Predict porosity from thin sections using DL

# Problem statement:

Thin sections are prepared for quantitative textual analysis of the rock samples to understand the mineral composition, fabric, and general makeup of a rock. This technique is used for study of grain size, grain shape and roundness, porosity, and permeability of the reservoir rocks.

A thin sliver is cut from the rock sample with a [diamond saw](https://en.wikipedia.org/wiki/Diamond_saw). Then they are cleaned in a vapor phase bath to remove solvable excess residual hydrocarbon. Next the samples are impregnated with blue epoxy to identify porosity and preserve textures, polished and mounted onto a glass slide. The samples are then ground down to a thickness of 30 microns and stained with a combined carbonate stain. Finally, a second glass slide is glued on the polished surfaces.

The prepared thin sections are examined petrographically. A polarizing microscope is used to take pictures of the thin section. When placed between two polarizing filters set at right angles to each other, the optical properties of the minerals in the thin section alter the color and intensity of the light as seen by the viewer. As different minerals have different optical properties, most rock forming minerals can be easily identified. Individual minerals are identified by their stained color and crystal structure (if visible).

To analyze a material’s composition, the technician identifies rock constituents at several “points” (~300-400) within a single thin section.  Petrographic analysis can be used to evaluate the pore system in a reservoir rock. These data are then used to calibrate 3d models of the subsurface and predict and estimate desired resources.

This project is aimed at developing a model that can predict porosity from the image of a thin section. Such a model can help to save on cost and human effort that is involved in thin section analysis.

# Description of the dataset

I’m using data from Volvo dataset. It is a published dataset for people to use for their research in oil and gas industry. The dataset has 26 wells. Three out of twenty-six wells have core data and two of these three have thin section analyses data (thin section images and corresponding porosity estimate). The thin section image data is available as pdf reports and must be wrangled and extracted to be used in this study. Each thin section has two photographs of the same thin section. There is a depth reference for the rock sample used to create the images. It is shown on the picture in-between the two images partially covering each of them. Picture below shows a picture for a single thin section taken in two different modes.

A spread sheet for conventional core analyses reports rock porosity for a particular depth. We can relate two datasets through the common variable – depth.

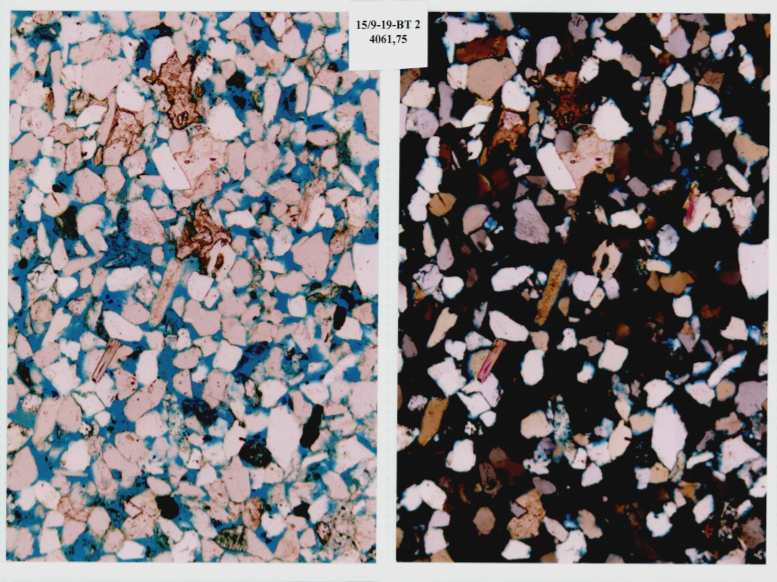


Figure: An example of a thin section photograph from a rock sample in well 15/9-19-BT 2 taken at depth 4061,75 m MD RKB which has the following rock properties: porosity – 23.7%

# Data wrangling

The first step is to extract thin section images from the pdf reports along with their associated depth references.

The image extraction from pdf with fitz package produces the result shown in Figure 2. We can see a pair of thin sections are show on the same image along with some background. The thin section photographs were presumably printed and placed in the sheet protector on some white background.

We first extract text from each of the images and search for depth related information using regular expressions. The process is somewhat successful, but still requires manual qc’s and filling missing depths. We save depth references for both thin sections as part of the image name when saving the images on the disk.

The next step is to extract individual images from each page. There are two thin sections on each pdf page with two different photograths (left and right). I call them top left, top right, base left, and base right. The location of the separations lines between individual images were used to crop each photograph from the page. The coordinates for the cropping are founds by summing the values of the pixels along horizontal or vertical directions. White separation lines between the photographs will have lower sums than lines going through the photos. Searching within a particular region for min or max of the low sum clusters will give the coordinates of the photos. Example of the extraction of the image coordinates for a page shown in Figure 2 is found in Figure 3.

