scrAiber:

Data Mining Driven Microscopic Reference Data Acquisition

M. Sc. Artem Leichter

Institute of Cartography and Geoinformatics
Leibniz University Hannover

Dr. Renat Almeev
Prof. Dr. rer. nat. Francois Holtz

Institut of Mineralogy
Leibniz University Hannover

Date of submission 13.05.2022

Required funding (**5 months**)

Abstract

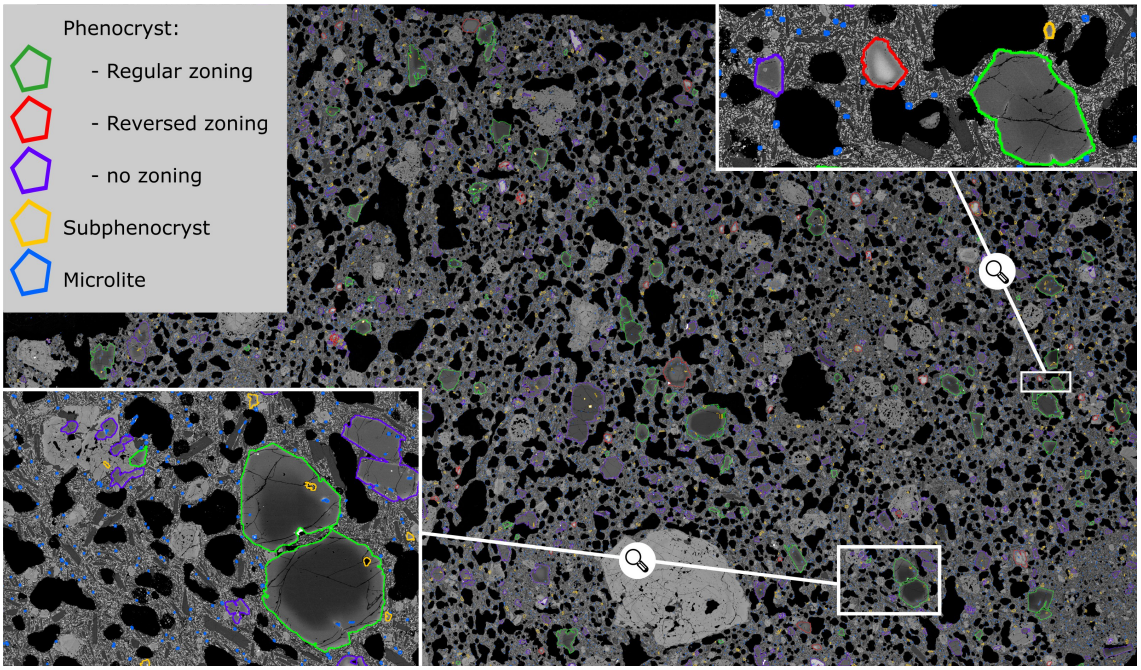
Creating training datasets for machine learning (ML) applications is always time consuming and costly. In domains where a high degree of expertise is required to generate the reference data, the corresponding costs are high and thus slow down the use of artificial intelligence (AI) systems. This proposal focusses on automated mineralogy and will provide tools to characterize the microscopic textural and mineralogical features of thin sections of rocks using back scattered electron images. Our goal is to address this problem with a data mining application where unsupervised methods in combination with expert users generate reference data without additional effort and cost for explicit labeling. The tools will be developed so that it can be used by scientists that have not a profound knowledge of ML.

I. *Introduction*

Availability of data and reference data is crucial for the success of ML algorithms. Creating reference data is costly and time consuming, especially for the characterization of rock samples in Earth sciences. The mineralogy and textures of rock samples can be characterized by a variety of analytical techniques, and detailed investigations are mostly based on the microscopic analysis of thin sections. The most common, fast and high resolution technique used to characterize thin sections is based on the acquisition of back scattered electron (BSE) images using electron microscopy. Such images can be used to extract information for an extremely large range of applications in mineralogy. However, the quantitative analysis of rough data requires experts, whose time is costly and limited. To address this problem, we propose a workflow combining unsupervised ML methods with expert input. Our prototype implementation of the approach will focus on the (unsupervised) *segmentation* of BSE images acquired on modern scanning electron microscopes with subsequent *clustering and classification* of objects of interest (e.g. minerals).
Unsupervised machine learning (ML) methods extract information without the need for reference data, although it is always advantageous to have reference data to validate the models or at least identify the suitable metaparameter. For example, a common strategy for creating reference data is to use unsupervised ML methods in the first step and then to refine the results through user

interaction. Such approach requires high expertise of the user. We propose to prototype a digital environment that provides productive work based on unsupervised ML methods so that users interact with it on their own initiative while working on their own tasks. In this way, there is no additional workforce investment necessary. The reference data is created implicitly by tracking the user interaction. This approach is state of the art in many commercial applications such as online stores and video platforms, but as far as we know it is novel in the context of digital petrological analysis.

