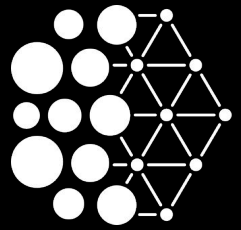
Institut
québécois
d'intelligence
artificielle



Mila

# Horoma project
# Block 2

Francis Grégoire
Mathieu Germain


horomaai
latin n.; vision
Télédétection et photo-interprétation automatisées
Automated remote sensing and photo-interpretation

# Horoma project

- **Task:** Analysis of images of forest canopy to determine:
  - Tree specie.
  - Tree density.
  - Tree height.

- **Main challenge:** Data labeling is made by human interpreters and is therefore costly and time consuming.

- **Goal:** Develop an unsupervised or semi-supervised machine learning system capable of predicting forest canopy properties with no or small amounts of labeled data.

# Project summary

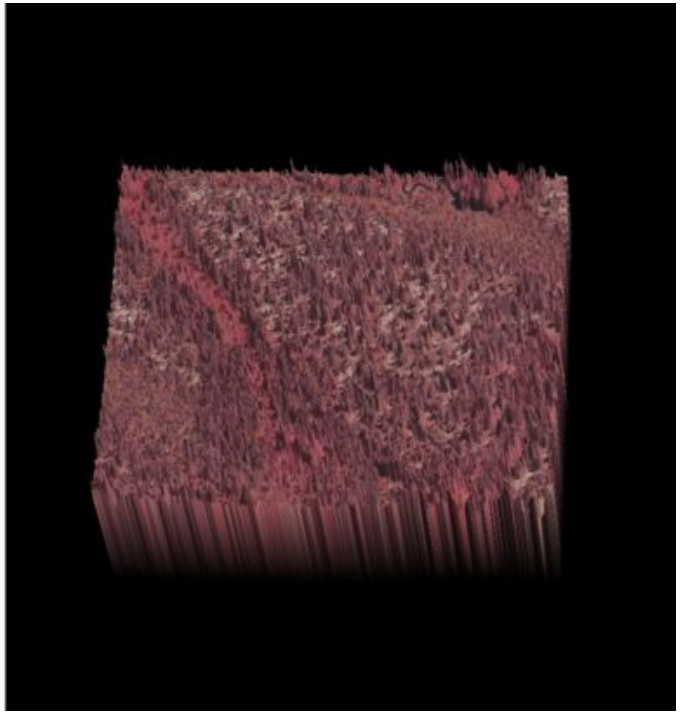
- **Overall task**:
  - Develop a model that predicts the labels of 32 x 32 pixel patches.
- **Block 1**:
  - Unsupervised training (i.e. labels are provided for the validation set; no labels are provided for the training set).
  - Structure data in a way that is useful to develop models for blocks 2 & 3.
  - Apply *K*-means clustering as a baseline model.
- **Block 2**:
  - Unsupervised training.
  - Develop more advanced clustering algorithms leveraging deep learning methods.
- **Block 3:**
  - Semi-supervised training (i.e. labels are still given for the validation set, and labels are also provided for a small number of training examples).
  - Develop semi-supervised classification algorithms.

# Objectives

- We expect you to:
  - Use only unsupervised training.
  - Explore and select relevant deep learning-based models.
  - Apply clustering algorithms.
  - Analyze your results.
  - Document your ideas and your experiments.

