# Planetary scale LiDAR classification from multi-scale voxelized PCA feature extraction: Optimization assessment

Antoine Lucas, Shen Liang, Stephen Porder, Gregory Sainton, Alexandre Fournier

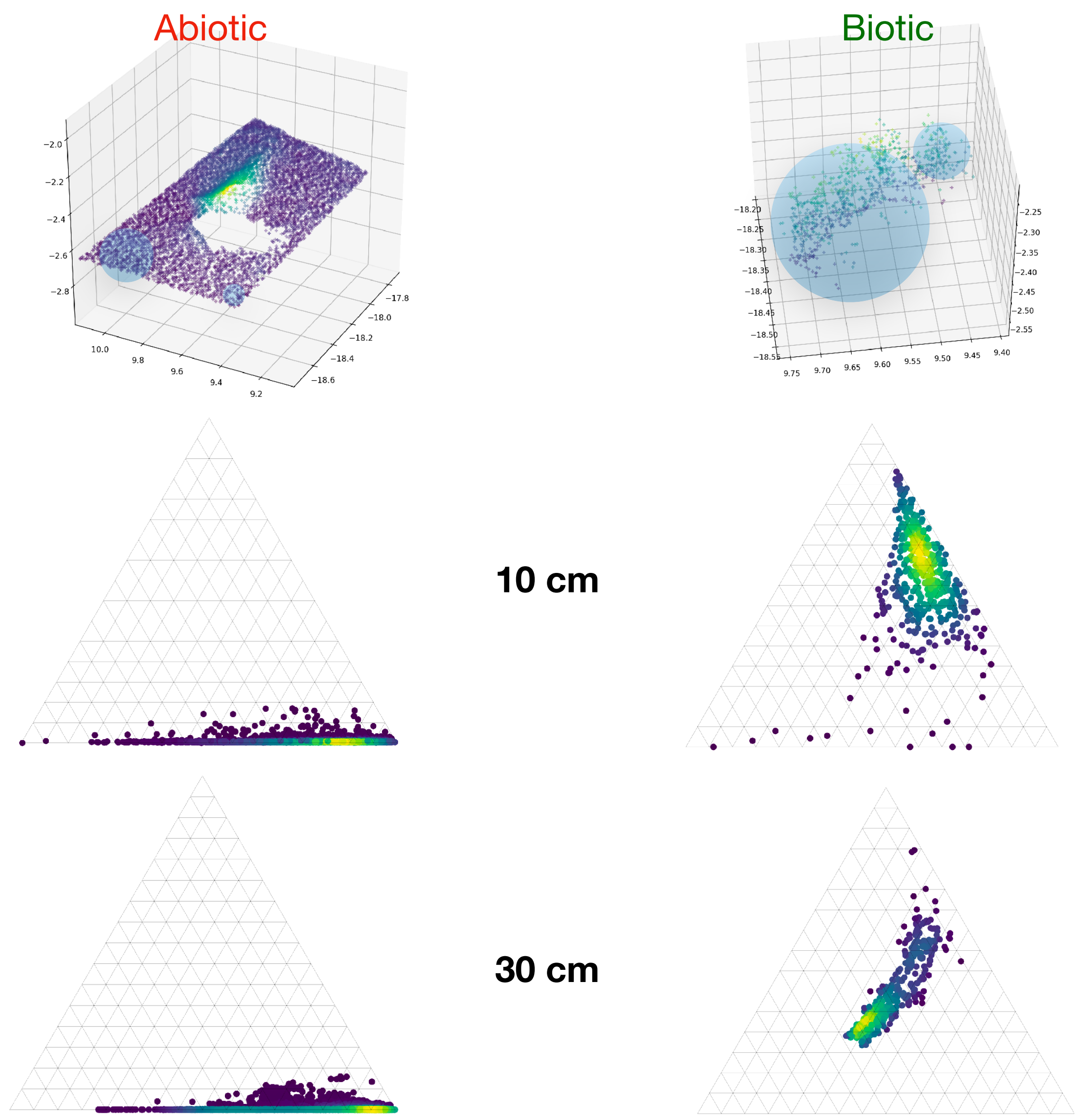
Topography, the surface configuration of a planet, bears the imprint of the processes that shape it. Deciphering the "text" encoded in topography has been the goal of scholars for centuries (Darwin, Gilbert, others). For scholars looking into space, topography is a window in the processes that shape alien worlds. Here on Earth, it offers clues to how our planet might be different, and if so why. Despite centuries of theory that the abundance of life is shaping the surface of our planet (Darwin, 1881), the study of topography has yet to uncover a unique signal on Earth.

