THE VRTILITY R PACKAGE

Harnessing *GDAL* for efficient ML4EO

Hugh Graham
Permian Global
University of Exeter

Christopher Philipson
Permian Global
belian.earth

Andrew Cunliffe
University of Exeter

AND WHY IT IS SO COOL!

- It is the de facto standard for geospatial data processing.
- It is the most widely used geospatial library in the world.
- It is open source and has a large community of contributors.
- It has so much powerful and underutilised functionality.

PERMIAN GLOBAL

TESS Lab

UNIVERSITY OF EXETER · LUCEM SEQUIMUR ·

belian · earth

VRTILITY

- vrtility is an R package that aims to make the best use of GDAL's VRT capabilities for efficient processing of large raster datasets.
- Modular in design
- enables simple access to GDAL VRT built-in and python pixel functions.
- Advanced compositing methods that maintain spectral consistency, such as the geometric median and medoid.
- Time series filtering functions to improve temporal consistency and reduce noise.
- Efficient parallel processing .

PERMIAN GLOBAL

TESS Lab

UNIVERSITY OF EXETER · LUCEM SEQUIMUR ·

belian · earth

VRTILITY

Why build this?

geometric median + hampell time-series filter

PERMIAN GLOBAL

TESS Lab

UNIVERSITY OF EXETER · LUCEM SEQUIMUR

belian · earth

EMBED ML MODELS IN VRTS

Now, what you've all been waiting for… raw XML! 🤓

PERMIAN GLOBAL

TESS Lab

UNIVERSITY OF EXETER · LUCEM SEQUIMUR ·

belian · earth

EMBED ML MODELS IN VRTS

SCL CloudMask

PERMIAN GLOBAL

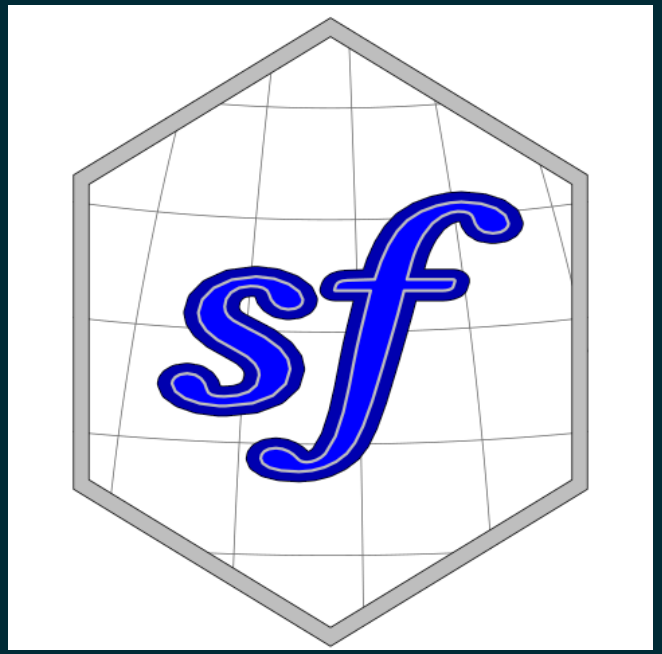
TESS Lab

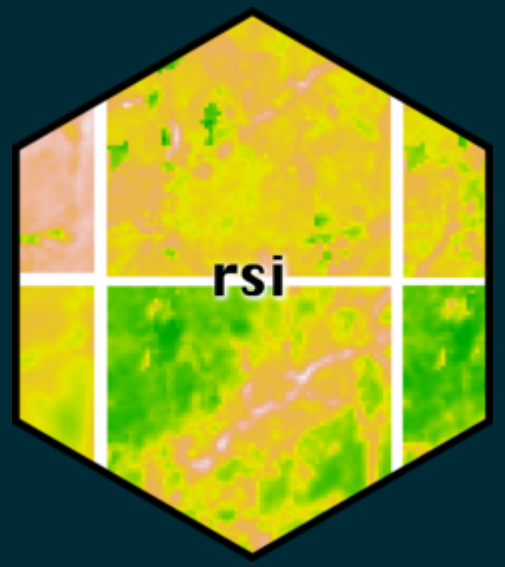
UNIVERSITY OF EXETER · LUCEM SEQUIMUR ·

belian · earth

SOFTWARE ABSTRACTIONS

I love them but sometimes we need more!





PERMIAN GLOBAL

TESS Lab

UNIVERSITY OF EXETER · LUCEM SEQUIMUR ·

belian · earth

ANALYSIS READY DATA (ARD)

- If it's too good to be true, it probably is!
- There is a plorifertation of ARD products. This seems very tempting on the face of it.
- Remember there are many decisions that are made in the production of these products. Make sure they are appropriate for your use case.
- If you can - build your own.
- Fully understanding data preprocessing can make building ML models much easier.

PERMIAN GLOBAL

TESS Lab

UNIVERSITY OF EXETER · LUCEM SEQUIMUR ·

belian · earth

AWESOME NEW RELATED STUFF (THAT I NEED TO LEARN ABOUT!)

- New GDAL CLI
- Zarrr: a format for the storage of chunked, compressed, N-dimensional arrays.
- With new ML methods will we be able to abandon this preprocessing altogether one day? see pyRawS (processing level-0 sentinel 2 data.)

PERMIAN GLOBAL

TESS Lab

UNIVERSITY OF EXETER · LUCEM SEQUIMUR ·

belian · earth

SUMMARY

PERMIAN GLOBAL

TESS Lab

UNIVERSITY OF EXETER · LUCEM SEQUIMUR ·

belian · earth

- All data processing has implications for downstream ML tasks. It helps to be aware of associated limitations and trade-offs.
- GDAL VRTs are a powerful tool for efficient processing of large raster datasets. you can work with them using vrtility or just use GDAL directly.
- embedding ML workflows directly into virtual files can be more efficient - potentially opening up many new applications.
- software abstractions are great but sometimes we need to get our hands dirty.
- Improving data processing pipelines will almost certainly have the largest impact on model performance.

PERMIAN GLOBAL

TESS Lab

UNIVERSITY OF EXETER · LUCEM SEQUIMUR

belian · earth