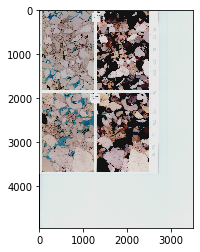
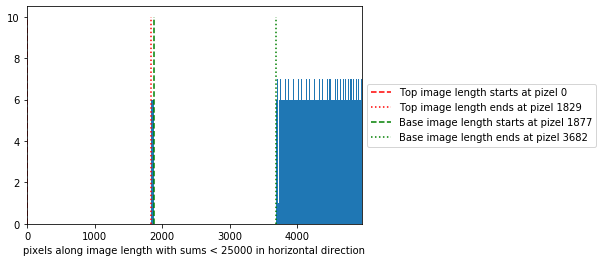


Figure 2: Image extraction from pdf report using fitz package and saved under name 15\_9\_19\_A\_p2-78-3837.55\_3838.50. The name for this image is a combination of well name (15\_9\_19\_A), page number (p2) in the pdf report, image id (78), and the depths shown for each thin section in the image (3837.55 for the top thin section, 3838.50 for the base thin section).



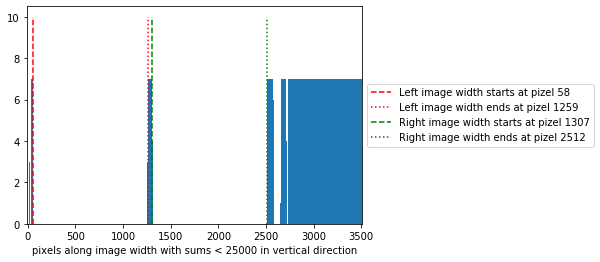
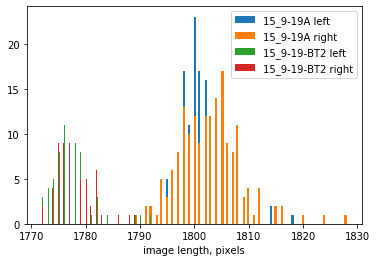


Figure 3: Vertical lines are showing the cropping coordinates for extracting four images from the Figure 2. Top figure is for extraction along the length of the page to separate top images from the base ones. The lower figure is for extraction along the width of the page to separate left images from right ones.



Figure 3: Example of the image extraction for images shown in Figure 2 (15\_9\_19\_A\_p2-78-3837.55\_3838.50). Four images are extracted: top left and top right are the photographs of the same thin sections at depth 3837.55 m MD RKB; base left and base right are the photographs of the same thin sections at depth 3838.50 m MD RKB

****

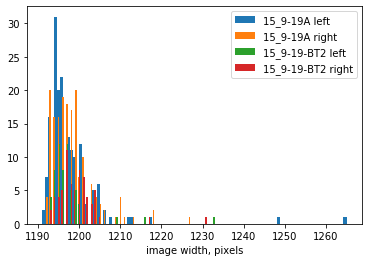
****

Figure 4. Histogram of image depths and widths for all images extracted for well 15\_9-19A (blue) and well 15\_9-19TB2 (orange).

The process of image extractions results with the images of different sizes. Figure 4 shows the histograms of image lengths and widths for all the images extracted from the reports. The widths of the images have a narrow distribution across two wells and both left and right photographs. The lengths have a bimodal distribution. Both thin section photos from the 15\_9-19TB2 are shorter than those from 15\_9-19A wells.

In order to use all the data in our analyses images were resized to the standard shape of 1024 by 1536.



Figure 5: Example of the image resizing to the standard shape of 1024 by 1536 for images shown in Figure 3 (15\_9\_19\_A\_p2-78-3837.55\_3838.50).

The final step in the data preparation is to find image labels (porosity values). As mentioned above porosities are reported in the excel spreadsheets for each depth. Porosity and corresponding depths are extracted from the core analyses spreadsheet into a data frame. Filenames are then added as a separate column by depth key. The resulting data frame contains porosity, depth values and filename for the thin section images.

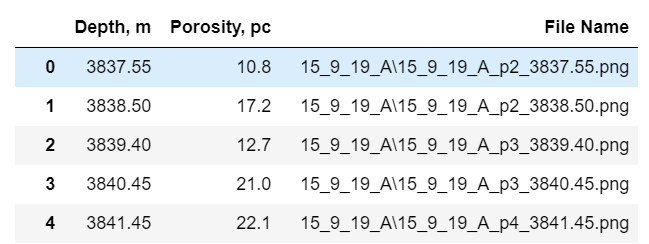


Figure 6. A snapshot of a final dataframe with all the information required for training a DL model. Labels are stored in “Porosity, pc” column. Input are images saves under file names in the “File Name” column.