Indeed as recently as 2006, Dietrich and Perron who asked the question is there a topographic signature of life? concluded that none could be detected from terrestrial data available back then. Today, with the revolution in Earth observations from space, we can approach this question with fresh eyes and extend it to other bodies of the Solar System. Doing so requires the interpretation of a wide range of data sets, and cutting edge mathematical approaches to their integration. This integration and interpretation are the focus of this proposal.

## Objective:

Finding topographical features relevant to specific processes including life

- Methodology:
  - Extract features from 3D point cloud data of earth and other planets (e.g. Mars, which as far as we know have no significant life activities);
  - evaluate the discriminating power of the features by feeding them into a classifier that distinguishes between earth and other planets.




## Challenges

- Feature quality: how to avoid extracting features that are discriminative yet uninteresting?
  - Features that yield high classification accuracies are not necessarily interesting. E.g., inconsistencies of data sampling are definitely not life signatures!
- Scalability: how to efficiently handle massive amount of data?

## 3D cloud Dataset Formation

- SRTM +GEDI (Earth)
- MOLA (Mars)

## Data Preprocessing (Normalization)

- Planetary Coordinate normalization with PyProj
- Fast sampling normalization with GPU-based voxelized AIDW interpolation

## Feature Extraction

- Multi-scale voxelized PCA feature extraction

## Feature Evaluation (Classification)

- Linear Discriminant Analysis (LDA)
- Support Vector Machine (SVM)

## Maximizing intra-class diversity

- Class labels: Earth, Mars, …
- Intra-class diversity: include data from different geolocations, landscapes, sampling strategies… (E.g. for the Mars class, include data from volcanos, craters…)
- Why: account for as many known interfering factors as possible to ensure feature quality (E.g. Suppose we only have data from one specific geolocation on earth and one on Mars, the extracted features may only reflect the differences between these two locations which may not apply to the rest of earth and Mars)
- **Current dataset:**
  - SRTM/GEDI: Earth data, about 19,000,000 points
    (SRTM are used for the prototyping of the algorithm for sake of efficiency)
  - MOLA: Mars data, about 428,994 points
  - More diverse data has already been added, ready for further studies!

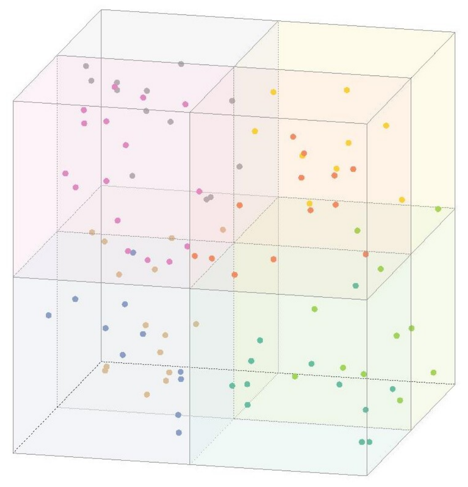
## Minimizing inter-class inconsistencies

- Intra-class inconsistencies:
  - Different geo-coordinate systems: Earth vs Mars
  - Different sampling strategies:
    - SRTM: regular (gridded) and dense
    - MOLA: irregular and sparse
- Why: mitigating as many known inconsistencies as possible to ensure feature quality (E.g. Suppose we keep the original sampling strategies, we may extract features related to sampling, not features related to the topography)
- Handling different geo-coordinate systems
  - Using the pyproj package to standardize the coordinates
- Handling different sampling strategies
  - Using interpolation methods to normalize the sampling points

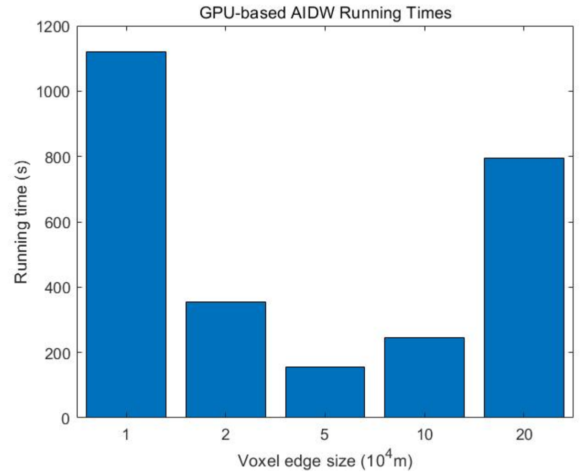
## Voxelization

- Obtaining the sub-regions
  - Classic method: Ball query
    - For each point in the region, let it and all points whose distance to it is below a given threshold as a sub-region.
    - This method is slow: the running time increases exponentially as the number of points increases.
  - Our method: Multi-scale voxelization
    - Divide the region into voxels of different edge sizes (hence "multi-scale"); each voxel is considered to be a sub-region
    - Our method is fast: the running time increases linearly as the number of points increases.
    - In the mean time, the multi-scale nature of our method is crucial to ensuring the quality of the extracted features! (More on this later)
  - Empirical study: extracting features for the MOLA and SRTM datasets
    - Classic method: 38,827s for MOLA; 128,288s for SRTM
    - Our method: 819s for MOLA; 6230s for SRTM/GEDI