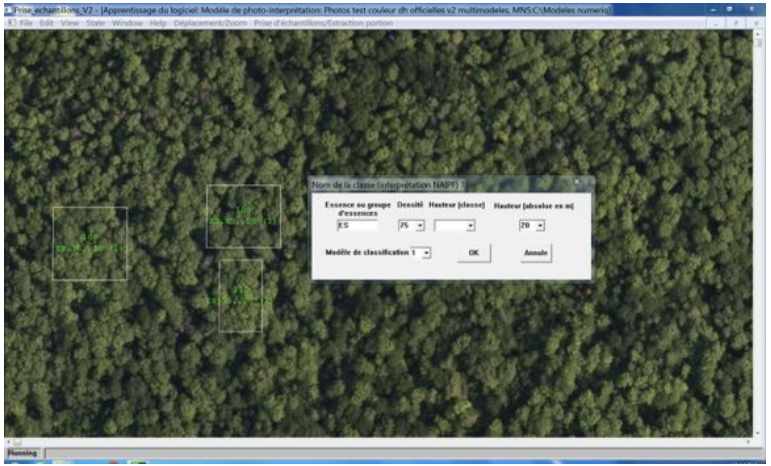
# Data (image)

- 3 Digital Surface Models (DSM) of forest areas in the Outaouais, a Western Quebec region.
- An image (or DSM) covers approximately 5km*3km (15km$^2$) of land, such that a single pixel represents an area of about 30cm*30cm.
- All images were taken during the same period of the year when trees are leafy (thus facilitating image labeling).



# Data (image labeling)

- Each image was labeled by a different human interpreter.
- Interpreters only labeled subsections of an image, with each subsection containing trees of the same overall specie, density and height (w.r.t the forest floor).
- The tree specie, density and height values were estimated by the interpreters.



# Data (image labeling)

- Tree density and height estimated by interpreters are approximate measurements.
- Interpreters base their measurements on the diameter and shape of treetops.
- Tree height measurement (which is rounded at 5m) is difficult to determine when there are no gaps in the image allowing the expert to see the forest floor.

# Data (pixel)

- Each pixel of an image has 4 values associated with it:
    - **RGB colors** (3 values). Those values are in [0, 255].
    - **Height** (1 value). The height values were obtained using photogrammetry and were georeferenced; they are measured w.r.t. the sea level.

# Data (pixel patches)

- 32 x 32 pixel patches were extracted from labeled image subsections.

- **Inputs**: each pixel in a 32 x 32 pixel patch has 4 dimensions:
  - RGB colors in [0, 255] (3 values).
  - Relative height originally defined as the height w.r.t. the sea level, where we subtracted from it the smallest pixel height within the 1024 pixels of the 32 x 32 pixel patch:

$$relative\ height_j^i = height_j^i - \min(height_1^i, height_2^i, \ldots, height_{1024}^i)$$

    where *i* = index of pixel patch and *j* = index of pixel in pixel patch *i*.

- **Outputs**: each 32 x 32 pixel patch has 3 labels:
  - Tree specie.
  - Tree density.
  - Tree height w.r.t. the forest floor.

# Data (pixel patches)

- A 32 x 32 pixel patch always contains trees of the same specie, density and height.
- **Image 1**: 430,768 patches.
- **Image 2**: 575,996 patches.
- **Image 3**: 1,011,008 patches.
- **Total**: 2,017,772 patches.

# Data (train, valid and blind test sets)

- **Blind test set**:
  - 43,077 patches randomly selected from the 1st image.
  - 57,600 patches randomly selected from the 2nd image.
  - 101,101 patches randomly selected from the 3rd image.

- **Validation set**:
  - 43,077 patches randomly selected from the 1st image (not in test set).
  - 57,600 patches randomly selected from the 2nd image (not in test set).
  - 101,101 patches randomly selected from the 3rd image (not in test set).

- **Train set**:
  - 344,614 patches (leftovers from the 1st image).
  - 460,796 patches (leftovers from the 2nd image).
  - 808,806 patches (leftovers from the 3rd image).

# Data (input format)

- Inputs are provided as binary numpy.memmap files in **float32**.
- Memory-mapped files are used for accessing small segments of large files on disk, without reading the entire file into memory.
- Each pixel patch has a shape of 32 x 32 x 4 (where 4 = [R, G, B, relative height]).
- **train_x.dat**: 1,614,216 x 32 x 32 x 4.
- **valid_x.dat**: 201,778 x 32 x 32 x 4.

