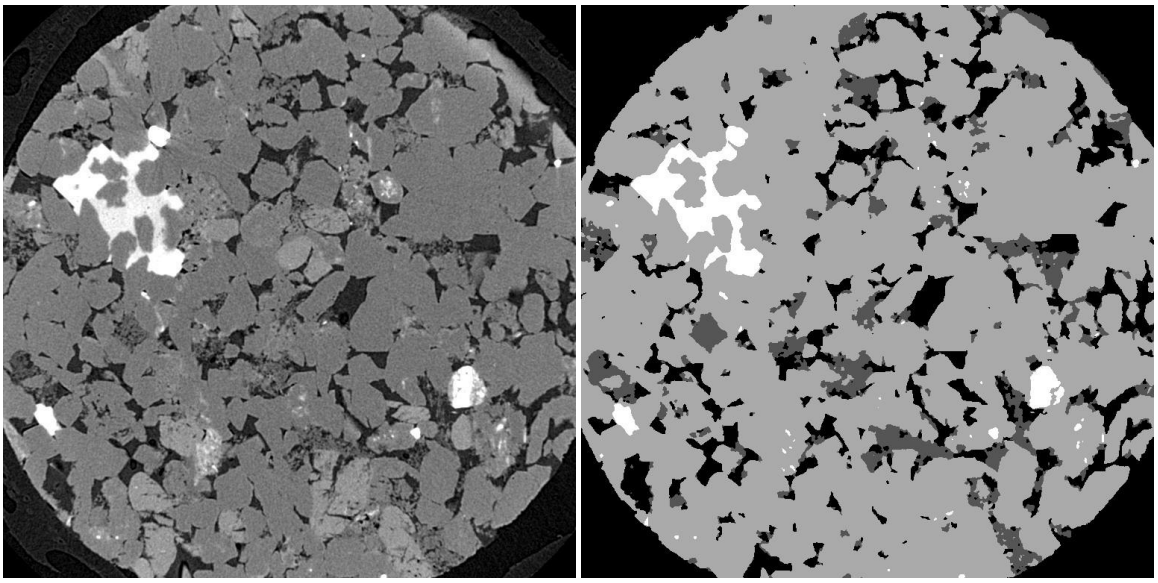
Description of the sandstone dataset

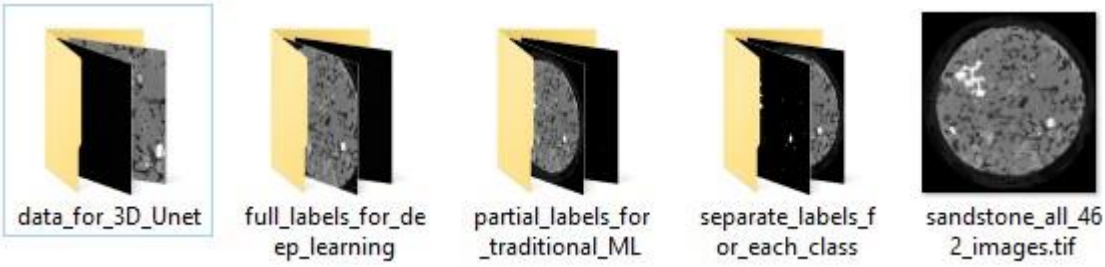


This data set can be used for semantic segmentation exercises. It represents a volume (tiff stack slice by slice) of 462 slices, each 996x1024 pixels. Voxel size is 2.2315 microns in each x, y, and z dimension, respectively. This means the pixel size to be 2.2315 microns in each 2D slice.

The volume represents the XRM (tomography) scan of a sandstone cylinder of size about 2 mm diameter. The region around the sandstone that shows up as dark pixels in the image is air. Also, the dark regions inside the sandstone represents air/void. In addition to air, the region has 3 regions of interest that need to be segmented, quartz (light grey), clay (darker grey with texture), and pyrite (bright pixels). Clay and quartz are minority classes but since quartz appears in bright, it is easier to segment. **Why do petrologists care about segmenting these types of images?**

Mineralogy is important to understand the value of a hydrocarbon reservoir. Lot of pore space is good as it potentially holds the hydrocarbons, quarts is good as it is brittle and easy to fracture, clay is bad as it does not fracture easily and also has potential to expand upon contact with water and clog the reservoir.

Description of Folders and Files:



Files: sandstone_all_462_images.tif - Tiff stack of all 462 images that need to be

segmented.

Folders:

separate_labels_for_each_class - Contains 9 training images as tiff stack and 4 other tiff stacks for labels corresponding to each class, air, clay, quartz, and pyrite, respectively. You can use these images to practice binary segmentation or for multiclass / instance segmentation that requires masks to be provided as separate labels.

partial_labels_for_traditional_ML - Contains 9 training images as tiff stacks and corresponding labeled image showing partial labels only. These labels were generated using www.apeer.com. This file can be used to practice semantic segmentation using Random Forest or XGBoost where we generate features from images and drop all pixels where we do not have any labels (value = 0). If the unlabeled pixels are not dropped the algorithm interprets background pixels (value=0) as a separate class. Please remember that deep learning expects full labels so you cannot use this dataset for that purpose.

full_labels_for_deep_learning - Contains 10 training images as tiff stack and corresponding fully annotated masks that can be used for deep learning. It also contains a subdirectory labeled '128_patches' containing tiff stacks of images and masks patched into 128x128 pixels. These patches can be directly used as inputs without the need for you to extract patches from large images.

Pixel value 1 = Background / Air
Pixel value 2 = Clay
Pixel value 3 = Quartz
Pixel value 4 = Pyrite

Remember to encode classes to pixel values 0, 1, 2, 3 in your code. This makes it easier to calculate IoU and other metrics using Keras.

data_for_3D_Unet - Contains a tiff stack with 448 images, each 512x512 in size. This is basically the original dataset that has been cropped into size that is divisible by 64 to make sure we do not run into any issues when working with 64x64x64 size patches. The folder also contains training images and masks representing a volume of 256x256x256 pixels. This dataset can be used to work with 3D Unet.

Additional information:

The labeled images will look dark when opened using regular image visualization programs such as Windows image viewer. This is because the maximum pixel value in masks is 4 and most programs assume full range (0-255) while displaying images. Therefore, to properly view the masks use imageJ and adjust brightness to 0-4. You can do this by navigating to Image-->Adjust-->Brightness/Contrast-->Set and changing Minimum to 0 and Maximum to 4.

To download imageJ: https://imagej.net/Fiji