- Concept
  - Due to the architecture of the GPU, it often not optimal to feed the data into it all at once. Rather, we need to divide the data into several batches and sequentially feed them into the GPU
  - Voxelization: Dividing the 3D space the data points are in into equal-sized cubes called "voxels". All points in each voxel form a batch that is fed into the GPU at a time.



- Effectiveness of Voxelization
  - Empirical study: interpolating the MOLA dataset
  - From 428,994 known points to 217,903,314 points
  - CPU-based: 3652s (optimized to have the fastest sequential processing speed)
  - GPU-based with no voxelization: 3227s
  - GPU-based with voxelization: 155s (voxel edge size = 50,000m)

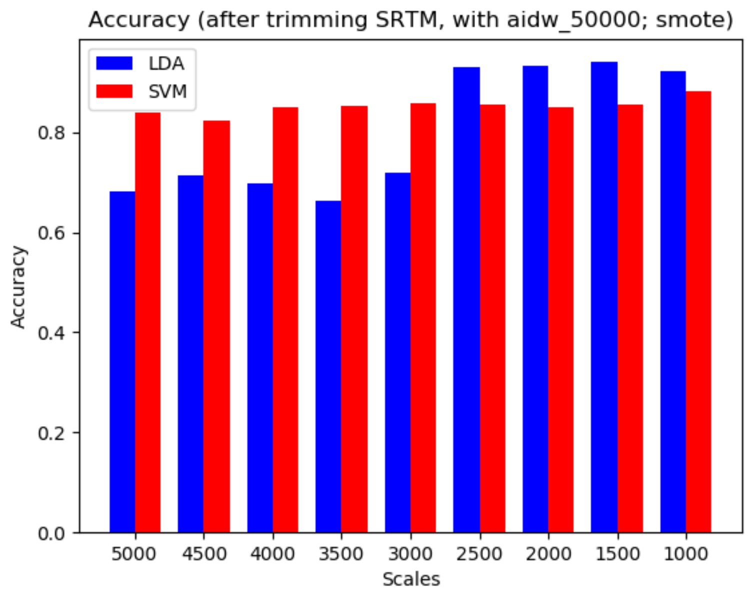



## Performances

**Scalability: Feature Evaluation (Classification) appiied to volcano landscape on Earth and Mars**

| #Features | Accuracy | Precision | | Recall | | F1-score | |
|---|---|---|---|---|---|---|---|
| | | SRTM | MOLA | SRTM | MOLA | SRTM | MOLA |
| 10 | 0.983 | 0.680 | 1 | 1 | 0.982 | 0.807 | 0.991 |
| 100 | 0.996 | 0.913 | 1 | 1 | 0.996 | 0.953 | 0.998 |
| 1000 | 0.962 | 0.514 | 0.995 | 0.895 | 0.965 | 0.651 | 0.980 |

| #Features | Accuracy | Precision | | Recall | | F1-score | |
|---|---|---|---|---|---|---|---|
| | | SRTM | MOLA | SRTM | MOLA | SRTM | MOLA |
| 10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 100 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1000 | 0.997 | 1 | 0.997 | 0.936 | 1 | 0.965 | 0.998 |




## Conclusions and perspectives

- Challenge I - Feature quality: how to avoid extracting features that are discriminative yet uninteresting?
- Current solutions:
  - Dataset formation: diversifying the datasets
  - Data Preprocessing: normalization of coordinate systems and sampling
- Future work:
  - Dataset formation: even more diversified datasets
- Challenge II - Scalability: how to efficiently handle massive amount of data?
- Current solutions:
  - Data Preprocessing: GPU-based interpolation with voxelization
  - Feature Extraction: Multi-scale voxelization
- Future work:
  - Feature Extraction and Feature Evaluation: Integrating current feature extraction method into a lightweight deep neural net, thus building a fully GPU-based workflow for topographical 3D cloud mining.

This work is supported by diiP, IdEx Université de Paris, ANR-18-IDEX-0001