# Data (output format)

- Outputs are provided as 3 text files (can be easily read from a terminal).
- Each file contains 201,778 values.
- **valid_species.txt**: tree specie (2 characters).
- **valid_densities.txt**: tree density (percentage*100 rounded at lowest unit of 5).
- **valid_heights.txt**: tree height w.r.t. the forest floor (rounded at nearest 5m).
- The *i*-th value of each file is associated to the *i*-th pixel patch in **valid_x.dat**.

# Data (tips)

- This is a hard unsupervised learning task.
- Find clever ways to exploit the information that is given to you to determine the number of clusters.
- You should document your different approaches and explain your ideas in your report.
- It’s important to think beyond the numbers and to analyze your results.

# Lessons learned from Block 1

- Issues resulting from the partitioning of the data (i.e. pixel patches in train/valid/test sets come sometimes from the same subsection ⇒ no clear cut separation between the 3 sets).
- Evaluating model overfitting is difficult.
- Data sets are noisy and unbalanced.
- It is preferable to apply multi-task learning by leveraging information from the tree species, densities and heights instead of using single task learning directly on the triples.
- Use the F1 score instead of the accuracy to evaluate models (due to data imbalance).
- Need to be careful during evaluation (do not only rely on numbers to make decisions).
- We encourage you to think about the data and the models that you are using.

# Block 2 instructions & expected timeline

| | Feb 13 & 15 | Feb 20 & 22 | Feb 27 & Mar 1 | Mar 13 & Mar 15 |
|---|---|---|---|---|
| Tasks | • Read about deep learning-based unsupervised models (see last slide)<br>• Select code from Block 1 | • Start developing deep learning-based models<br>• Apply clustering algorithms on latent representations | • Finish deep learning-based models<br>• Evaluate models against the best performing model from Block 1 using valid set | • Write a short report summarizing the work, and results<br>• Peer review the code of other teams |
| Objectives / Deliverables | • Understand the data<br>• Understand deep learning-based unsupervised models<br>• Build on top of code from Block 1 | • Understand clustering algorithms | • Models and clustering ready (beginning of week after spring break)<br>• Data visualization ready (beginning of week after spring break) | • Produce documented code and report summarizing the experimental work<br>• Provide model for blind test set evaluation<br>• Complete peer code review |

# Deadlines

- Each team needs to provide the deliverable (report + code + best model) corresponding to a block at the latest on **Friday 11:59pm** of the last week of the block.

- Any block deliverable that is provided past **Friday 11:59pm** of the last week of a block will automatically get 0% for the peer evaluation.

# Deadlines

- Any block deliverable that is provided past **Tuesday 11:59pm** following the last week of a block will automatically get 0% for the UdeM evaluation.

- Peer evaluation must be completed by **Monday 11:59pm** following the last week of a block.

# Evaluation grid for block 2

- **25% of the final score**.
- 10% Code review [5% of averaged peer evaluation + 5% UdeM].
- 12% Report evaluation [UdeM].
- 3% Model performance evaluation on blind test set [UdeM].

# Code review - Peer evaluation

- **10% of the final score** [5% of average peer evaluation + 5% UdeM].
- Random assignation of code reviews.
- The code provided by a team will be evaluated by at least 2 other teams.