From our preliminary work based on the analysis of BSE image data, we have a number of implemented algorithms that yielded extremely promising results but that can only be carried out by users with strong programming skills (Leichter et al., 2022). Based on our experience gained so far and in the context of this project, we plan to construct a platform which can be used for image analysis by non-experienced user.



1. Examples of successful deep learning (DL) segmentation and rule based classification applied on a thin section of a volcanic rock (Leichter et al., 2022). In this example, ML was used to identify automatically all olivine crystals present in the thin section (more than 20.000) and to classify the minerals depending on the compositional zoning as well as the type of minerals (phenocrysts, subphenocrysts, microlites). Phenocrysts with compositional zoning (ca. 800 crystals) were subsequently used to apply diffusion chronometry. Without application of ML learning techniques, it is impossible to treat manually (and objectively) this amount of data. For more examples visit icaml.org/olmap/.

II. *Incubator Project description*

This project consists of two main packages, namely (1) implementation of a productive online tool for analyzing BSE data and (2) acquisition of user tracking data and its prototypic evaluation.
(1) Implementation of a productive online tool (3 Month): The online tool, hereafter referred to as "scrAiber", is crucial for the possibility of collecting data on segmentation. For this purpose, scrAiber should allow the user to solve specific problems (e.g. semantic segmentation of minerals, clustering of crystals, unsupervised pixelwise segmentation) when working with BSE images. In order to meet the compact framework of this project, the system is implemented as a web application, which facilitates the management of dependencies and increases availability for users. The implemented functionality is based on already existing solutions like preprocessing of BSE raw data, segmentation of minerals with deep learning (DL) framework or unsupervised segmentation. The user interface (UI) is implemented in a minimalistic way with focus on functionality and a low entry barrier (immediate start of work).
(2) Collecting user tracking data (2 Month): User tracking data allows applications such as online stores to identify similar and complementary items based on user activity. Actions such as completed purchase of a product and joint purchase of multiple products serve as indicators for the analysis. In this project, such indicators must first be identified and integrated into the scrAiber workflow. In this step, mineralogy expert(s) actively work with scrAiber. The outcome of their work session is documented in short interviews and questionnaires. The gathered data is used to prototypical evaluation of the tracking data.
For the processing of the two packages, partner institutions bringing experts in Earth Sciences (mainly petrology/mineralogy) and geoinformatics (data engineering competence) together are required. The processing time of the first package is estimated to be three months and consists of two man-months work for an expert in Data Engineering (Institute of Cartography and Geoinformatics) and one month for a petrologist (Institut of Mineralogie) . For the second package, the working time is composed of one man-month work for each partner.

III. *Relevance for the NFDI4Earth*

Machine learning opens up unique opportunities for the analysis of *electron microscopy image data* (this proposal) but also for a variety of analytical sensors in Earth sciences in general. To be able to use these possibilities, remote data are mandatory. The strategy proposed here can also be applied to other domains to acquire reference data with little additional effort.
The main users of scrAiber are expected to be petrologists investigating and characterizing rock systems. The application of scrAiber is expected to be helpful in applied geosciences and for mineral processing engieneers, allowing a fast characterization of a series of rocks such as drill cores (e.g. drill cores used for the characterization of ore deposits) as well as for other more fundamental scientific projects based on mineralogical analyses (e.g., structural geology, volcanology, metamorphic and magmatic rocks). The tool will also be extremely useful for teaching courses in Earth Sciences and can also be used to introduce the advantages of ML.

So far, to our knowledge, there is no structured open access database available for BSE images of thin sections. Thus, at this stage, our proposal will not establish a tool that can be directly applied to available databases. However, the tool developed in this proposal, allowing users to extract extremely quickly a huge amount of information from BSE images, is expected to initiate and promote the creation of databases for the microscopic analysis of thin sections.
The main repositories that can be used are rocks/thin sections from drill cores or field expeditions collected in the frame of several national research initiatives or international expeditions. These rocks are stored in various places in Germany. The access to data and to rocks from national drilling cores, stored at repositories under the responsibility of the Bundesanstalt für Geowissenschaften and Rohstoffe (BGR) is possible in the frame of the cooperation of the Leibniz University of Hannover and the BGR. Cores and thin sections that have been characterized in detail can also be obtained by contacting the ICDP and IODP coordinators in Germany (GFZ Potsdam and BGR, respectively). The petrologists of the Institut of Mineralogy have also numerous cooperations which would allow a rapid access to numerous samples and BSE data from thin sections.

IV. *Deliverables*

1. scrAiber: A tool for segmenting SEM data that is accessible to all via the internet.
2. scrAiber Code: opensource project for the scrAiber online tool, wich allows the adoption to different datatyper and tasks.
3.The data gathered by scrAiber will be published as open data and can be integrated to NFDI4Earth repositories.

V. *References*

Leichter, A., Almeev, R. R., Wittich, D., Beckmann, P., Rottensteiner, F., Holtz, F., & Sester, M. (2022). Automated Segmentation of Olivine Phenocrysts in a Volcanic Rock Thin Section Using a Fully Convolutional Neural Network. Front. Earth Sci, 10, 740638. | https://doi.org/10.3389/feart.2022.740638