and why you made your choices.

**2. Source code:** Submit your code in the form of a Jupyter notebook and/or an executable Python project.

The overall goal is to provide a minimal solution, not to over-optimize on some metrics. We are interestedin your approach more than the actual results.

If you have any questions, feel free to contact rsleimi@hydrosat.com

***Note:*** Please don't make the results of your challenge public. Making a github repo is fine, but then make it private.