| Code quality (peer evaluation + UdeM evaluation) | /8 |
|---|---|
| Coherent and modular code/file organization (e.g. data processing, model definition, model training, model inference are in different files/modules; no code duplication) | /1 |
| Code respects the PEP8 standard | /1 |
| Comments are relevant (see article) | /1 |
| Proper management of input arguments in the training script (see argparse, python fire, configparser) | /1 |
| Proper utilization of GitHub (e.g. branching, relevant commits and messages, usage of pull request) | /1 |
| Meaningful variable and function names | /1 |
| Executable scripts with a “main” function (see article) | /1 |
| Reproducible experiments (e.g. seed) | /1 |

| | |
|---|---|
| Introduction | /2 |
| Introduction to the project | /1 |
| Brief introduction to the methods that will be used in the report | /1 |
| Methodology | /6 |
| Description of the algorithms and the experiments (including a description of the approaches used to fine tune the hyperparameters, select the best “model” using checkpointing, etc.) | /3 |
| Data description and data selection (train/valid/test, number of samples, shape/structure of data points) | /3 |
| Results and discussion | /6 |
| Presentation of results (tables, figures, etc.).<br>Note that this should include:<br>• A comparison with results from the previous block.<br>• Figures showing the loss value across epochs/checkpoints and models (using tensorboard). | /2 |
| Discussion of results | /4 |
| Conclusion | /2 |
| Recommendation for next steps | /1 |
| Summary of project state (what was done, what needs to be done) | /1 |
| Quality of the report | /2 |
| Report format (title with team member names, clear sections, flow between sections, figures and tables titled, axes titled, etc.) | /1 |
| Report is short and to the point (5-7 pages including references, font size 11) | /1 |

# Report evaluation

- **12% of the final score**.
- 5-7 pages (including figures, tables, and references).
- Single column.
- Font size 11.
- Use the NeurIPS LaTeX format.

# Blind test set evaluation

- **3% of the final score**.
- If the best model provided by a team crashes or provides results that are statistically worse than those of the baseline model provided by the TAs, the team gets 0%.
- Otherwise, if the best model provided by a team is statistically equivalent to the baseline model, the team gets 1%.
- Otherwise, if the best model provided by a team is statistically better than the baseline model:
    - The team gets 3% if the model is the best performing one or is statistically equivalent to the best performing model provided by another team.
    - Otherwise, the team gets 2%.

# Code execution - Blind test set evaluation

- The **test set** will be structured identically to the **valid set**.
- You will not have access to the **test set** and we will be executing your code on the **test set** ourselves.
- We will provide explicit instructions and examples for you to enhance an evaluation skeleton script that will be provided to you. You will need to complete this script. We reserve the right to give 0 if we cannot execute your code.

# Official evaluation metric

- For each patch, you need to predict its tree **specie**, **density** and **height** (3 labels).
- A patch is correctly classified if the predicted tree **specie**, **density** and **height** are all equal to the ground truth, i.e.$(\hat{s}, \hat{d}, \hat{h})_i = (s, d, h)_i$.
- Use ground truth labels found in **valid_species.txt**, **valid_densities.txt** and **valid_heights.txt** to calculate the F1 score of your predictions (see sklearn.metrics.f1_score).
- Note that you should convert the multi-label problem to a binary problem (is the predicted triple equal to the ground truth triple?).
- An evaluation script will be provided.

# Informative evaluation metrics

- You have to explore widely used metrics for clustering performance evaluation such as **Normalized Mutual Information** (NMI) and **adjusted Rand index** (ARI) (see scikit-learn documentation).

- F1 score of tree specie:

  $F1_s$ = `sklearn.metrics.f1_score(specie_pred, specie_true)`

- F1 score of tree density:

  $F1_d$ = `sklearn.metrics.f1_score(density_pred, density_true)`

- F1 score of tree height:

  $F1_h$ = `sklearn.metrics.f1_score(height_pred, height_true)`

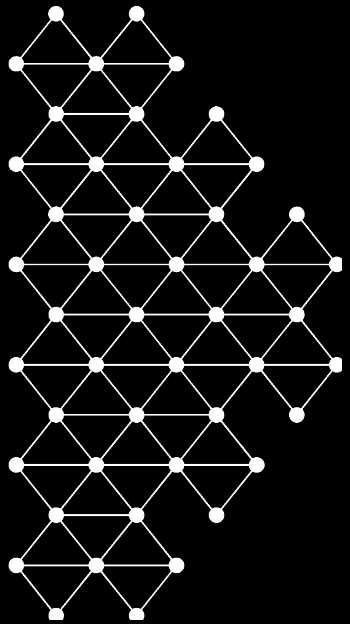
# Precision, recall, F1 score

- **Precision** is the proportion of correctly classified triple pairs among all classified triple pairs.

- **Recall** is the proportion of correctly classified triple pairs among all triple in the validation/test set.

- **F1 score** is the harmonic mean of precision and recall.

Institut
québécois d'intelligence
artificielle



# Autoencoders

# Unsupervised learning

- We assume that a data generating process contains internal variables that are not observed, but that still influence the data.

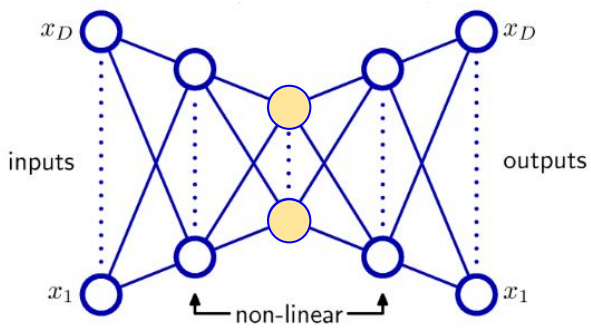
- These unobserved variables are called hidden variables or latent variables.

- We postulate that there is a latent variable $\mathbf{z}_i$ associated with each training example $\mathbf{x}_i$.

# Autoencoders

- A basic autoencoder is a feedforward neural network trained to reproduce its inputs (i.e. outputs = inputs).
- The size of the hidden layer is smaller than the input's size.
- This leads to reconstruction errors, which we seek to minimize during training.
- We obtain a latent representation of the inputs in the hidden layer.
- Traditionally used for dimensionality reduction or feature learning.

# Autoencoders

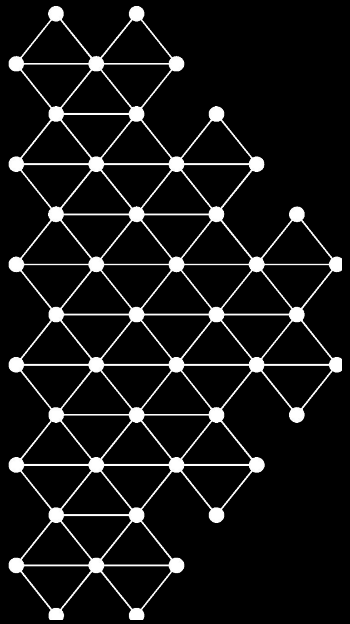



- An autoencoder can be viewed as consisting of two parts:
  - An **encoder** that maps the inputs to a latent representation in the hidden layer.
  - A **decoder** which creates a reconstruction of the inputs from the hidden layer.

# Autoencoders

- Reference: Chapter 14 - Autoencoders of the Deep Learning book.
- https://www.deeplearningbook.org/contents/autoencoders.html
- Intuitive introduction to VAE [blog][video]
- Overview of multiple types of autoencoders

Institut
québécois d'intelligence
artificielle



# *K*-means clustering

# *K*-means clustering

- Most widely used clustering algorithm.
- We consider $\mathbf{z}_i$ as a discrete latent variable.
- The inference of $\mathbf{z}_i$ is known as clustering because it leads to assigning each training example $\mathbf{x}_i$ to a group among $K$, where $\mathbf{z}_i$ represents the assigned group (cluster).
- We use one-hot encoding to model $\mathbf{z}_i$.
- e.g. if $K = 5$ and $\mathbf{z}_i$ is associated to group 4, then $\mathbf{z}_i = [0, 0, 0, 1, 0]$ ($z_{i4} = 1$).

# *K*-means clustering

- We suppose that there are *K* prototypes $\mathbf{u}_1, \dots, \mathbf{u}_K$ which represent the means of the clusters (centroids).
- The dimension of a prototype $\mathbf{u}_k$ is equal to the dimension of $\mathbf{x}_i$.
- Each $\mathbf{x}_i$ belongs to the cluster with the nearest prototype (i.e. shortest Euclidean distance).
- The loss function of *K*-means clustering minimizes the average distance between the training examples and their associated $\mathbf{u}_k$:

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \|\mathbf{x}_i - \mathbf{u}_k\|^2$$

# *K*-means clustering

- The iterative learning process corresponds to finding the prototypes $\mathbf{u}_k$ as well as the associations $\mathbf{z}_i$ that minimize $J$.

- It is common practice to initialize the $\mathbf{u}_k$ by randomly choosing $K$ training examples and using these as initial values.

# *K*-means clustering

- Finding the optimal $\mathbf{z}_i$ :
  - By knowing the $\mathbf{u}_k$ we can optimize all $\mathbf{z}_i$ independently.
  - *J* is minimized if $z_{ik}$= 1 with the closest $\mathbf{u}_k$ ($\mathbf{z}_i$ = [$z_{i1}$, $z_{i2}$, … , $z_{iK}$]) and 0 otherwise.
  - The optimal solution is:

$$z_{ik} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_i - \mathbf{u}_j\|^2 \\ 0 & \text{else} \end{cases}$$

# *K*-means clustering

- Finding the optimal $\mathbf{u}_k$ :
  - By knowing the $\mathbf{z}_i$ we can minimize $J$ by setting the derivative w.r.t. $\mathbf{u}_k$ equal to zero:

$$2\sum_{i=1}^{N} z_{ik}(\mathbf{x}_i - \mathbf{u}_k) = 0$$

  - The optimal solution is:

$$\mathbf{u}_k = \frac{\sum_i z_{ik}\mathbf{x}_i}{\sum_i z_{ik}}$$

  - The optimal $\mathbf{u}_k$ is the mean of all examples associated to the cluster $k$.

# *K*-means clustering

- The learning process consists of iterating the inference of $\mathbf{z}_i$ and the update of $\mathbf{u}_k$ until the convergence of $J$:

1. Initialize $\mathbf{u}_k$

2. Repeat until convergence {

Find the optimal $\mathbf{z}_i$

Find the optimal $\mathbf{u}_k$

}

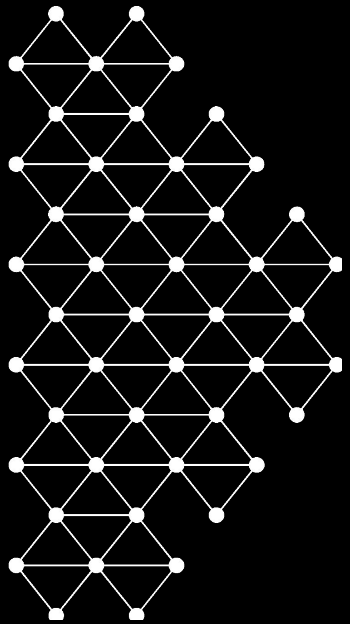
- This procedure is guaranteed to converge because $J$ is positive and at each iteration, either $J$ decreases or does not change.

# *K*-means clustering

PLAY WITH ME!

Institut
québécois d'intelligence
artificielle

Mila



# Recommended papers

# Recommended papers (non-exhaustive)

- Foundations:
  - Masci et al. - Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction
  - You are invited to search for relevant papers😎 .
- Recent approaches:
  - Xie et al. - Unsupervised Deep Embedding for Clustering Analysis
  - Yang et al. - Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering
  - Caron et al. - Deep Clustering for Unsupervised Learning
  - Chazan et al. - Deep Clustering Based on a Mixture of Autoencoders
- Survey:
  - Min et al. - A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture