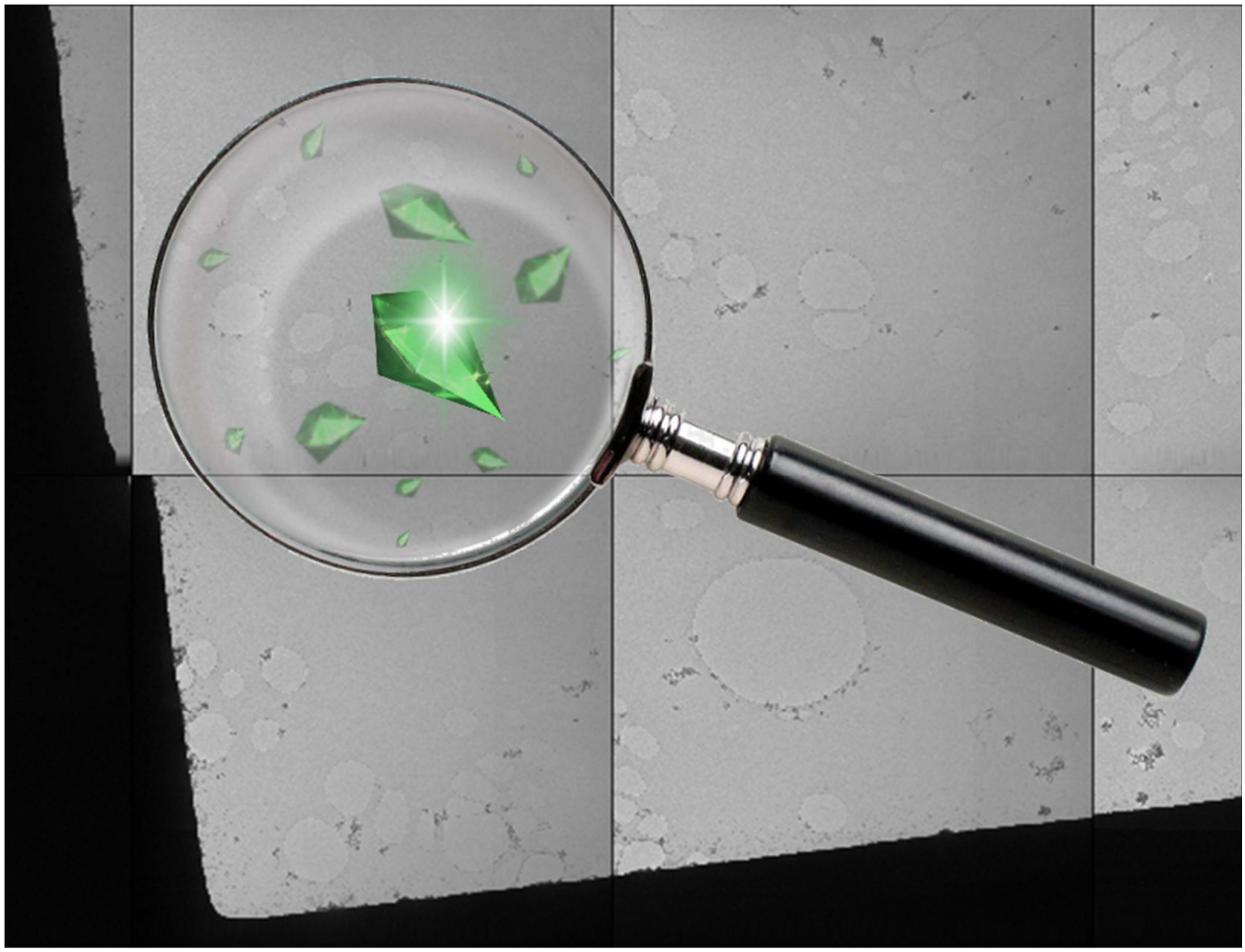
JADS Data Expert Program 2018

The Micro-Crystal Locator

Finding needles in the haystack of electron microscopy



Michael Janus

11 December 2018

JADS Jheronimus Academy of Data Science

ABSTRACT
The work presented in this document reports on a data science based technique to detect particles in images from electron microscopes. With this technique, particles are located in a smarter and faster way than the current, mainly visual detection methods, saving up to 90% search time.

The work was conducted as a final assignment in the context of the Data Expert Program at the Jheronimus Academy of Data Science. The goal of this assignment was three-fold: to learn how to apply data science techniques, to develop entrepreneur skills and to contribute to a solution for a real-life problem.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 WHAT IS ELECTRON MICROSCOPY?

Electron microscopy (EM) has been a revolutionary technique for past 80 years to obtain high resolution images at magnifications far beyond the capabilities of light microscopes. EM enables scientists and engineers to study microscopic structures in fine detail, opening up the world of nanoscale materials and allowing biologists to study the building blocks of life at the scale of single molecules an atoms.

Modern electron microscopes are mostly computer-controlled and their capabilities are not limited to 2D images. With a plethora of detectors, advanced acquisition techniques and dedicated software, additional properties of the structure at study can be characterized and revealed in 3D. Electron microscopes are highly complex pieces of equipment and require skilled personnel to operate them.

Figure 1 Different types of electron microscopes. From left-to-right: a mid-range TEM, a high-end TEM, older type of TEM (still used a lot) and a SEM.

1.2 CUSTOMER SEGMENTS

Though electron microscopy is a very specialized field not well known to the public, electron microscopy is a multibillion dollar market and applied in a wide range of disciplines.

Various marketing reports [1, 2] segment the EM market in one of the following ways:

a) by application field (Life Sciences, Material Sciences, Semiconductors, Nanotechnology)
b) by main instrument type (Scanning Electron Microscopy, Transmission Electron Microscopy), which is subdivided in high-, mid- and low-end instruments
c) by end-user type (industries, academic institutions, and others)
d) by method of sample preparation (cryo-genic samples, chemically fixed samples, ..)

In reality, the world is not that black-and-white and whichever segmentation you choose, many customers do not fall into just one category. From a development perspective, we have to recognize that even high-end electron microscopy is gradually transitioning from an experts-only field to more mainstream. This implies that the level of automation, service and ease-of-use has to go up (see e.g. Crossing the Chasm [3]).

1.3 Life Sciences : Determining The Structure of Proteins

A rapid evolving field in Life Sciences is structural biology. Researchers use a plethora of techniques to unravel the details of cells down to the level of molecules and atoms, in order to understand the machinery of life and develop new medicines.

Proteins play a critical role in the processes that take place in living organisms. They are large, complex molecules and understanding their three-dimensional structure is key to infer how they function. Protein structure determination is actually one of the most important fields of modern biology and finds application in fields like drug discovery and design.

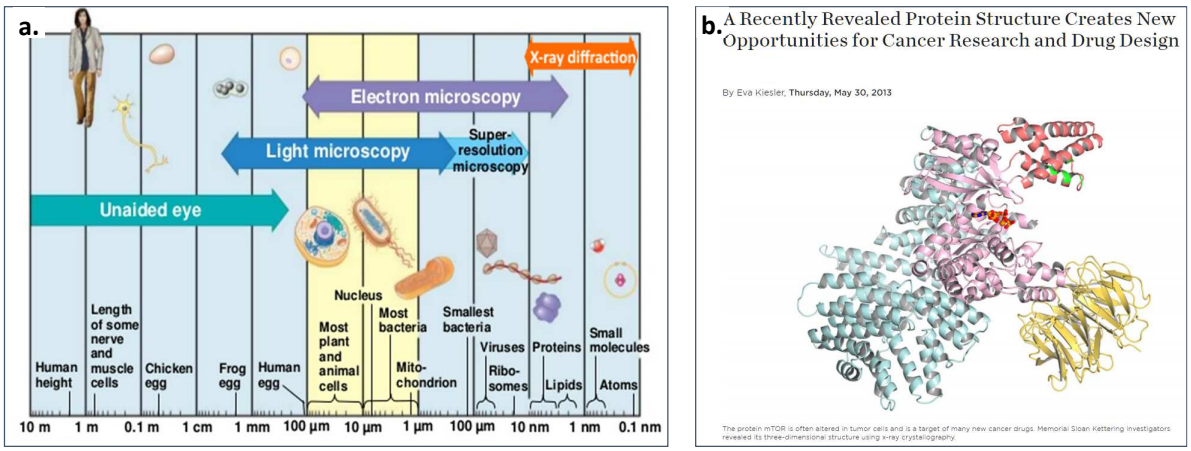


Figure 2 **a.)** A chart of the relevant structures in life sciences and their associated length scales plus an indication which technique is used for the research. **b.)** An example of a protein molecule and its impacts medicine.

1.4 Structure determination with *X-Ray crystallography*

X-ray crystallography is the dominant technique for determining the atomic and molecular structure of biological macromolecules like proteins. Probably the most famous structure determined by this technique is DNA. Modern *x-ray crystallography* is conducted with large synchrotron set-ups that have the size of multiple football fields.

In x-ray crystallography, a specimen is exposed to high energy x-rays and its scattered signal is collected. The specimen has to be crystalline, which requires the proteins to be grown into large, high-quality crystals, a complicated and delicate preparation process. Through advanced computer algorithms, the collected signals are processed to resolve the atomic structure in 3D.

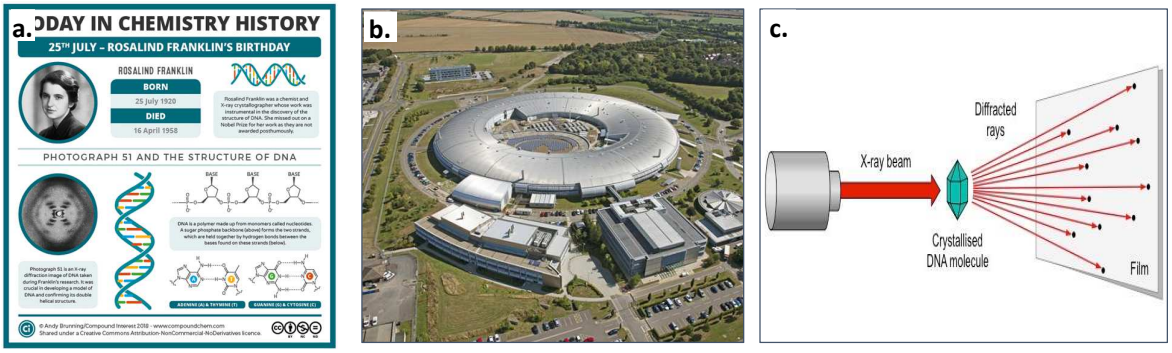


Figure 3 **a.)** One of the most famous break-throughs with x-ray crystallography was the determination the structure of DNA. **b.)** Modern x-ray crystallography is conducted in large synchrotron facilities (picture of Diamond Light Source, United Kingdom). **c.)** The principle of x-ray crystallography has not changed: a crystallized sample of molecules is radiated with x-rays and the scattered x-rays are recorded; the diffraction pattern is sort of a footprint of the atomic structure in the molecule.

1.5 CRYSTALLOGRAPHY WITH ELECTRONS: *MICROED*

Micro-Electron Diffraction, or *MicroED*, is an emerging technique in electron microscopy to resolve the 3D molecular structure of molecules similar to *x-ray crystallography*, but then using electrons instead of x-rays.

This technique has caught a lot of attention lately as electron microscopes are less costly to operate than the large synchrotron facilities. Moreover, with *MicroED* it is possible to determine the molecular structure from much smaller crystals than are needed for *x-ray crystallography*; micron- or submicron-sized crystals are sufficient. Such small crystals are much easier to make.

In *MicroED*, the electron beam is accurately positioned onto a tiny crystal, where electron diffraction patterns are collected from multiple angles (e.g. by rotating the specimen). From these diffraction patterns, advanced computer algorithms can reconstruct the structure of the protein in 3D down to atomic resolution.

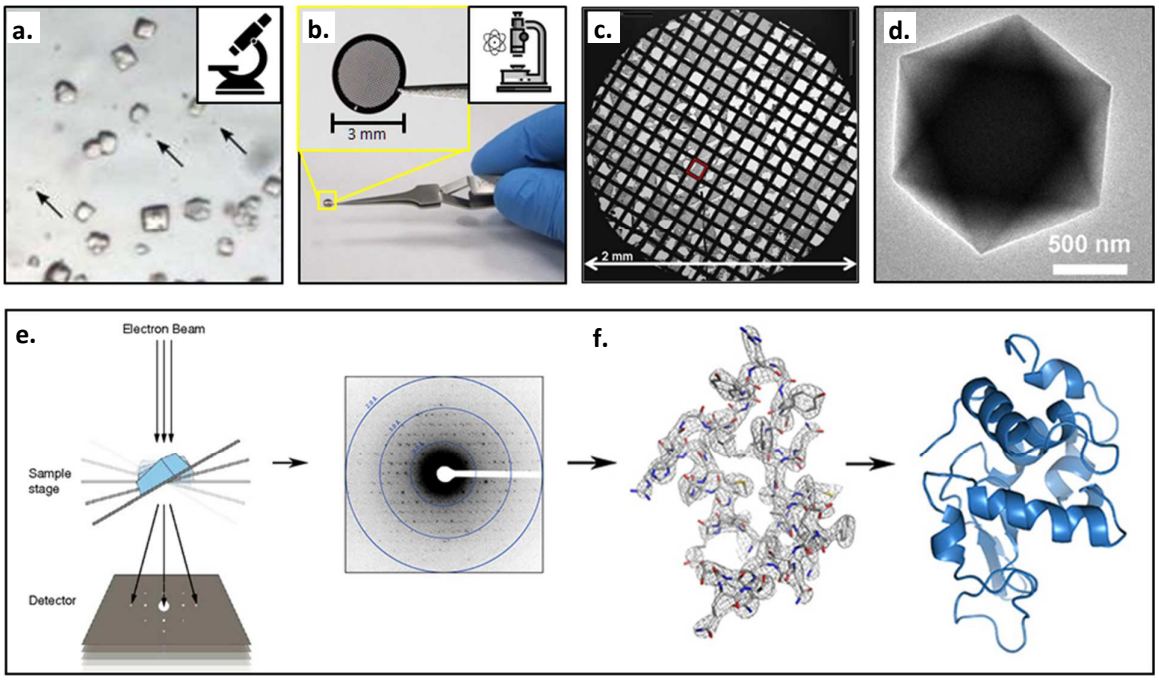


Figure 4 The MicroED Workflow. **a.)** With a light microscope, suitable micro-crystal candidates are identified; **b.)** the micro-crystals are deposited on a *TEM grid* and the specimen is transferred in to the electron microscope; **c.)** the specimen is screened at low magnification to locate the micro-crystals; **d.)** the electron beam is accurately positioned on to a micro-crystal **e.)** the *MicroED* data collection is executed; **f.)** the collected diffraction images are post-processed and the 3D molecular structure is reconstructed with sophisticated software.

1.6 What are Micro-Crystals?

With the proper preparation technique, protein molecules arrange themselves into an ordered pattern, forming a crystal. However, growing large crystals - as needed for x-ray crystallography - is a challenge. Smaller crystals are more easily obtained and when these crystals are micron-sized, they are called **micro-crystals,** which are suitable for *MicroED* experiments. If the crystals are much smaller than a micron, they are also referred to as *nanocrystals*.

Micro-crystals come in varying shapes and sizes, depending on its constituents and the preparation technique. On a specimen, the micro-crystals are typically randomly distributed, hence, they first need to be pinpointed before *MicroED* data can be collected.

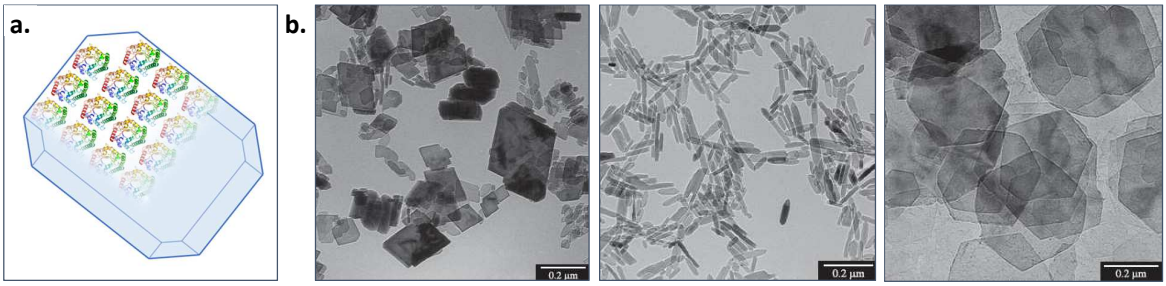


Figure 5 **a.)** Schematic drawing of how protein molecules line up into small crystals. **b.)** Micro-crystals com in varying shapes and sizes as illustrated by the electron microscopy images of three different types of micro-crystals

1.7 TERMINOLOGY

Terminology from the domain of electron microscopy as used throughout this report are listed below.

Table 1 Glossary for Electron Microscopy

Å *Ångström*	A unit of length equal to one ten-millionth of a millimeter (i.e. a 10^{-10} meter). This is roughly the size of an atom.
Acquisition	See Image Acquisition
EM	Abbreviation for Electron Microscopy
FOV, *Field-of-View*	The area that the microscope is currently looking at. This is a combination of the magnification and the location on the specimen.
LM, *Low Magnification*	Refers to the magnification range of roughly 10x to 1000x .
HM, *High Magnification*	Refers to magnifications far above 1000x and can go up to more than million times.
Image Acquisition	The process of collecting images on an electron microscope
Mapping	The process of acquiring images at regular array of positions, in order to image an area larger than the field-of-view. The result look very similar to a map with a grid.
MED, MicroED	Micro-Electron Diffraction, an emerging technique in electron microscopy to resolve molecular structures in 3D. The technique is akin to x-ray crystallography, but then using electrons instead of x-rays.
Micro-Crystal	A small crystal, about a few microns in size
Micron / Micrometer	A unit of length equal to one thousandth of a millimeter (i.e. 10^{-6} meter).
Nanometer	A unit of length equal to one millionth of a millimeter (i.e. 10^{-9} meter).
SA	Originally abbreviation for 'Selective Area', but used to indicate a specific high magnification mode in TEM. Often used in a context to indicate the range in-between Low and (very) High Magnifications.
Sample	Another word for *specimen*
Sample Prep	The (delegate) activity or process of preparing a *specimen* for electron microscopy. Sample preparation is a field on its own and consists of a plethora of techniques, including cryogenic freezing, chemical fixation, staining and many more.
Screening	Refers to evaluating a number of *specimens* (or parts of a single specimen) for suitability of the more involved experiments. This usually involves scanning the specimen(s) quickly at a low magnification.
SEM	A **S**canning **E**lectron **M**icroscope acquires its images by scanning the electron beam over the sample in a way similar to old fashioned television tubes.
Specimen	A specimen is a small sample of something, that is taken and prepared for studying in a (electron) microscope.
TEM	A **T**ransmission **E**lectron **M**icroscope acquires its images by 'shining through' the specimen like a slide projector. TEM's are the most powerful electron microscopes, able to resolve structures down to atom level, but specimens need to be very thin and often require special preparation.
Tomography	A technique where images are acquired section by section, e.g. by rotating the specimen in-between acquisitions, and combining the images with advanced algorithms into a 3D volume. More commonly known examples of this technique MRI and CT in medical imaging.
X-Ray Crystallography	One of the most important techniques for scientists to determine the atomic and molecular structure of matter. DNA is probably the most famous structure determined by X-ray crystallography. It is based on high energy x-rays and the *diffracting* properties of crystals. Modern x-ray crystallography is conducted with large synchrotron set-ups with the size of multiple football fields!

2 THE IDEA

2.1 THE PROBLEM: FINDING THE NEEDLE IN A HAYSTACK

As data collection is a time-consuming and expensive activity in electron microscopy, the amount of data acquisition should be limited to only the relevant parts of a specimen. Hence, most experiments start by acquiring overview images at relatively low magnifications to identify potential areas of interest. In consequent steps, the microscope operator gradually zooms into the areas of interest until the relevant microscopic structures are found. Only there the actual, high-quality data collection schemes are executed.

This process of finding the relevant areas can be daunting task for non-expert users, but even for an experienced microscope operators this is a time-consuming activity and a tedious exercise. As such, more automation and user guidance for this process are highly desired.

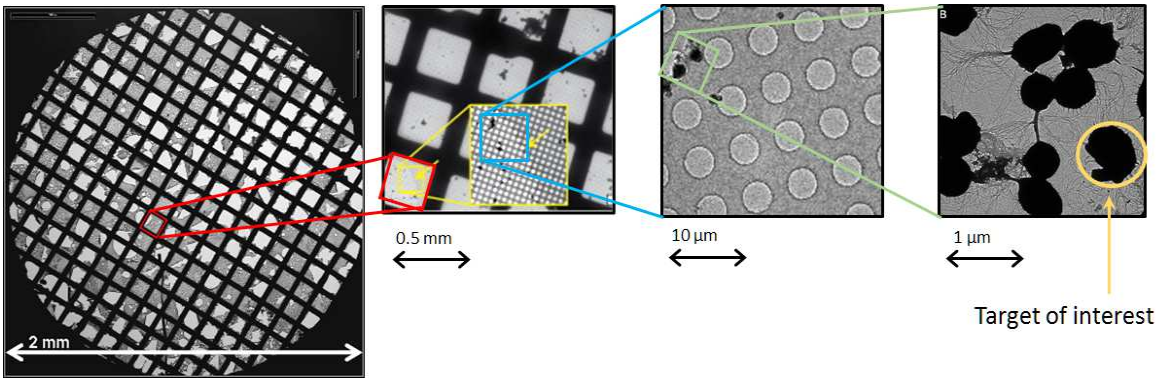


Figure 6 The process of gradually zooming in on the specimen to find the relevant structure to study.

2.2 EXISTING SOLUTIONS

For a number of use cases, conventional image recognition is used to identify the areas-of-interest. However, this approach only works well if the characteristics of the specimen are well known upfront and little variation is to be expected. In addition, such techniques are tuned by experts to the problem at hand, which is only viable for repetitive experiments. An example of such use case is wafer inspection in chip industry. For experiments with much greater variation in specimens, these techniques are not suitable and a more adaptive technique is needed.

2.3 THE CASE OF 'MICRO-CRYSTALS'

One use case that requires more flexibility is the detection of so called micro-crystals for *MicroED* experiments, a novel technique in electron microscopy to resolve molecular structures in 3D (see paragraphs 1.5 and 1.6). This technique requires the microscope to be pinpointed exactly at the very small crystals. As these micro-crystals vary a lot in shape and size, even within a single specimen, conventional image recognition is not suitable and a more adaptive technique is required.

2.4 PROPOSED SOLUTION: THE MICRO-CRYSTAL LOCATOR

Use modern and adaptive machine learning techniques to automate the process of finding the micro-crystals in a faster, smarter and more reliable way. The concept is explained in more detail in chapter 4.

3 BUSINESS IMPACT

3.1 BUSINESS CASE

The emerging technique of *MicroED* is mentioned in a number of scientific publications as an alternative or complementary technique to *X-ray crystallography* (see e.g. Nannenga *et al* [3] and Nicolopoulos [5]). Its potential and importance for the pharmaceutical industry is recognized in various business reports on 'protein crystallization and crystallography' [4, 6], a growth market which annually exceeds one billion dollar.

Ease-of-use, automation and throughput are mentioned as hurdles for wide scale adoption of this technique. The automatic detection of micro-crystals contributes to all three[1].

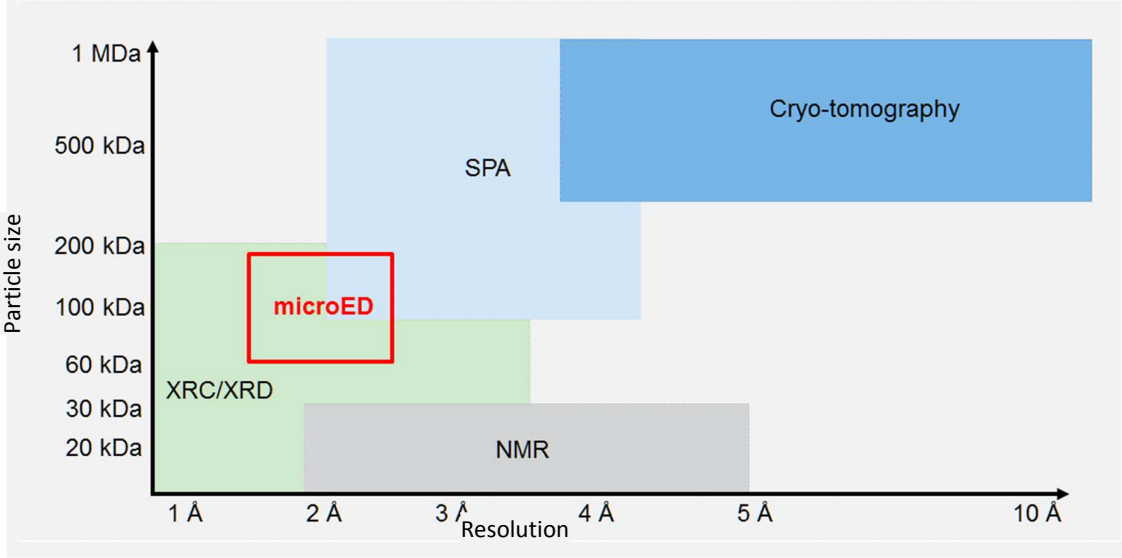


Figure 7 A plot of the dominant imaging techniques for characterizing protein molecules and their applicability. The y-axis indicates the size of the molecules that can be studied with the technique, the x-axis shows the resolving power expressed in Angstrom. Graph courtesy to Lingbo Yu.

3.2 ADDITIONAL POTENTIAL

As mentioned in paragraph 2.1, many experiments in electron microscopy start by pinpointing the relevant locations. Hence, automated detection is beneficial for many experiments or workflows that require batch acquisition. As such, an adaptive detection technique may also be applied to other use cases.

[1] In electron microscopy, software automation packages are sold mostly in combination with the equipment, enabling more advanced use cases or functionality. Making a profitable software business independently has proven to be difficult due to the small number of electron microscope owners and the price cap that exists on software at many research facilities.

3.3 CUSTOMER VALUE

From a customer perspective, automated detection reduces the time needed at the microscope and relieves the operator from a tedious and repetitive task. For small numbers of micro-crystals, manual searching is acceptable, but for large numbers this is not viable.

The automatic detection of micro-crystals is not the only part of the *MicroED* workflow, other steps in the workflow need to be automated first. An estimate of the reduction experiment time for different levels of automation is listed in Table 2.

Table 2 The benefits of automating the MED workflow for increasing number of micro-crystals

	1 µ-crystal		**10 µ-crystals**		**100 µ-crystals**	
	time	**costs***	**time**	**costs***	**time**	**costs***
Manual Workflow, base line***	30 min	$50	5.0 hr	$ 500	50 hr **	$ 5000
Expected time improvements:						
Manual Crystal Selection + Automated Data Collection	10 min	$16	1.5 hr	$ 166	15 hr	$ 1666
Automated Crystal Detection + Automated Data Collection	2 min	$ 3	20 min.	$ 33	3.3 hr	$ 333

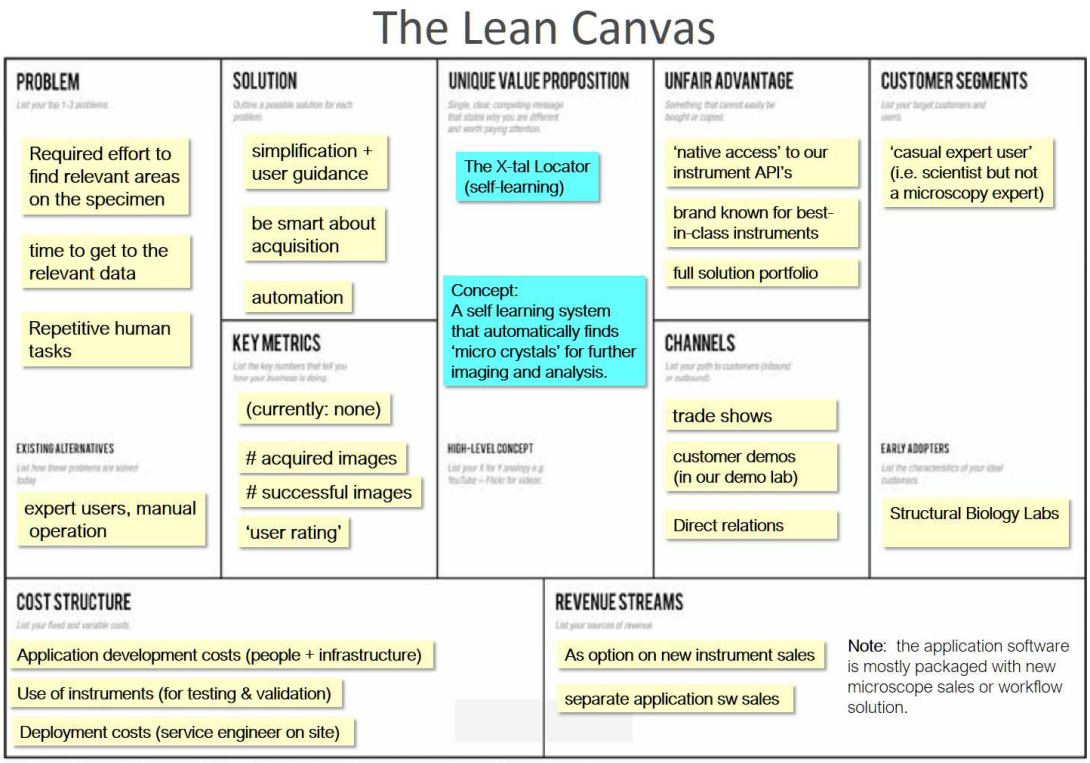
* Costs based on $ 100 per hour for a skilled microscope operator
** Large numbers of micro crystals in a manual workflow is hardly viable
*** The base line was determined at a customer location

Note that reduced experiment time is not only about labor costs, but also about microscope utilization and time-to-data

3.4 The Business Model

The business model for micro-crystal detection is summarized in the business canvasses below.

a.



b.

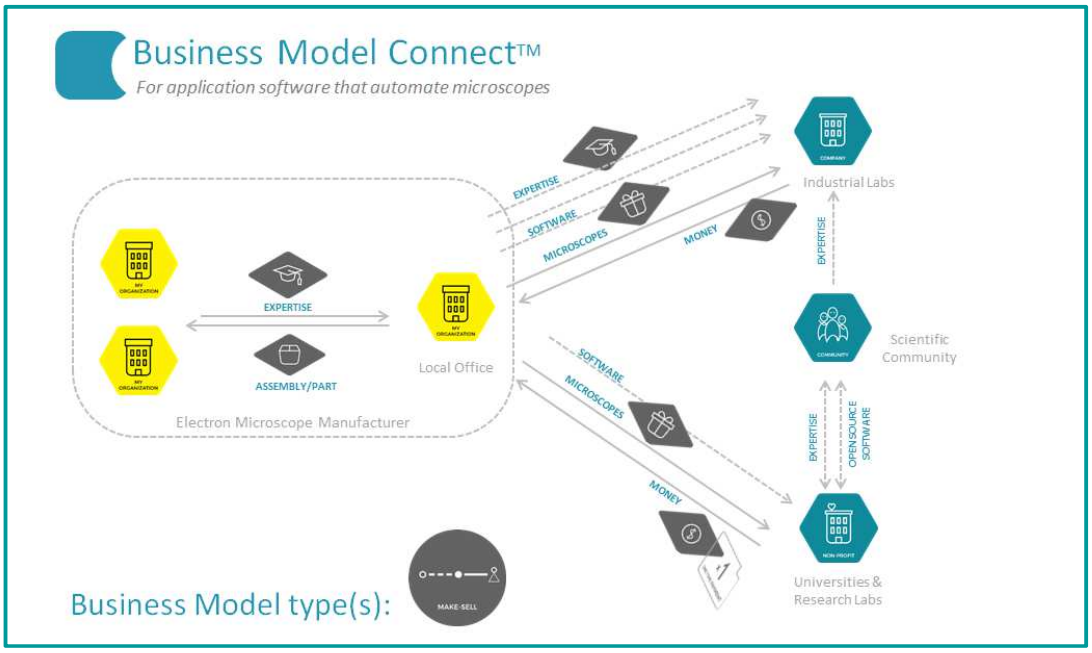


Figure 8 The Lean Canvas **(a)** and Business Model Connect **(b)** summarize the business model that applies to micro-crystal detection and automation software for electron microscopes.

4 THE SOLUTION EXPLAINED

4.1 THE CONCEPT

The solution presented in this report is based on sub-dividing the microscope images into smaller regions and using regular image statistics[2] to assess if these regions contain a *micro-crystal* or not. Based on this assessment, each region can be assigned a color, which can be plotted on top of the original image as a heatmap. The concept is schematically depicted in Figure 9

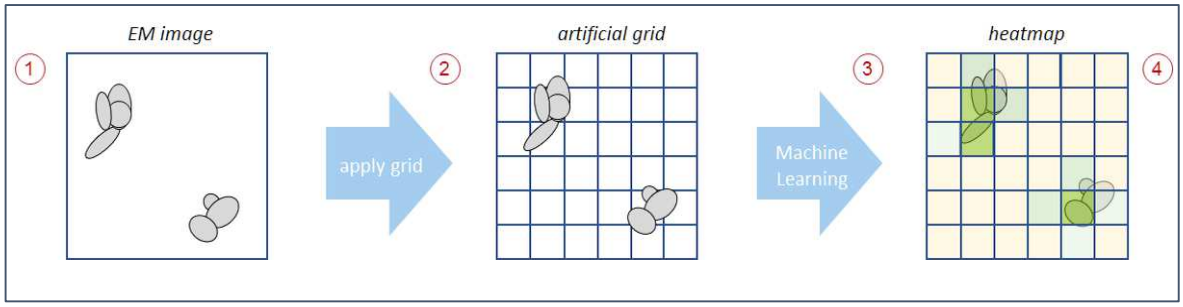


Figure 9 Schematic representation of the crystal detection. First, an electron microscopy image is recorded (1). Then, an artificial grid is applied on the image and for each cell of the artificial grid, image statistics are extracted (2); these image statistics are fed to an unsupervised learning algorithm to cluster the grid cells into groups (3); each group is assigned a color and plotted as an overlay onto the original image (4). Finally, the user should indicate the cluster of interest (not shown).

4.2 UNSUPERVISED MACHINE LEARNING

Which specific image statistic will reveal the presence of the particles or not, is highly dependent on how the micro-crystal images look and will vary from specimen to specimen. This is where machine learning comes in, using the most discriminating statistics from the data at hand to divide the sub-regions into common looking groups (also referred to as *clusters*).

As the data is not *labelled* upfront, it is not yet known which of the clusters contains the particles; a user interaction is required to indicate the relevant group. However, knowledge from prior experiments with similar images can be re-used, limiting the amount of user interaction needed. Alternatively, a rule-based approach could be used for that assessment.

4.3 WHY NOT DEEP LEARNING?

In the last years, many advancements have been made for image recognition with so called *deep learning* networks and one could argue that those techniques are a better fit to this problem. However, deep learning requires a lot of *labelled* data to train, which is currently not available and most customers are very protective on their data. For such scenarios, *machine learning pipelines* outperform deep learning networks (see Appendix A).

In addition, the current computers that control the microscopes are not equipped for training deep networks and as they are disconnected from the internet (for protective reasons), using cloud services for the learning is not an option either at this point in time. However, this is changing and in a few years' time from now, indeed deep learning networks may be a viable solution.

[2] Image statistics are based on the pixel values in the image, for example the average gray value, its standard deviation, the *kurtosis* and *skewness* of the distribution of gray values, etc.

5 THE DATA SETS

The data for the work presented in this report consists of images that were acquired with an electron microscope, at various magnifications and on different test samples containing micro-crystals. An example of such a data set is shown in Figure 10. Note the huge difference in scale between the low and high magnification images. Micro-crystals exist in many shapes and sizes; three types are shown in Figure 11, but many more exists. The full data sets are listed in Appendix B .

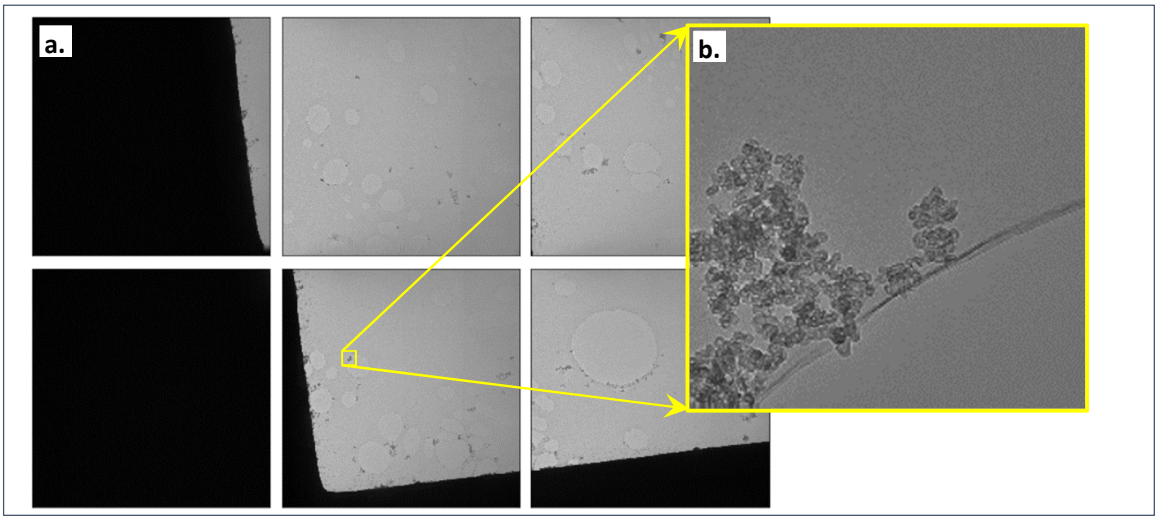


Figure 10 An example data set of one type of micro-crystals. **a.)** A (relatively) low magnification overview of part of the sample (obtained by mapping); each individual image covers 18 microns. At this magnification level, the micro-crystals are only visible as small dark dots. **b.)** A high magnification image of a micro-crystal, spawning a field-of-view of approximately 1 micron.

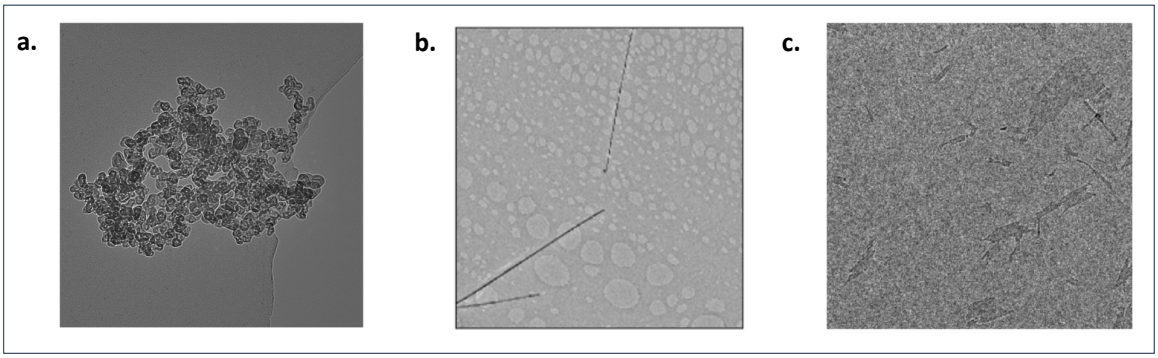


Figure 11 Examples of different micro-crystals, with a large variety in shape, size and texture. **a.)** round and 'blobby' graphite particles; **b.)** elongated and rod-shaped asbestos fibers; **c.)** hardly visible polymer crystals with a 'patchy' structure.

REMARKS

- The images are typically 2k x 2k pixels in size and of 16 bit gray scale format (8 MB per image); the number of images in the analyzed data sets varies from 4 to 16
- The grayscale values can vary from image to image due to different lighting conditions and pre-processing to maximize dynamic range.

6 THE MACHINE LEARNING PIPELINE

6.1 OVERVIEW OF THE ML PIPELINE

The figure below depicts schematically the set-up of the machine learning pipeline that has been implemented in the prototype. The individual steps are described in more detail in next paragraphs.

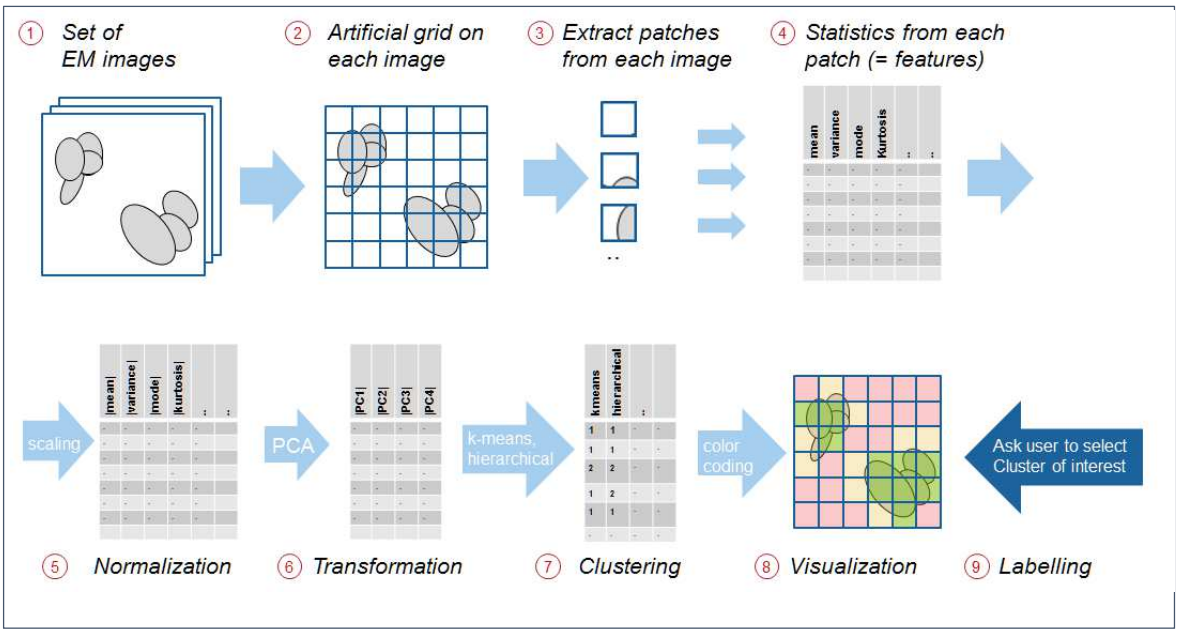


Figure 12 Schematic representation of the full machine learning pipeline that has been implemented in the prototype.

6.2 STEP-BY-STEP DESCRIPTION

1. Set of EM images

The input of the pipeline is a set of images on disk, originating from an MED experiment (see paragraph on Data). The images are in 16 bit grayscale format with a resolution of 2048x2048 or 4096x4096 pixels. For performance reasons, some of the data sets have been scaled down to 1024x1024 or 512x512 pixels.

2. Artificial Grid

The grid is not really applied (only conceptually), but is a specification of a N x M array for the patch extraction (next step).

3. Patch Extraction

Each image is loaded from disk and sliced up into a N x M array of 'sub-images' (the patches). The sub-images are kept in memory for image statistics (next step)

4. Image statistics (= features)

For each image patch, basic image statistics of the pixel values are calculated (like the mean, standard deviation, mode, kurtosis, skew). The exact choice of which statistics to use followed from domain knowledge, exploratory data analysis of the histograms (see chapter 12) and experimentation.
The calculated statistics are stored in a table, each column representing a statistic and one row entry per patch.

5. Normalization

Before feeding the statistics to the machine learning algorithm, the numbers are standardized to zero mean and a standard deviation of 1 (per statistic). The primary purpose of this normalization is to give each statistic the same order of magnitude, which is required for many machine learning algorithms (for more on normalization, see e.g. the Wikipedia article 'Feature Scaling' [10]) . The normalized statistics of each patch can be considered as the components of a feature vector describing the patch in feature space.

6. Transformation

The goal of the transformation step is to find an alternate representation of the data that facilitates the clustering process (next step). One such technique is PCA (see e.g. Introduction to Statistical Learning [9] for more background on this technique), which also allows for dimensionality reduction and feature selection (in the new representation) by filtering out only the most significant contributors.

7. Clustering

At the clustering step, the machine learning algorithms are applied to find similar groups in the data. The outcome of this step is an additional column in the table, assigning a group number to each patch. Multiple unsupervised techniques were tried out, but as best results were obtained with K-means and Hierarchical / Agglomerative, the other algorithms were dropped.

8. Visualization

In order to visualize the grouping as a heatmap, each group number is assigned a color coding. Note that at this point, each group is equivalent and the color assignment is arbitrary. The goal is to indicate the different type of areas in the image (in terms of image content).

9. Labelling

In a final step, the user is asked to indicate the group that is relevant. It is that final step which labels the data and makes one group more important than the others. Note that this step has not been implemented in the prototype, but is fairly trivial. Alternatively, the group importance could be assessed based on prior experience or user interaction, but that is more of a refinement of the concept.

6.3 LOCATING ACCURACY

With the provided approach, the micro-crystals are only located coarsely, depending on the granularity of the artificial grid. That is good enough for determining the location where to record the next image at a higher magnification. If a higher localization accuracy is required, e.g. for the final data collection at high magnification, the granularity of the artificial grid can be increased.

However, that has its limits: if the grid cells become too small to accommodate image features, no meaningful image statistics can extracted (in the most extreme case of a one pixel grid cell, there are no statistics at all!). Hence, for more exact localization, an alternative is proposed (next paragraph)

6.4 Refined search with a 'sliding window'

To pinpoint the micro-crystals more accurately, a variant of the technique can be used in which the artificial grid is replaced by a 'sliding window' as depicted below

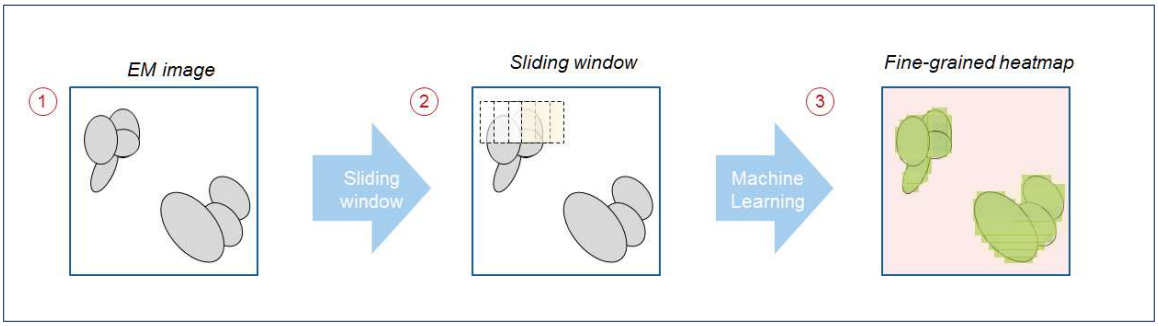


Figure 13 Schematic representation of the fine search. First, an electron microscopy image is recorded (1); then, a 'window' is moved over the image pixel-by-pixel, extracting statistics for the image contents below the window at each window location (2); the image statistics of each window location are fed to an unsupervised learning algorithm to cluster these into groups (2); each group is assigned a color and plotted as an overlay onto the original image at each window location (3).

6.5 Further Reading

The machine learning pipeline was developed incrementally; chapter 12 reports in more detail on the development and implications.

7 THE RESULTS

This chapter shows some of the highlights. More results and detail are shown throughout chapter 12.

7.1 RESULTS ON DISTINCTIVE DATA SETS

In these data sets, the micro-crystals are easy to identify. The ML pipeline is able to detect the micro-crystals in these images as well (shown in the figures below).

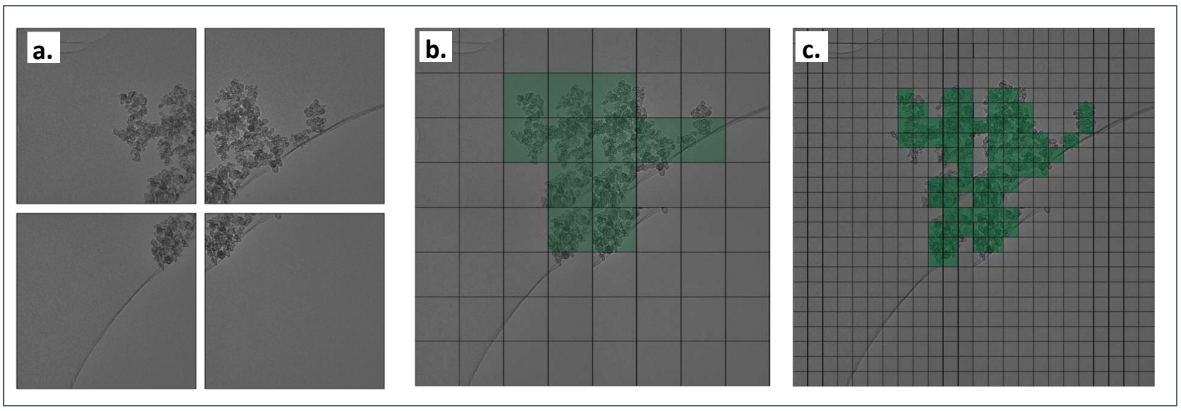


Figure 14 Results on data set 1, containing nanocrystalline graphite particles at high magnification. **a.)** the original images; **b.)** detection with a coarse grid; **c.)** detection with a fine grid

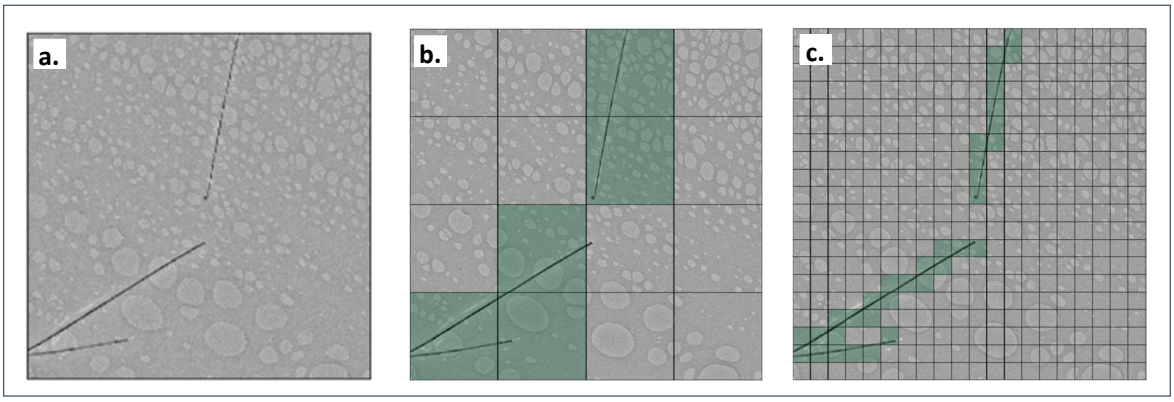


Figure 15 Results on data set 3b, containing asbestos fibers at a medium high magnification. **a.)** the original image; **b.)** detection with a coarse grid; **c.)** detection with a fine grid

7.2 RESULTS ON MORE INTRICATE DATA SETS

In these data sets, the particles are much smaller in the image and the images contain black areas which required special treatment. A two-step approach is applied in order to detect the paricles: first filtering out the black patches, using the same ML pipeline, and then applying a grid that is fine enough.

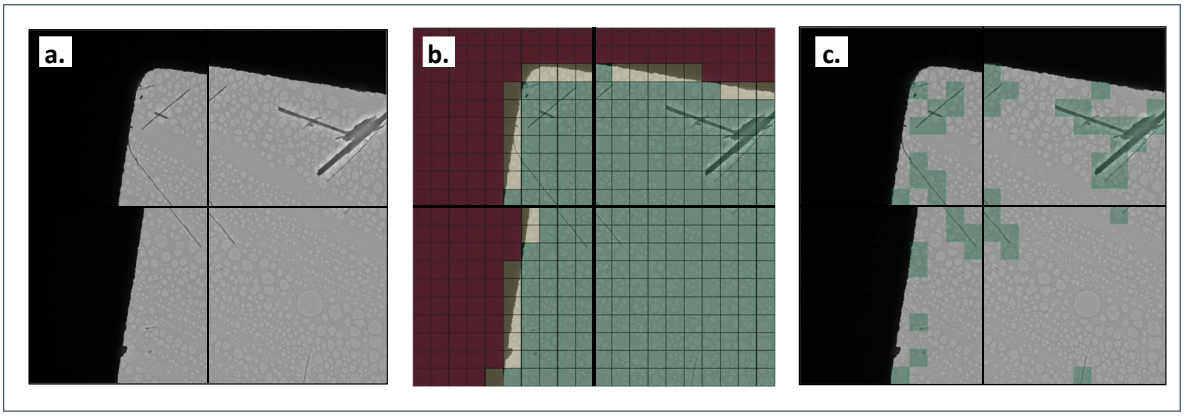


Figure 16 Results on dataset 3a, an image set of asbestos fibers at low magnification. **a.)** The original images; **b.)** In the first step, the black and partial black patches are identified; **c.)** after the filtering out the (partial) black patches, the fibers are detected with a fine grid.

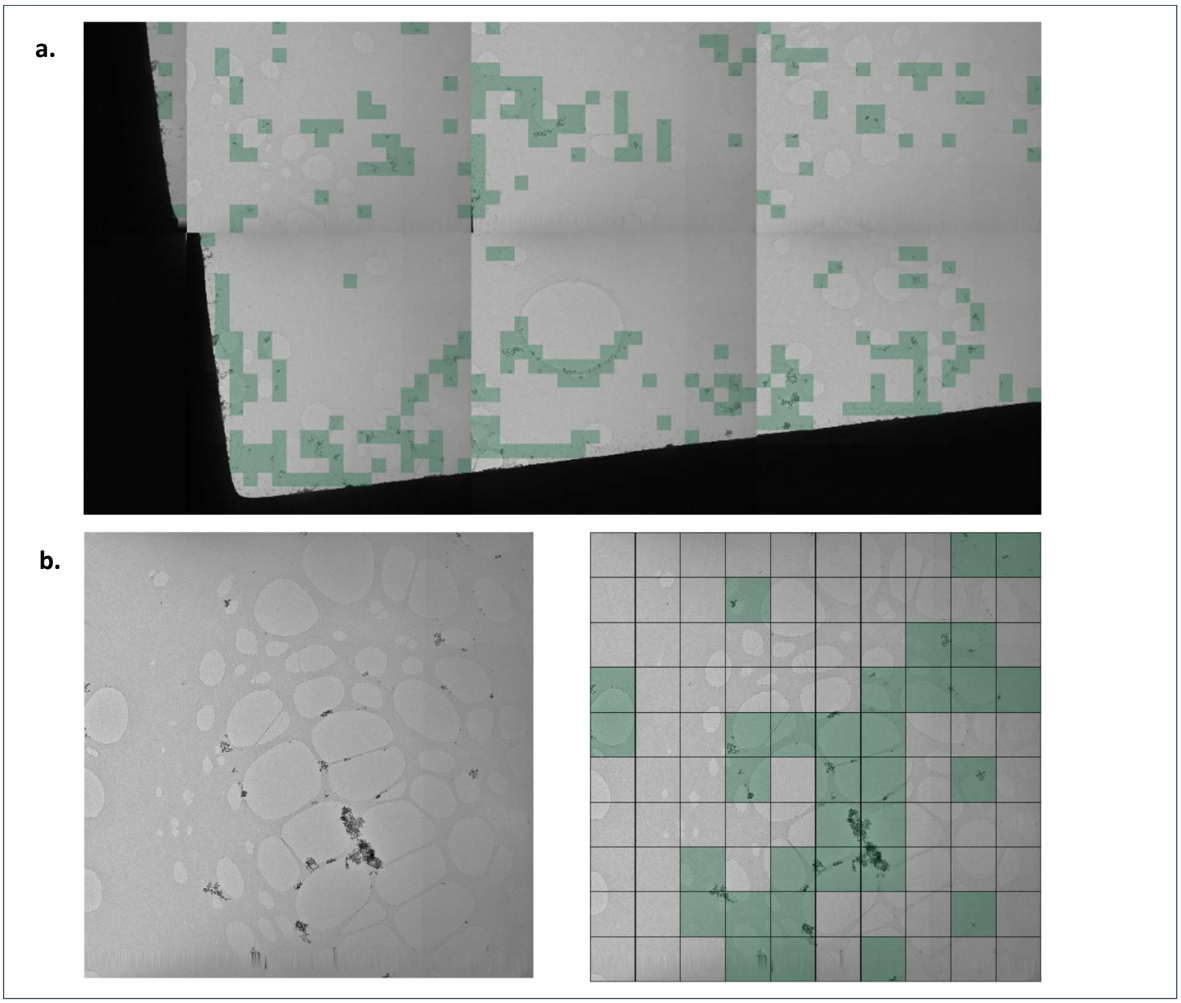


Figure 17 Results on data set 2, containing nanocrystalline graphite particles at low and medium magnification. **a.)** in the large overview, consisting of multiple images, the particles were detected with the two-step approach, first filtering out the black parts. **b.)** on the medium magnification images without black areas, the particles could be detected directly with a fine grid

7.3 Results with Fine search

With the variant that uses a 'sliding window', the particles could be detected accurately. This approach even gave some results on very difficult data set where the particles are hard to distinguish from the background (Figure 20); but this worked only on selected areas and by careful tuning.

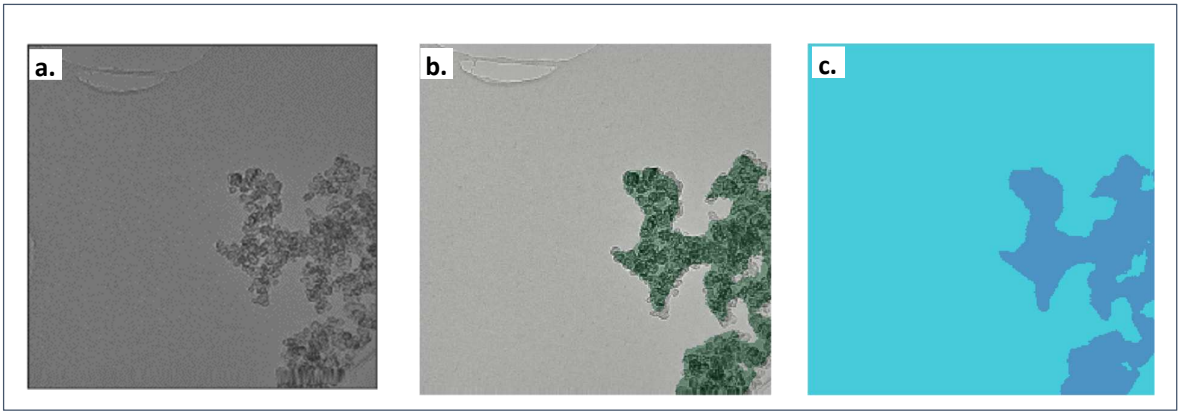


Figure 18 Results of the sliding window approach on a single image of graphite particles. **a.)** The original image; **b.)** the detected cluster plotted as an overlay on the original image; **c.)** the binary image of the cluster corresponding to the particle.

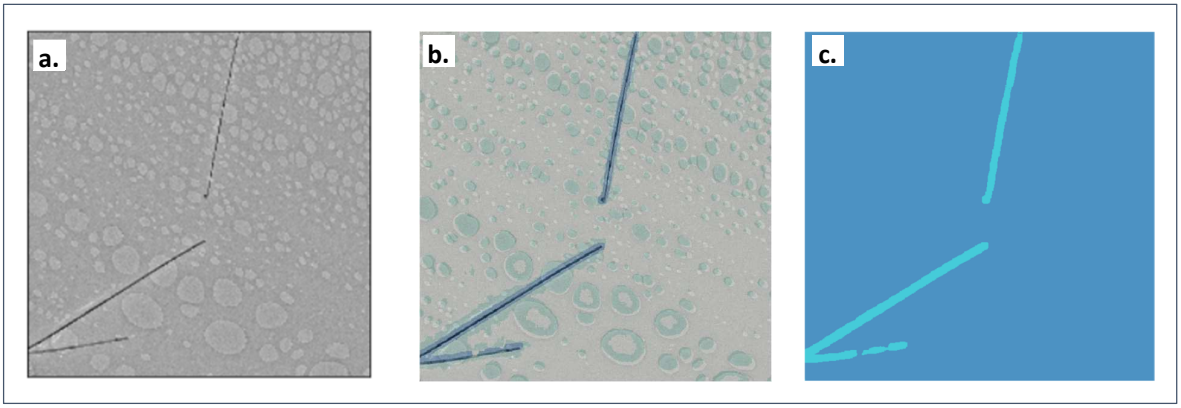


Figure 19 Results of the sliding window approach on an image of asbestos fibers. **a.)** The original image; **b.)** the detected fibers plotted as an overlay on the original image; **c.)** the binary image of the cluster corresponding to the fiber

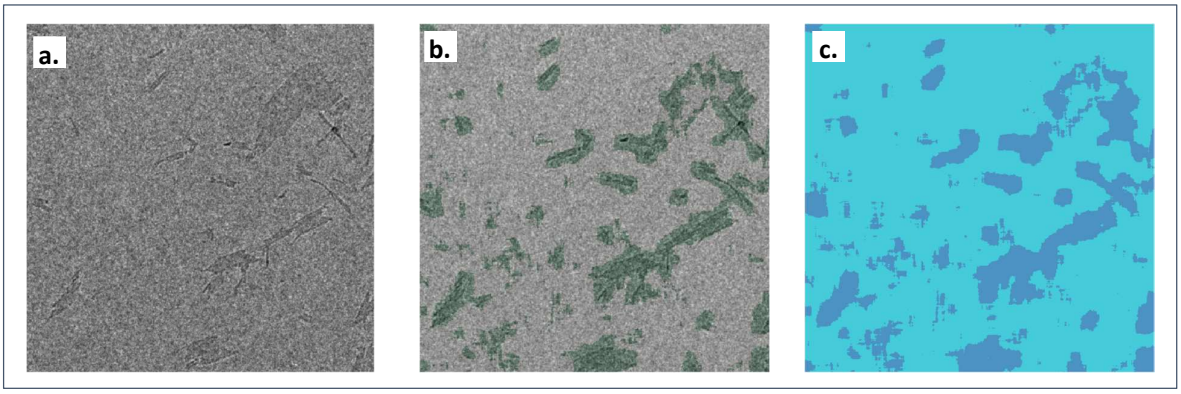


Figure 20 Results with the sliding window approach on a polymer data set, in which the polymers are hardly visible. This set did require sanitizing the data and careful tuning

8 FINDINGS & DISCUSSION

The topics discussed here are a summary of the findings and observations encountered during the development of the machine learning pipeline. More detail on these topics can be found in chapter 12.

8.1 OUTLIERS OR A NEW TARGET?

The (partial) black image patches were a challenge as they perturbate the statistics significantly. Filtering them out in a two-step approach is a successful strategy to counter this. Detection of the (partial) black images can based on the same clustering techniques used for detecting the micro-crystals.

The detection of the black parts has a value on its own: black areas are quite common in TEM images and need to be excluded from further images in many scenarios.

8.2 HOW MANY CLUSTERS TO SEARCH FOR?

The optimal number of clusters to search for is data dependent, but the following guidelines apply:

- Start with 3 clusters, then the '*maybe* cases' tend to get separated out in their own group, making it easier to select the positive case (instead of the *maybe* cases ending up as false positives).
- Reduce to 2 clusters if the particles are very distinct on an even background;
  (this mostly worked well with *k-means*; *hierarchical clustering* still required 3 in most case)
- Try 4 clusters if the image contains black regions (patch size needs to be small enough). If that does not work, use the two-step approach to first filter out the black regions.
- For filtering out black regions, 3 clusters are required (with 2 clusters only, some of the partial black regions will not be filtered out).

Automatically tuning the number of clusters with a metric did not result in a reliable assessment; visualization and user confirmation are more reliable (see also next section 8.7).

8.3 ACCURACY & GRANULARITY

One of the most critical parameters for crystal detection with this technique is the number of patches and thus it's size. If the patches are too big, the particles are not revealed by the statistics. On the other extreme, if patches are too small, they don't show any statistics at all. A good starting point is a patch size comparable to the expected size of the particles to be detected.

The 'sliding window' variant is able to locate the crystals with much higher accuracy, but is also computationally much more intensive and getting close to the first layer of a *convolutional neural network*. The window size is a tuning parameter which has not been studied, but expected particle size seems a logical guideline as well. Note that the granularity does not come from the windows size, but from the pixel-by-pixel steps of the sliding window. A sliding window with a larger step size will speed up the search considerably at the cost of some granularity, but still with more accuracy than a fixed grid.

8.4 WHICH UNSUPERVISED TECHNIQUE?

A number of unsupervised learning techniques was tried out, but quickly the selection was reduced to *hierarchical clustering* and *k-means*.

Hierarchical clustering gave the best results on some of the data set and its intuitive interpretation makes this an attractive method. However, this method does not scale well for large number of data points (a known limitation), which I also encountered when aiming for small patch sizes (= many data points per image).

K-means with PCA was able to deal with the largest number of data sets. As k-means is also a relatively simple and fast algorithm, this is the preferred approach for now.

8.5 Which image statistics?

Although many statics were tried, especially on the more difficult sets, the best results were obtained with the most basic ones:

mean, standard deviation, relative standard deviation, kurtosis, skewness, mode, range.

However, the data sets looked at mostly contained high-contrast particles. which may explain the success of these very simple statistics.

For dealing with particles that are harder to distinguish from other image content, this can be an area of improvement. One should consider other statistics that are more sensitive to texture or geometry, for example *Haralick texture features* [13].

8.6 Alternative methods

During the development of the machine learning pipeline, multiple methods have been assessed to find the micro-crystals in the images. An interesting observation is that even simple approaches like hand-picked statistics or 'feature similarity' can already help to identify relevant areas in an image. As these simple methods do not require any machine learning algorithm, they can be easily integrated into a product as a first approximation.

8.7 The human touch

Unsupervised learning is mostly used on data that is not labelled. But without a ground-truth or golden standard, it is hard to measure the performance. A number of intrinsic scoring methods exist, but in effect they only measure how well the algorithm could find clusters in the data, not whether the clusters make sense.

In other words, though unsupervised learning can find patterns in the data, it cannot judge the outcome in a way that a human can. Hence, some form of human judgement is required to make it successful. This should be sufficiently addressed in an enclosing application with proper visualization and user interaction.

9 FUTURE WORK

9.1 A USER-FACING APPLICATION

The machine learning pipeline can be further optimize and enhance, but the first next step to pursue should be integrating the pipeline into a user facing application. Direct user feedback is key for identifying which group contains the micro-crystals.

Moreover, by attaching such application directly to the microscope, the concept can be evaluated within the real workflow of finding small particles. This may provide new insights which should drive the future direction of improving the algorithms.

9.2 PERFORMANCE ASPECTS

When integrating the pipeline into a user-facing application, performance aspects will play an important role. Ideally, users should get almost instantaneous visual feedback when they adjust the number of clusters or patches. This is not viable within the current Python notebooks. The bottleneck is not the machine learning, but the image processing and visualization. However, this can be easily overcome by building the next prototype onto the infrastructure of production applications, in which image processing and visualization have been optimized.

9.3 IMPROVED MODELS AND/OR MORE MACHINE LEARNING

Collecting more data sets and interactive tuning, allows for further analysis on how to detect the particles in the image statistics. Maybe with well-chosen image statistics, a simple rule-based algorithm can be derived to automatically identify which cluster holds the micro-crystals.

Note that when the user indicates the group associated with the particles, he/she is essentially labeling the data. Labeled data opens up the opportunity for *supervised* machine learning to obtain better models.

An interesting concept to look into is to use these labeled image patches for training deep learning networks, which hold a great promise for particle identification. By applying *transfer learning* (i.e. reusing a pre-trained network from another domain), training such networks should be viable even with a small number of images.

10 CONCLUSIONS

The work presented in this report shows that the proposed concept is viable for crystal detection. Additional measures have to be taken to deal with some artifacts that can be present in the microscope images, but can be countered with additional measures.

Of course, the technique still has to prove itself on a wider range of data sets and may need to be matured. But that can best be done by integrating the pipeline into a real application that acquires images directly on the microscope.

Several colleagues have shown interest in my work and it is fair to expect that this innovation will become part of future roadmaps. At this moment in time, priorities in product development are already set, but when the *MicroED* workflow has matured, there is room for new time saving developments. Furthermore, this innovation may find its way in other application areas than *micro-crystal* research.

Most and for all, this project was an invaluable learning experience. In the DEP program I learned more than I could have hoped, finally making my first steps into the field of data science. And my journey does not end here, it has just begun…

Part II - The Journey

Lessons Learned along the way

11 The Entrepreneurial Journey

This chapter describes the project from an entrepreneurial perspective, starting from the initial ideation phase to building a case and (trying to) get stakeholder buy-in and customer involvement.

11.1 Idea Generation & Selection (Nov 2017 – Feb 2018)

Before I started the JADS DEP 2 course, I already had an idea for using image recognition and machine learning for a project I was working on. In the first few months of the JADS course, I iterated to a modified version of my initial idea.

Using the brain storming techniques that we learned in the entrepreneurial sessions and during the boot camp weekend, I generated more ideas and categorized them. Then I assessed the business value of each idea and talked to various people within my company to find running projects to which I could attach the ideas. I also performed a 'COCD Box' analysis on feasibility and originality.

Finally, adding 'personal drive & ambition' to the equation, I came to the 'The Micro-Crystal Locator'.

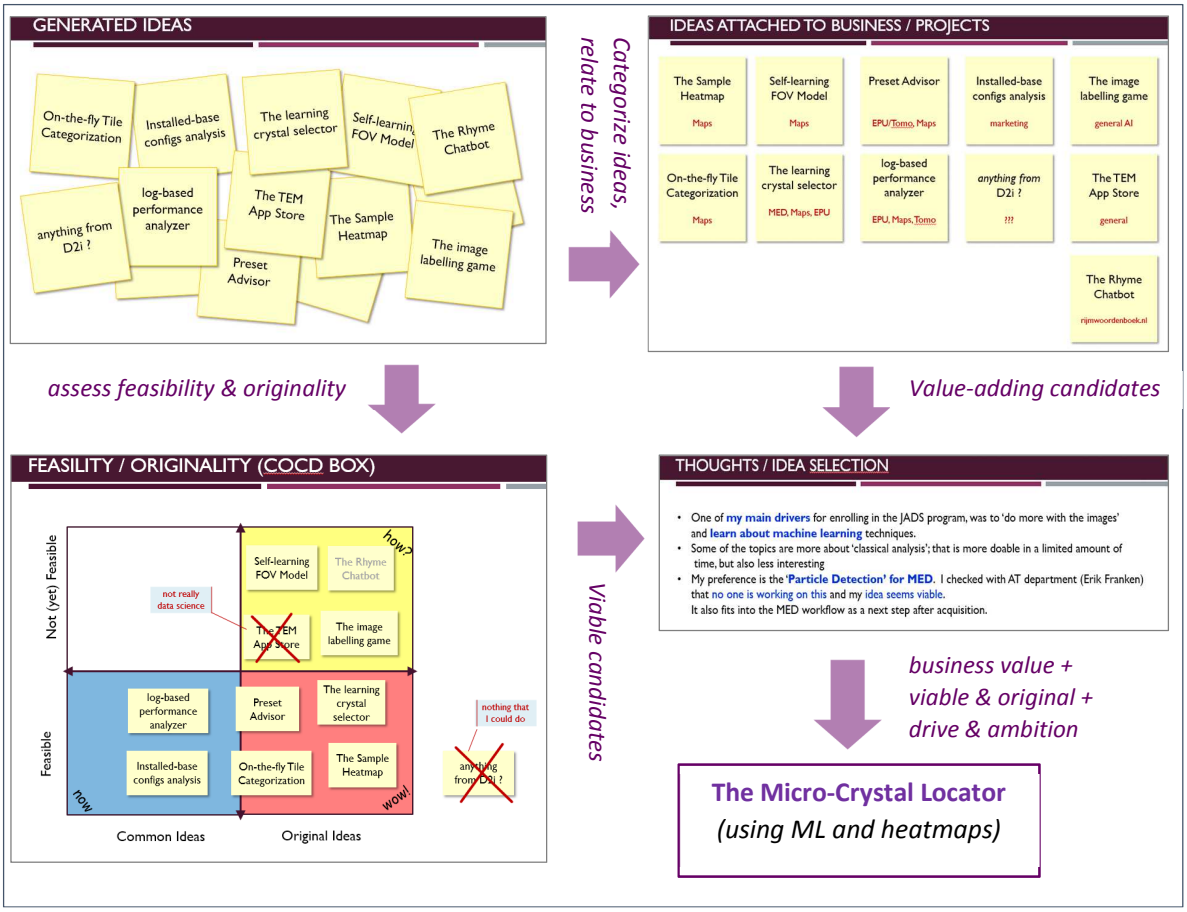


Figure 21 Finding a topic for the JADS graduation project, from initial idea generation to final selection.

11.2 Customer Validation (Apr 2018)

Goal:

Confirm if customers are interested in the idea of automatic micro-crystal detection within the context of *MicroED*, a technique for which my company is developing a new product.

Customer visit:

In April, me and a few colleagues from my company visited a potential beta customer for the *MicroED* application that we were about to start developing. Main goal of the visit was to validate our insights on how our application should support the *MicroED* workflow, which could be complemented by automatic micro-crystal detection.

Outcome:

They indeed confirmed the need for automatic micro-crystal detection, but first the advanced data collection of *MicroED* needs to be automated and under control.

They also suggested to use heatmaps to indicate relevant locations; interesting enough, that was actually part of one of my other initial ideas ('The sample heatmap'), which indeed is related. Hence I decided to incorporate the concept of heatmaps into the 'micro-crystal locator' as well.

An interview with another (pioneering) customer had similar out-come.

Lessons learned:

- Customer(s) do see value in automated crystal detection
- Other parts of the *MicroED* workflow need to be in place first.

11.3 Getting buy-in from stakeholders

Despite my efforts to make a compelling story (the pitch at JADS) and customer interest in automated detection, it seems not viable at this point in time to get buy-in from my management to work on this idea now. Main reasons are other short-term objectives and priorities and doubts about the direct business value.

As I still believe in the value of automated crystal detection, I need to find (own) time to pursue this endeavor; maybe initial results are needed for credibility and convincing stake holders of its value.

11.4 Defining metrics (May 2018)

As the 'micro-crystal locator' cannot be seen in isolation of the *MicroED* workflow, I defined metrics for the entire application that we are developing, not just for crystal detection. In this way, I can still perform a baseline measurement and start tracking performance.

The three metrics I defined are: total-time-per-microcrystal, customer satisfaction, and - when a full workflow application is in place - resolved resolution per micro-crystal (which is the outcome of a *MicroED* experiment)

Lessons learned:

- Defining viable and maintainable metrics upfront for a product that does not yet exist is difficult.
- It is hard to prove the value of such metrics to my (internal) stakeholders without an initial track record of the metric.

11.5 Revising scope & Adjusting Target customer (may/june 2018)

Due to limited bandwidth available for this project, I had to revise the scope: I will focus now on an isolated prototype to show the concept works instead of new functionality integrated into a product.

I also had to change my target audience from an external beta customer to an internal client as the beta program had been postponed until November (due to a dependency on the unavailability of a special detector which is critical for *MicroED*).

Lessons learned:

- Tying my JADS graduation topic to a project with many dependencies that are outside my scope of control may not have been the best choice (hind sight).
- Without some first actual results, I will have a hard time to get people on board for my idea.

11.6 CAUTIOUSLY SHARING FIRST RESULTS (JULY/AUG 2018)

As I was getting some results from my work, I gradually started to show my work to some colleagues to pitch their interest and get feedback. I got useful technical advise from one of my direct colleagues who is knowledgeable in the field of data science.

11.7 SEIZING OPPORTUNITIES TO SHOW RESULTS TO A WIDER AUDIENCE (SEPT/OCT 2018)

As we had to give two demos to business stakeholders on the progress of our application development efforts, I 'high jacked' the latter part of the demos to show my recent work on micro-crystal detection. Responses were mixed; some challenged the needs, other were more enthusiastic about the idea.

Lessons learned:

- Having some results makes it so much easier to share and promote your plans
- Apparently, the business value of micro-crystal detection is not obvious enough at this point in time.

11.8 OUT IN THE OPEN (OCT 2018 AND ONWARDS)

As work progressed, I now have compelling material to show that the concept is viable. I have shared my results with different audiences, so people are aware of my work. The topic is discussed now and then and, maybe, it will end up on the product roadmaps in the second half of 2019...

Lessons learned:

- If you cannot find ways to take-off quickly, your run-way needs to be really, really long!

12 THE DATA SCIENCE JOURNEY

This chapter describes in chronological order how the machine pipeline was developed. It reports in-depth on the data analysis performed, on the technical hurdles and on the lessons learned at each step.

12.1 TESTING ASSUMPTIONS : ARE (SMALL) PARTICLES OBSERVABLE IN THE IMAGE STATISTICS (MAY 2018)

Goal:

Conclude if it is viable at all to use basic image statistics (i.e. statistics on the gray values of the pixels in the image) to reveal if an image contains a particle or not?

The experiment:

To answer this question, a hypothetical test set was created, consisting of two types of images: with and without a particle. This hypothetical set was created by cutting out regions from an arbitrary set of microscope images, each region containing either a dust particle or a featureless area. Then, the experiment as summarized in the figure below was conducted with python scripts

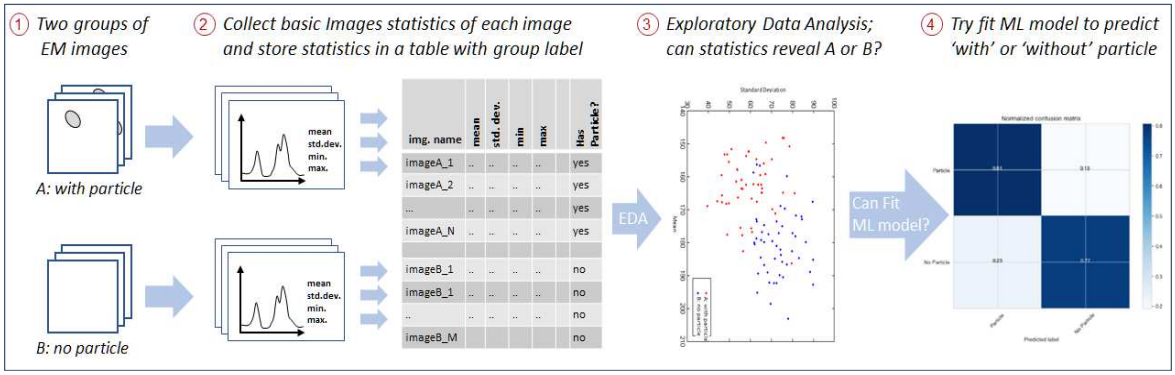


Figure 22 Schematic representation of the experiment to assess if the presence of a particle in an image is revealed in the statistics of pixel values. 1). Two sets of images: with and without a particle; 2.) For each image, calculate simple statistics of the pixel gray values and store in a table; 3.) Use exploratory data analysis (histograms, scatter plots) to assess if images with or without particles have (combinations of) different statistics; 4.) Fit a simple model to the data and check its prediction accuracy.

Exploratory Data Analysis:

From inspecting the histograms and scatter plots (shown in Figure 23), I could confirm that even simple statistics can reveal the presence of a particle. Special care is needed to handle the outliers, which originate from (partial) black images that obscure the statistics. After removing these outliers, the two groups (with and without particle) could be distinguished by basic statistics. I could fit a *k-nearest neighbor* model to the data with a predication accuracy of ~ 80-%; this is not perfect, but good enough to validate the assumption that particles can be revealed in the image statistics.

Lessons Learned:

- A simple data set shows that particles indeed can be revealed from the image statistics
- Outliers easily obscure the image statistics; especially (partially) black images should be accounted for as they do occur in real data sets as well ('grid bars' on the specimen)
- Analyzing and understanding the data points, requires inspection of the corresponding image; this is cumbersome without more assistance: **need better visualization & infrastructure**

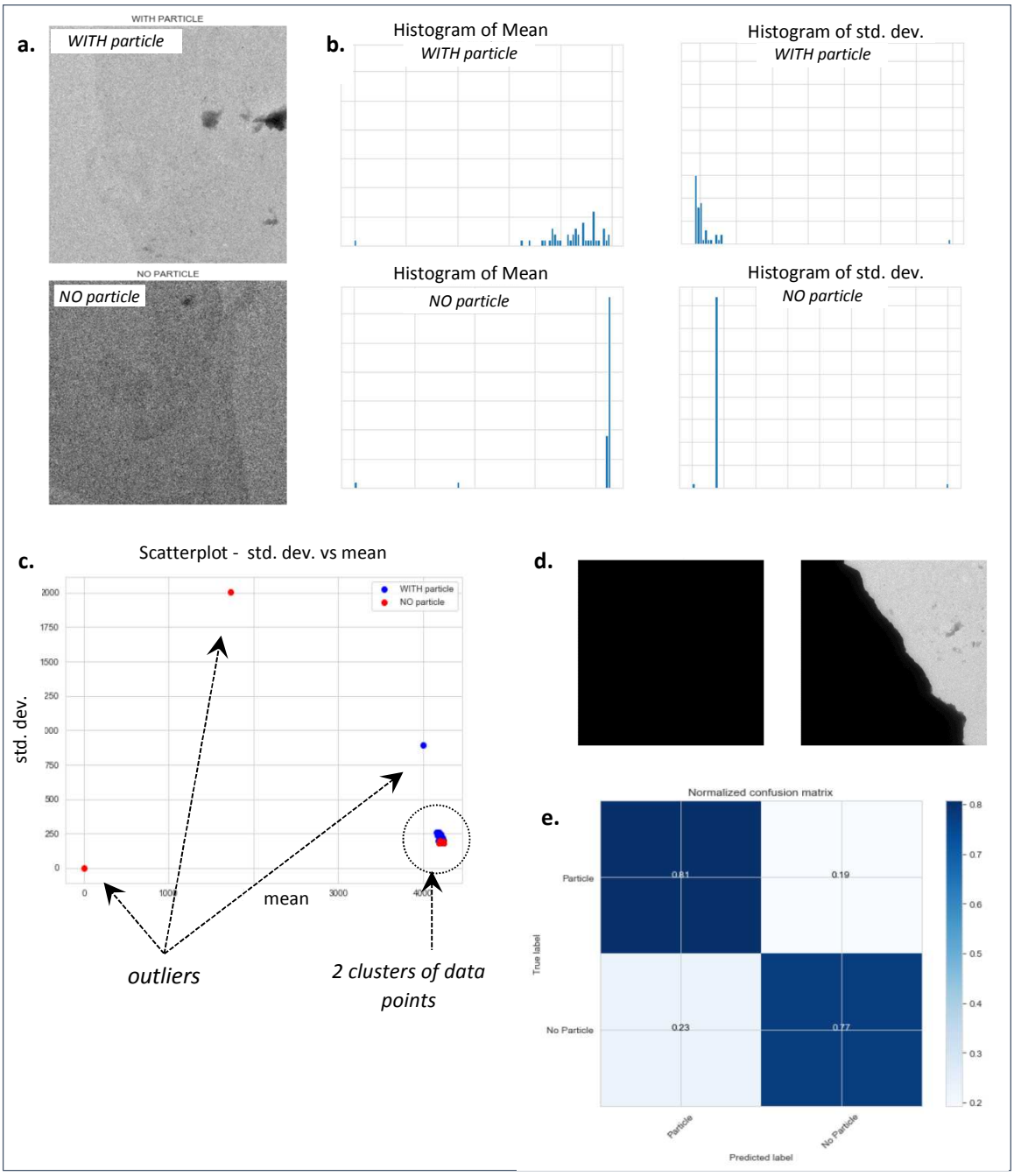


Figure 23 Exploratory data analysis on the images with and without particles. **a.)** Examples of a images with and without a particle. **b.)** The distribution of the mean and standard deviation of the pixel values of *all* images with and without a particle. On first sight, the histogram for images with and without particles look different, but this picture may be obscured by the outliers. **c.)** Scatter plots of the mean vs. standard deviation. Two clusters can be identified, but appear very close together as a few outliers blow up the scales of the x- and y-axis. **d.)** Inspection of the images corresponding to the outlier data points revealed the cause: fully or partially black images; such black images originate from 'grid bars' on TEM specimens.
e.) After removal of the outliers, a simple model could be trained on the data; the plot show its confusion matrix.

12.2 Analyzing First real data set & building infrastructure (June 2018)

Goals:

- Analyze a first real data set
- Build up infrastructure for visual data analyses and automate some of the data engineering

Data import automation

The data import and feature extraction on the first real data set is depicted below

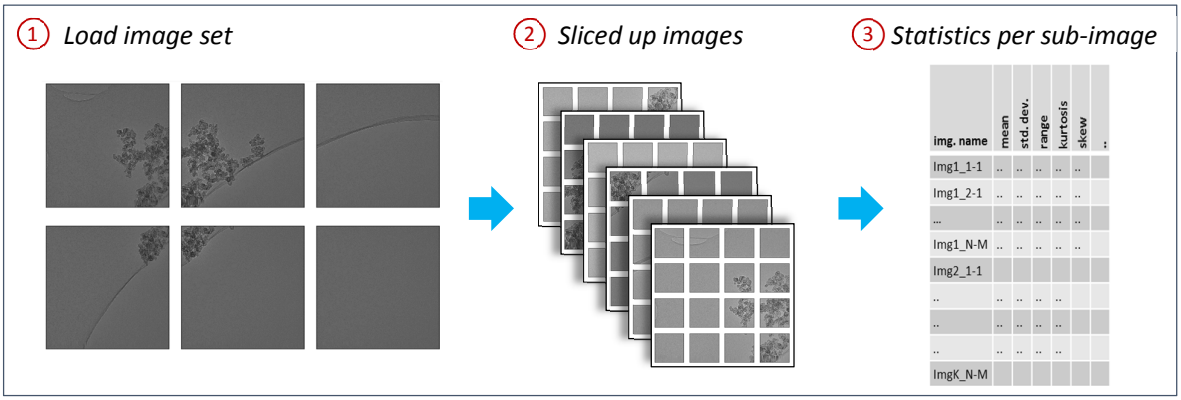


Figure 24 Schematic representation of the data import and feature extraction. 1.) Loading images from a *mapping* set (the images in this set have some overlap); 2.) slice up each image into patches; 3.) extract multiple statistics from each patch and store these in a table for further analyses.

Exploratory Data Analysis:

In the figure below, different combinations of statistics are plotted in order to see if patterns or clusters can be revealed. There appears to be some grouping of datapoints in the plot of *mean* vs *standard deviation*, but this needs closer inspection. Similar grouping is found in the scatter plots of *grayscale range* vs *standard deviation* (not shown).

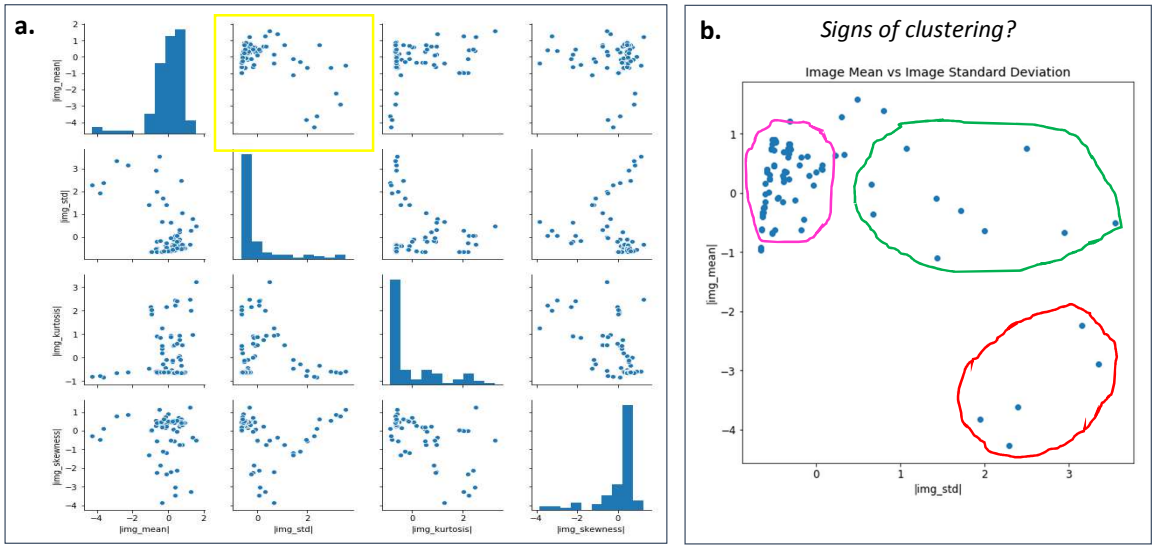


Figure 25 Pair-plots of a number of the image statistics that had been extracted from the patches. There seems to be groups in the mean vs standard deviation (enlarged in **b.**), but further inspection is required to assess if this is real or accidental.

Normalization:

Note that I normalized each statistic using *standard scaling,* even though the statistics are not normally distributed. Still, this is a common technique and greatly improves clustering analysis. A good reference on this topic is Wikipedia Feature Scaling [10])

Inspection with improved visualization:

To inspect the scatter plots more closely, I developed interactive graphs that shows the corresponding image patch of the selected data point.

And indeed, after inspection with the interactive plots, image patches with or without a micro-crystal can be loosely associated with groups in the data. *This looks promising!*

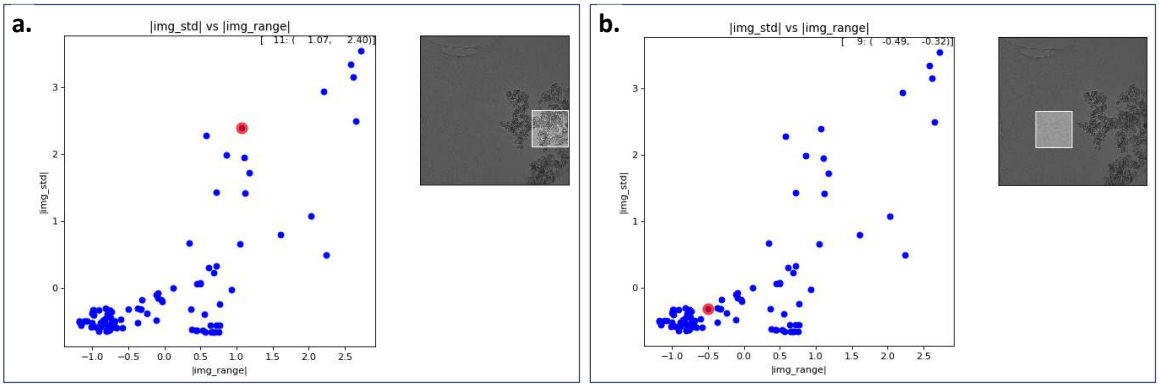


Figure 26 The interactive scatter plots show the corresponding image patch of the selected data point

Hand-picked statistics for a heatmap:

Even without machine learning, for this dataset a hand-picked statistic (i.e. the *standard deviation*) is already a good first indicator for the presence of micro-crystals. Hence, I could use this hand-picked statistic to create my first heatmap as shown below.

The visualization reveals that the hand-picked statistic is not without flaws: it also highlights other features in the image.

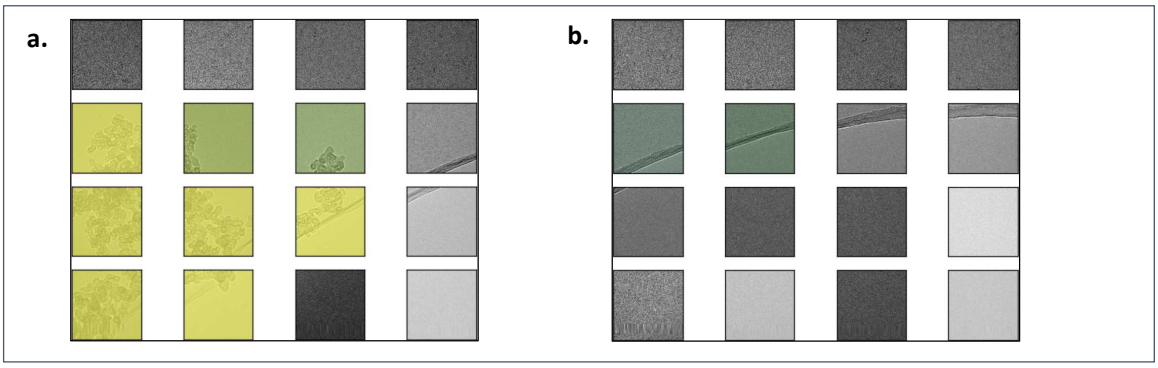


Figure 27 Heatmaps for two images based on a single statistic. The statistic does a proper job in identifying the micro-crystals (as shown in **a.)**. However, the statistic cannot distinguish between patches that only contains parts of a crystal or other background features (like the edges in the second row of image **b.**)

Note that the visualization still needs some improvement; the patches are shown with different contrast/brightness which obscures the perceived heat. I will refine that later.

Lessons learned:

- The build-up infrastructure and visualization greatly facilitates the data exploration
- On this first dataset, even a hand-picked statistic can be used as a first approximation
- Visualizing the results as a heatmap reveals that hand-picked statistics is not perfect
- Next step: try to find more (combinations of) more distinctive statistics

12.3 Try Find better statistics (July 2018)

Goal:

Find statistics that are less sensitive to background features (*standard deviation* also picked up edges).

Results:

I tried out many combination and variants, but unfortunately none of those statistics were more distinctive than the basic statistics used in previous experiment (i.e. mean, standard deviation and grayscale range). Some of my attempts are shown.

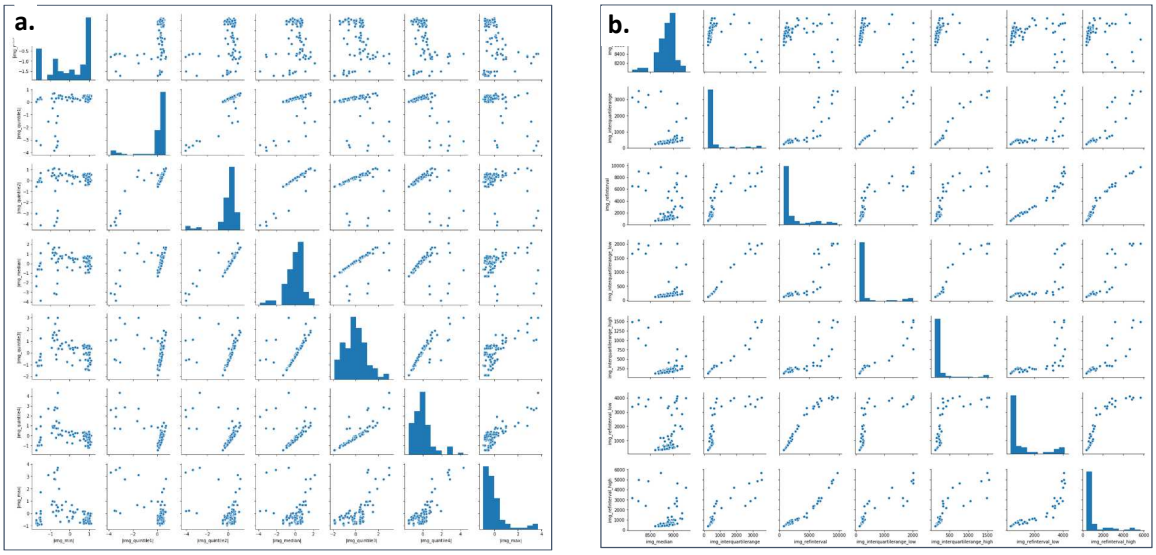


Figure 28 Exploring different types of statistics for data set 1. **a.)** pair-plots of statistics based on *quintiles*; **b.)** pair-plots based on *5 number summary statistics* and *interquartile ranges*. Notice the high correlation between many of them, making those less suitable for separating clusters.

Lessons learned:

- Finding distinctive statistics is more challenging than prior successes were indicating.
- Are my assumption wrong? Are the statistics of particles and background features too? Or have I just not found the right statistics yet? Need to analyze another data set

12.4 Struggling with the Second data set (July 2018)

Goal:

Repeat the experiment on a larger data set with smaller crystals and where black borders are present.

Results

Disappointing. The heatmap shows that the statistics are not able to catch the particles. It does separate the areas with (partial) black images from the area without borders, but not the crystals.

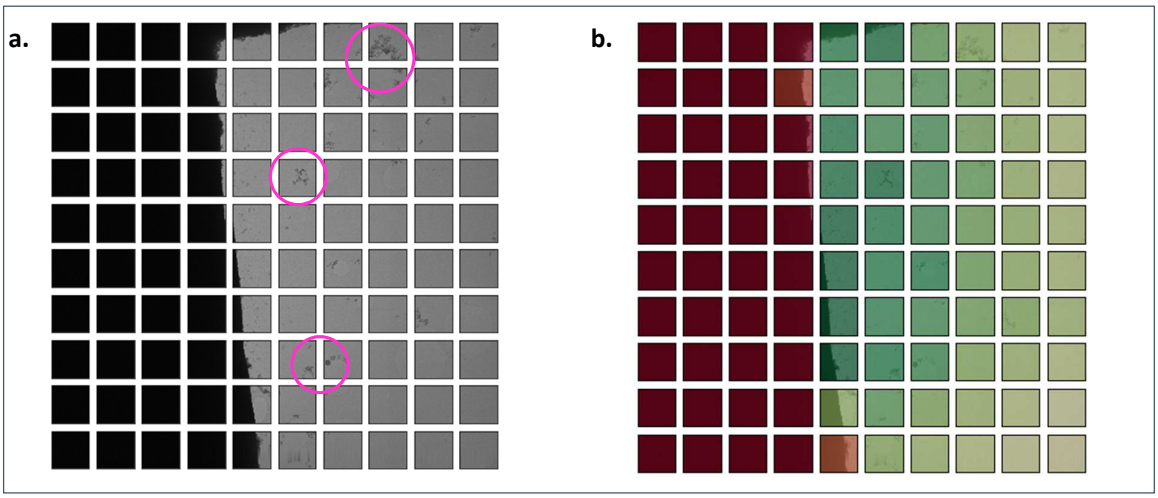


Figure 29 **a.)** One of the images in the data set, sliced up into patches. The purple circles indicate areas with larger number of micro-crystals. **b.)** A heatmap generated from this image based on a hand-picked statistic; it does not reveal the micro-crystals.

Analysis

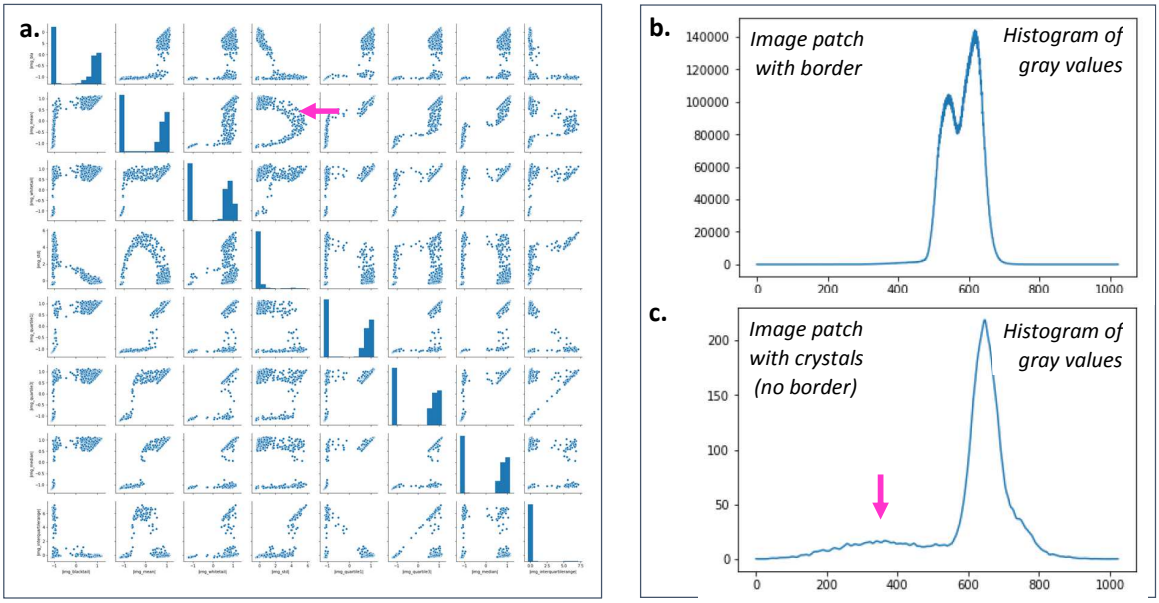


Figure 30 **a.)** Pair plots of a range of statistics of data set 2. The purple arrow indicates one of the 'continuous' groups that show up in some statistics, originating from the partial black images with different amounts of black. The large population of black images is clearly present in in many of the statistics. **b.)** A histogram of the gray values of an image patch with a black border; the two peaks correspond to the (normally distributed) black values and non-black values respectively. **c.)** A similar histogram an image patch with micro-crystals (without black borders); the main peak is the normally distributed background. The subtle side peak (indicate by the arrow) may be the signature of the small crystals.

The results can be explained by looking at the statistics on the image patches as show in Figure 30. The black and partial black images are very prominent in the statistics, the black images as single groups, the partial images a 'continuum', originating from the different amounts of black present in the images.

However, even without the black images, the particles are hardly visible in the histograms of individual images in this data set (Figure 30c) and may be too subtle to be 'seen' by summary statistics.

From the earlier experiment with artificial images, it already came up that (partial) black images obscure the statistics. So we may need to filter out the (partial) black images in advance. This is relatively simple as they clearly show up in the statics.

As combinations of three statistics show more structure, maybe they can be revealed by assessing feature importance or PCA.

Lessons learned:

- Early success on data set 1 is not replicated on data set 2
- Black and partial black images remain a problem, but can be filtered out via simple statistics
- Need to expand to new techniques (PCA, higher dimensionality, feature importance), but first on the easier data set 1

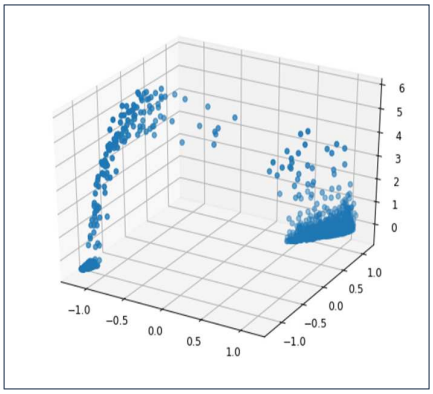


Figure 31 Combinations of three statistics show more structure.

12.5 Dimensionality and feature importance (July 2018)

Goal:

Analyze the features in higher dimensions and assess feature importance

The experiments:

1. Use hand-picked statistics (see paragraph 12.2) to label the patches from data set 1 into 3 categories:
   A:No Crystal , B: Partial Crystal, C: With Crystal
2. Inspect the labelled data set with 3D scatter plots with different combinations of statistics
3. Assess 'feature importance' by fitting a *random forest classifier* to the labelled data
4. Inspect the higher dimensional feature vectors with dimensionality reduction techniques

Results / Lessons learned:

- From inspection of this data set with 3D scatter plots, the 3 categories can be separated using certain combinations of three statistics (see Figure 32a)
- This is in agreement with the 'feature importance' by the random forest classifier (see Figure 32b)
- From the dimensionality reduction techniques, PCA and Isomap are able to separate the clusters in 2 dimensions

Remarks:

The data is most separable if three features are combined based on 'feature importance'. However, determining feature importance with random trees requires the data to be labelled, which is not the case for the 'micro-crystal locator' (this set was only labelled for analyses). Hence a technique like PCA is a more obvious choice for selecting features.

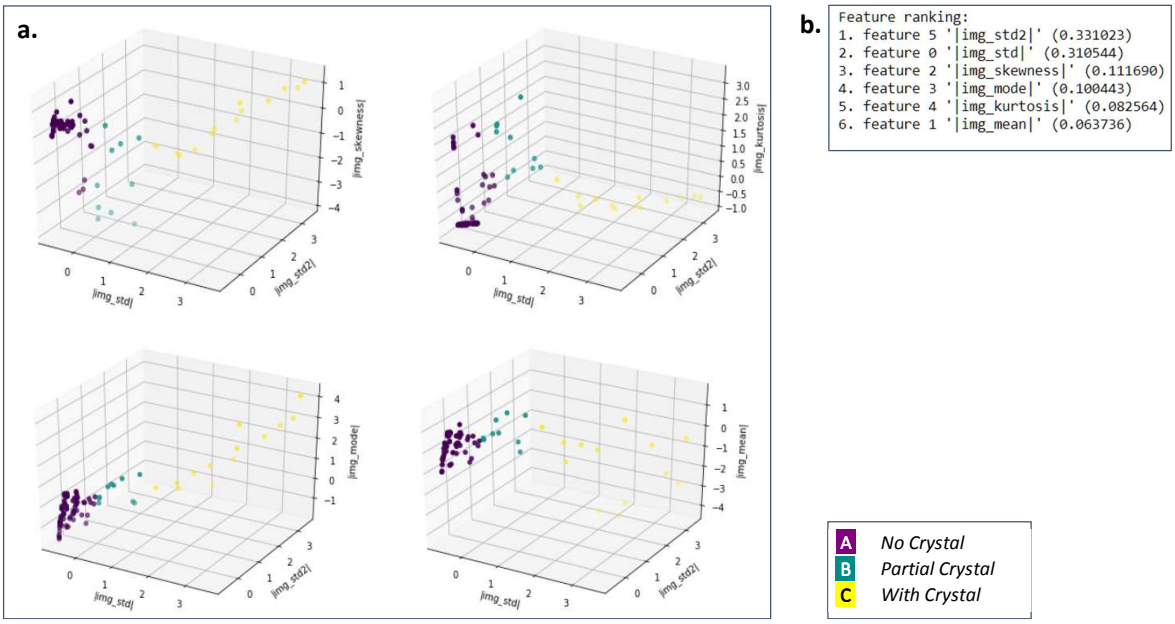


Figure 32 **a.)** Inspecting the basic image statistics in three dimensions. For this data set, any combination which contains the *standard deviation* and the *relative standard deviation* (named std2 in the graphs), can separate the three groups, but a third dimension is needed. **b.)** The result of the feature importance assessment. This is in line with the observations in the three dimensional plots, marking the *standard deviation* and the *relative standard deviation* as most important

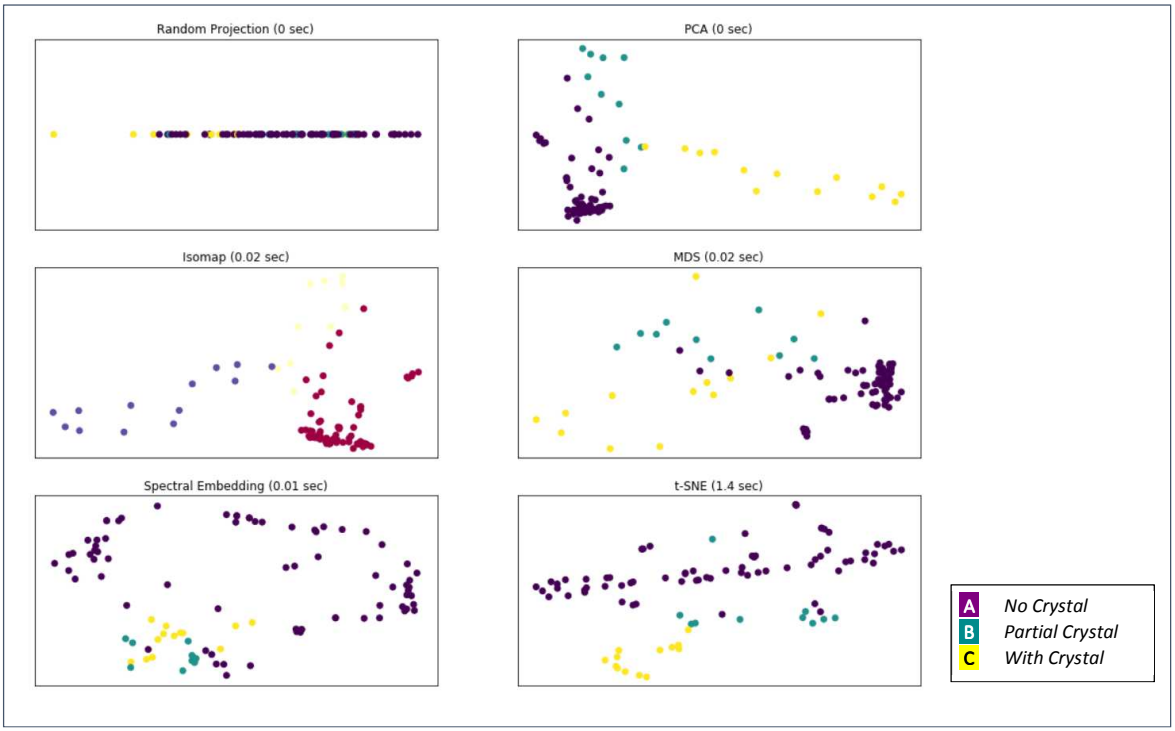


Figure 33 A number of dimensionality reduction techniques applied on the labelled data set. On this data set, *PCA* and *Isomap* were most successful in creating separable clusters.

12.6 INTERMEZZO: EDA WITH BOX PLOTS (AUG 2018)

Goal:
Inspect the features (i.e. the image statistics of the patches) of data set 1 with *box plots* (inspired by one of the on-line course of *DataCamp.com*).

Results:
The box plots are shown below. What is striking in these plots, is the asymmetry and number of *outliers* for each statistic (the points outside the blue box and its 'whiskers'). Clearly, the statistics are not normally distributed, which is what we want.

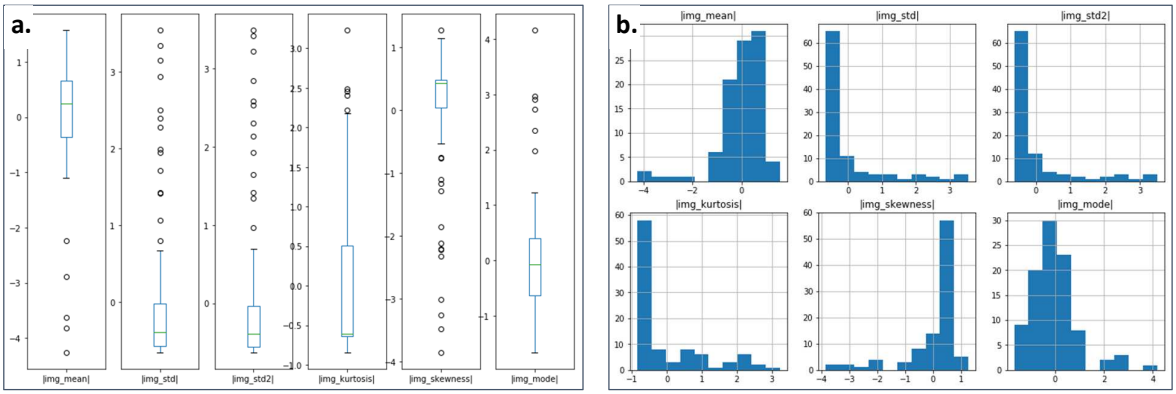


Figure 34 **a.)** Box plots of the (normalized) image statistics of the patches in data set 1. The asymmetry and large number of outliers for each statistic, are a clear sign of non-normal distributions. **b.)** The corresponding histograms, which indeed show strong asymmetries. Due to the prominence of the main mode in the histograms, it is hard to judge if there is a second mode or just an asymmetry.

In effect, if the statistics are indeed signatures of two (or more) distinct groups, a *bimodal* (or *multi-modal*) distribution is to be expected and it should be possible to separate them again.

Inspection of the corresponding histograms shows indeed the asymmetry, but a second (or third) mode is hard to distinguish due to the prominence of the main mode. Hence, combining a number of statistics to find the additional modes is still a logical strategy to pursue.

12.7 UNSUPERVISED LEARNING (AUG 2018)

Goal:
Explore if unsupervised learning algorithms are able to identify the areas with micro-crystals.

(All methods used are based on the *Scikit-Learn* tool set;

The experiment:
1. Load in the extracted features from data set 1 (i.e. the image statistics of the patches)
2. Manually classify the data set, dividing the patches into 3 groups (see paragraph 11.5)
   A: No Crystal , B: Partial Crystal, C: With Crystal
3. Try out a number of unsupervised algorithms to cluster the data set into 3 groups
   (methods tried: *k-mean, dbscan, spectral clustering, hierarchical clustering*).
4. Compare the 3 groups found by the algorithms to the manual classification

Results 1 – Exploratory evaluation:

- Both **k-means** and **hierarchical clustering** gave **reasonable to good results** as shown in Figure 35; there are some small differences between their outcome and the hand-labelling (discussed later on).
- DBScan and Spectral clustering did not perform at all on this data set (result not shown); this may be explained from their documentation, stating that these algorithms perform well on highly uniform shapes (which is not the case for this data set)
- An attractive property of *hierarchical clustering* is the possibility to visually inspect how it clustered the datapoints via its *dendogram* as depicted in Figure 36. The *dendogram* should be examined bottom-up, merging similar points into ever bigger groups. The *cut* determines the number of clusters found (or vise versa, specifying a number of clusters defines the cut).

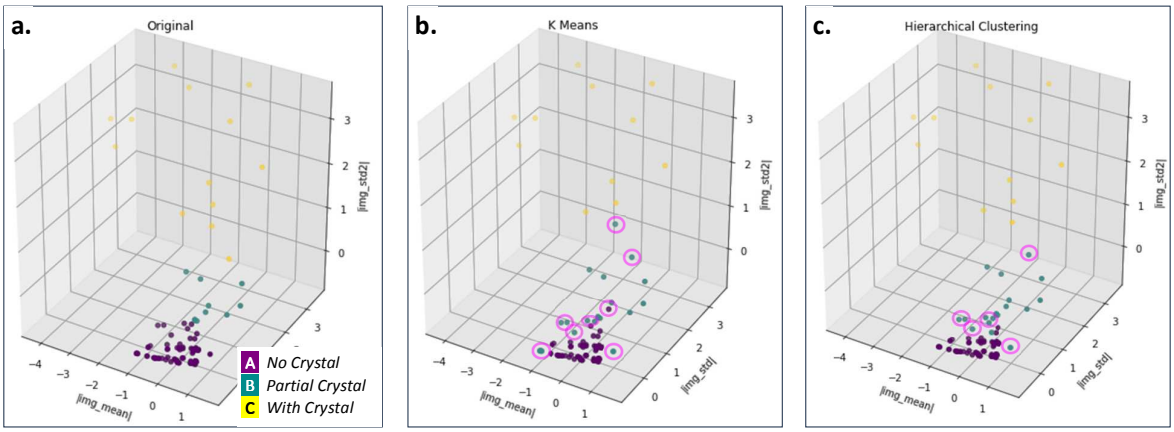


Figure 35 Data set 1 after clustering into three groups by different methods; the axis are the most important features (previous exploration indicated 3 statistics are required to clearly separate the groups. The color coding indicates the different classes. **a.)** Clustered by hand (the 'ground truth'); **b.)** and **c.)** Show the clustering results of with *k-means* and *hierarchical clustering* respectively; the purple circles indicate the data points that deviate from the 'ground truth'. Note that for the unsupervised results, the colors have been manually assigned to the cluster numbers (a cluster numbers are arbitrary) in order to compare the results with the ground truth.

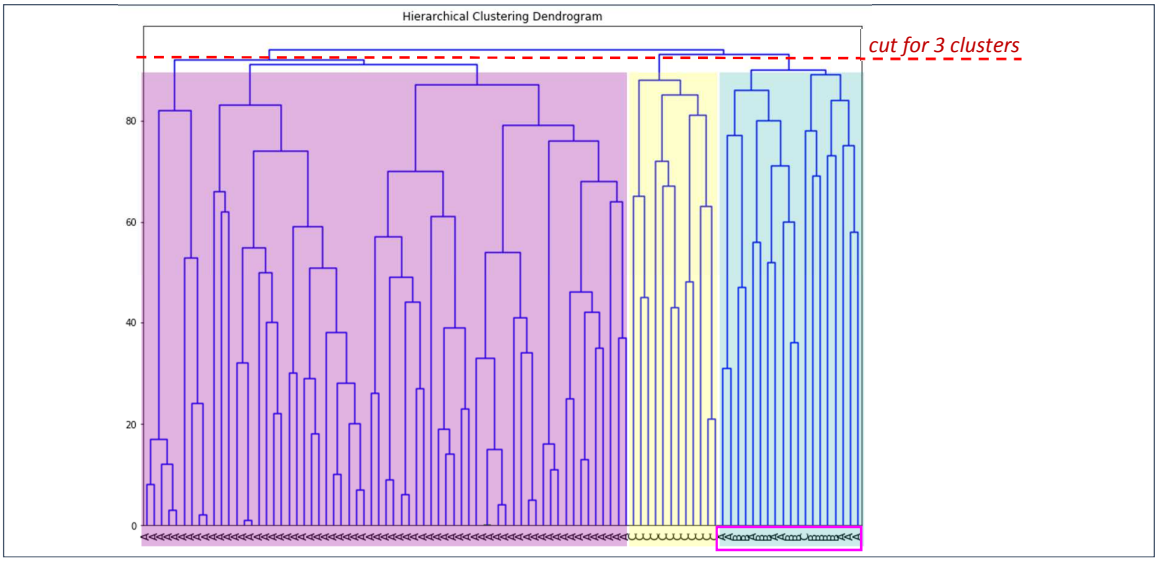


Figure 36 *Dendogram* of the hierarchical clustering result on data set 1. The coloring has been manually added to assist the comparison with the other graphs. The red dashed line indicates where 'the tree is cut' for 3 clusters. Note that the contents of the 3rd cluster – as indicated by the purple rectangle - deviate with the 'ground truth' .

Results 2 – Quantitative evaluation:

- Evaluating performance of clustering algorithms is not trivial, since the label assignment to each is arbitrary. Hence, any evaluation metric should not look at the value of the cluster labels.
- Two ways to evaluate the results of clustering algorithms quantitively are:
  a. Compare if the clusters have a similar separation as some 'ground truth' (if available);
  b. Assesses if the members that belong to the same class are more similar than members of different classes according to some similarity metric.
- Using the manual classification as the 'ground truth' option a. can be used. Multiple variants exist, results of scoring metrics available in the *Scikit Learn* toolkit [12] are shown in Table 3.
- According to all methods, **hierarchical clustering scores best**, closely **followed by k-means**; this is fully in agreement with the visual evaluation.

Table 3 Quantitative evaluation of clustering results on data set 1 using a number of scoring methods

clustering method: / *scoring method:*	**k-means**	**spectral**	**dbscan**	**hierarchical**
Adjusted Random Index	0.614064	0.257207	0.276250	0.722755
Mutual Information	0.534936	0.320783	0.327220	0.634900
Homogeneity	0.665933	0.493346	0.337546	0.762652
Completeness	0.547747	0.335585	0.350754	0.645305

Intermediate conclusion:

- Both *hierarchical clustering* and *k-means* are promising to further evaluate.
- The other methods will not be included for further evaluation as they did not perform well
- Furthermore, *hierarchical clustering* and *k-means* are simpler to interpret and only have one hyper-parameter (i.e. the number of clusters)

Results 3 – Visualization with heatmaps:

- The goal eventually is to indicate the micro-crystal locations visually as a heatmap.
- Such visualization also allows for proper *human judgement* of the results.
- In Figure 37, the results of k-means and hierarchical clustering are show as heatmaps in comparison to the manual classification. What it also show, is that…
- …there are **errors in the human classification!**
- The visualizations also show what the algorithms did: they picked up the edge in the background (just like the human did when not paying enough attention)
- Also note that ***with crystal*** and ***partial crystal*** is **not always clear-cut**; looking at the patch marked with the purple dot: who is right, *k-means*, *hierarchical*, or the *human*?
- Now that our 'ground-truth' is not fully true and there is this ambiguity, **scoring becomes difficult**.

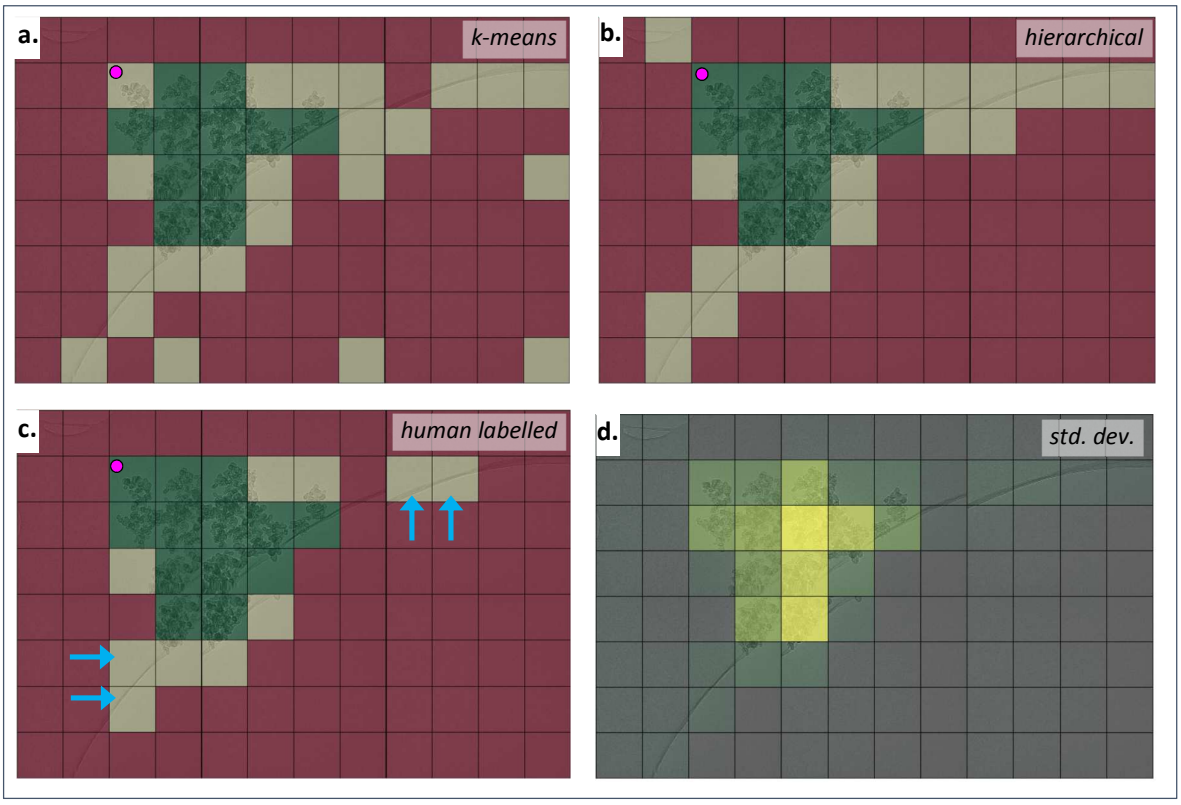


Figure 37 Heatmaps of the clustering results on data set 1 using different approaches. Note that the colors have been assigned manually (as cluster numbers from the algorithms are arbitrary). **a.)** The heatmap resulting from the *k-means* algorithm b.) The heatmap resulting from *hierarchical clustering* **c.)** The heatmap of the manual classification; looking closer, 4 patches are mis-labelled (indicated by the blue arrows), containing the edge in the background instead of part of a crystal. **d.)** The *standard deviation* of each image patch plotted as an overlay; the manual classification was done quickly and mostly based on this statistic; as a consequence, some patches were erroneously assigned to the category for 'partial crystal'.

Lessons Learned:

- Unsupervised learning with *hierarchical clustering* or *k-mean* works well on data set 1
- Using an automatic score for the quality assessment is problematic for a number of reasons.
- We need to fall back to human evaluation (supported by heatmap visualization);
- Track progress quantitively, would require counting squares each time; this is not viable.

12.8 Assessing the number of clusters (Aug 2018)

Goal:

Assess the influence of the number of clusters on the outcome of the unsupervised learning.

Results:

- On this data set, with hierarchical clustering, 3 clusters is the optimum (Figure 38)
- On the same data set, with k-means and a smaller patch size 2 clusters is the optimum.
- Remark: The dimensionality inspection 0 hinted towards using ***PCA*** to make the data better separable, hence I've added a ***PCA*** step before clustering. This improves the results for *k-means* a bit; for hierarchical clustering, hardly a substantial difference is noticed.

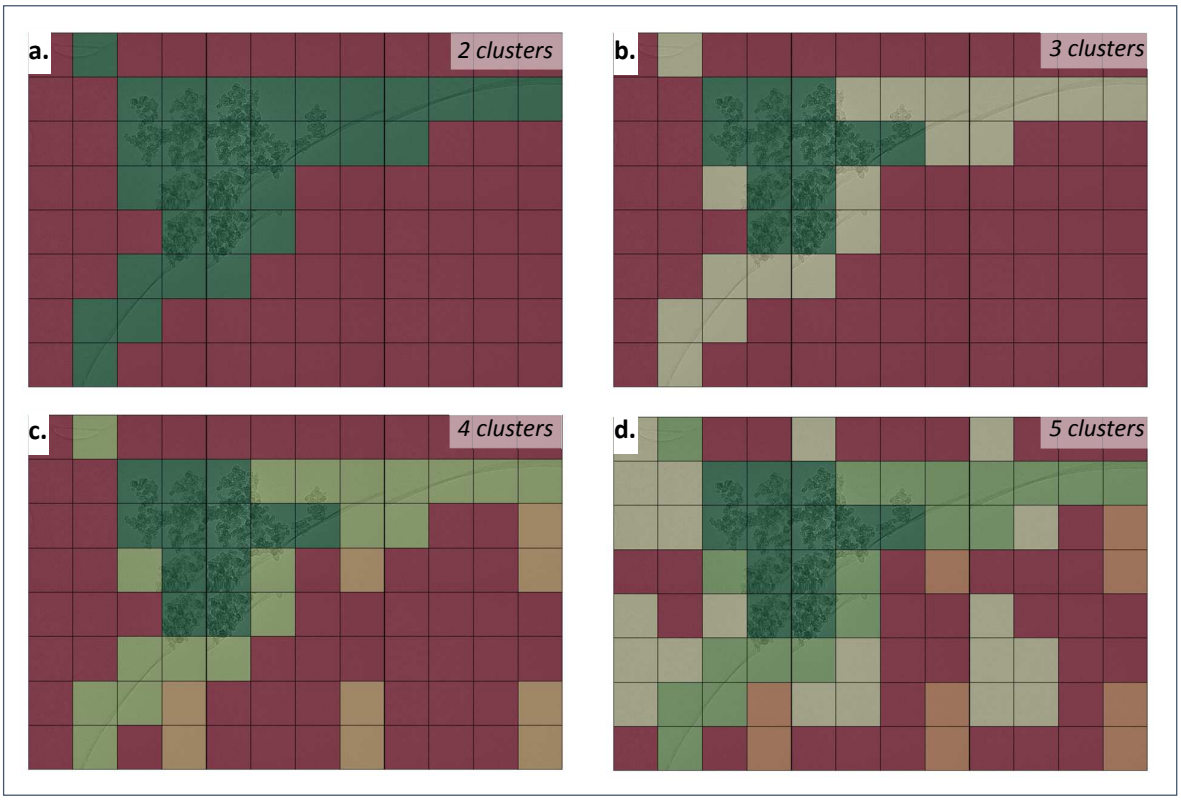


Figure 38 Varying the number of clusters for *hierarchical* clustering on data set 1. In this case, 3 clusters is optimal where only the green cluster should be used to indicate the micro-crystals.

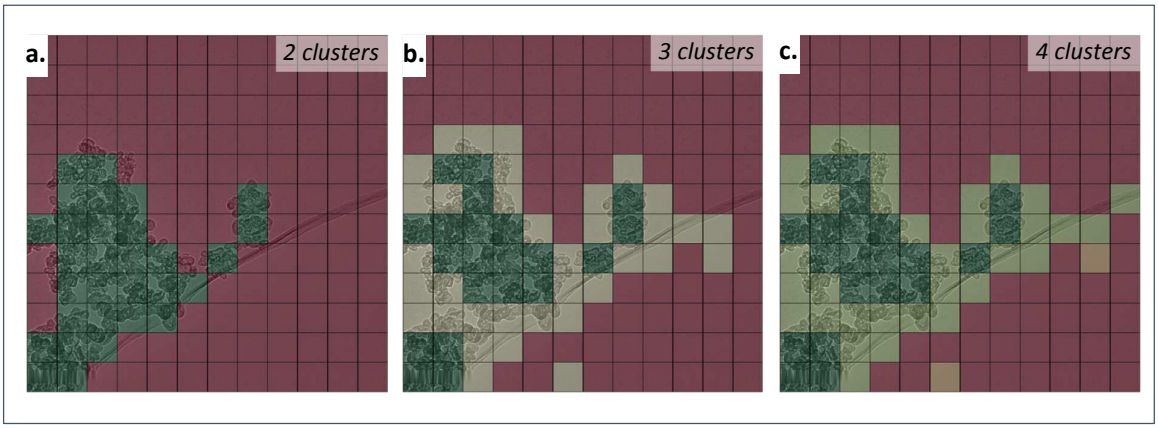


Figure 39 Varying the number of clusters for *k-mean* clustering on data set 1 with a smaller patch size, shows the most accurate results of 2 clusters are used.

Lessons Learned:

- What the optimum number of clusters is, depends on the algorithm
- For the problem at hand, either 2 or 3 work well.

12.9 Alternative Technique: NMF and Similarity

Goal:
Try out *NMF* and similarity based on *cosine distance* to identify regions with micro crystals

About the technique:

- *Non-negative Matrix Factorization* (*NMF*) plus *cosine distance is* a technique commonly used for topic modeling in *Natural Language Processing* (often referred to as *LSA* or *Word2Vec*).
- *NMF* resembles *PCA*, but constrains the components and matrix elements to be non-negative, which has some advantages. Similar to PCA, the feature vectors are transformed to a new basis.
- Consequently, the similarity of each data point to a reference data point is determined by calculating the dot product of the corresponding feature vectors (in the new basis).
- A good reading on the topic is Keita Kurita Blog [11]

The experiment:

1. Load in the extracted features from data set 1 (i.e. the image statistics of the patches)
2. Apply NMF on the extracted features]
3. Select a reference patch (in this case, one with a micro-crystal)
4. Calculate similarity of each data point with the reference patch and plot as a heatmap
5. Repeat steps 1-5 *without* the NMF transformation

Results:

- On this data set, NMF + similarity worked pretty well.
- Even without NMF – only using similarity of the feature vectors – gives good result
- This is interesting, as it does not require a (computationally intensive) learning algorithm

Lessons Learned:

- Feature vector similarity can also produce reasonable heatmaps.
- Due to its simplicity and low computation costs, this method is a valuable, quick first-order approach

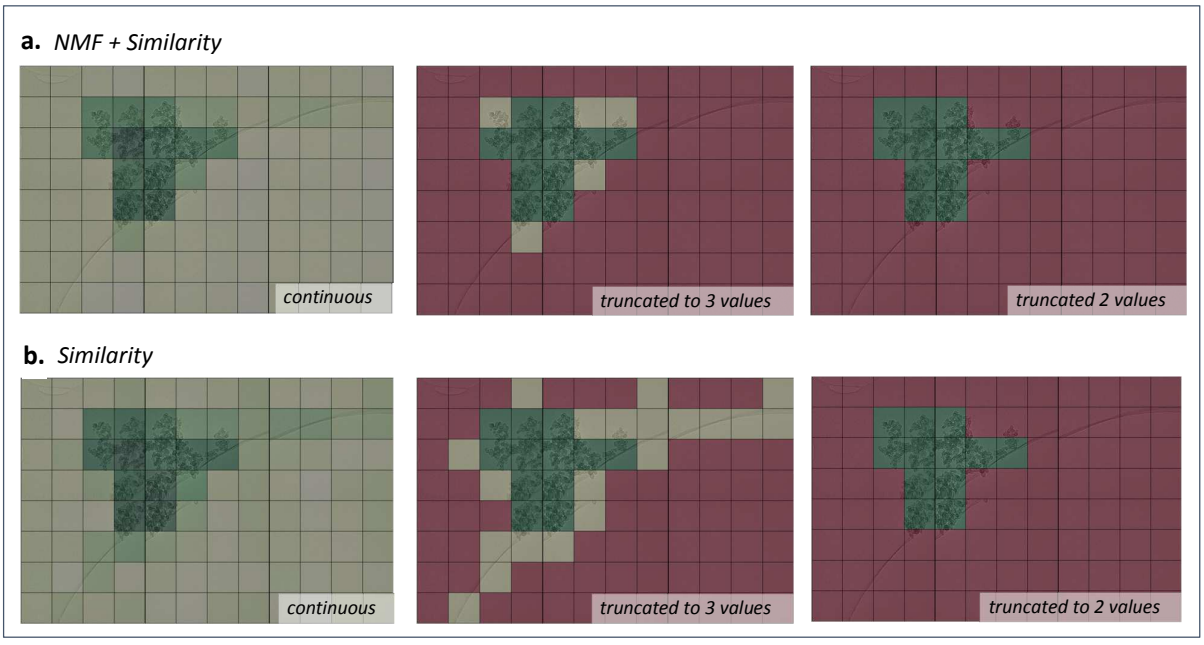


Figure 40 Heatmaps based on *feature similarity* with *NMF* **(b)** and without *NMF* **(b)**; the images on the left show the continuous similarity values, which can be any value between 0 and 1; the middle and the right images show the similarities when truncated to 3 or 2 values respectively, mimicking classification in 3 or 2 clusters.

12.10 The Full Pipeline (Aug 2018)

Goal:
Glue all pieces together into one pipeline script and run on multiple data sets

The experiment:
Described schematically in Chapter 6

Lessons Learned / Results:

- Successful runs on different data sets (Figure 41) as long as there are no black areas.
- Data sets with the black areas are still an issue, the micro-crystals are not detected (Figure 42)
- However, the algorithm does separate black areas and borders, which has value on its own
- With the full pipeline in place, it becomes easier to tune parameters (discussed in next paragraph)

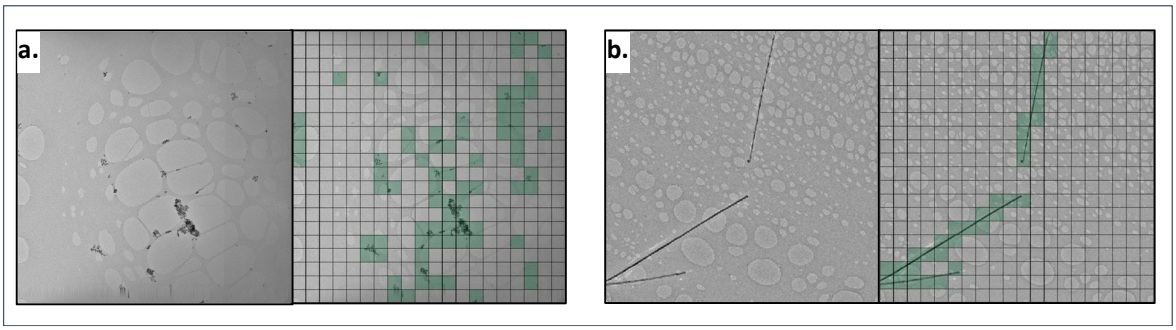


Figure 41 Results of running the full pipeline on different data sets. **a.)** the small, black particles are located in a single image of data set 2; **b.)** the elongated, rod-like particles are located in a single image of data set 3

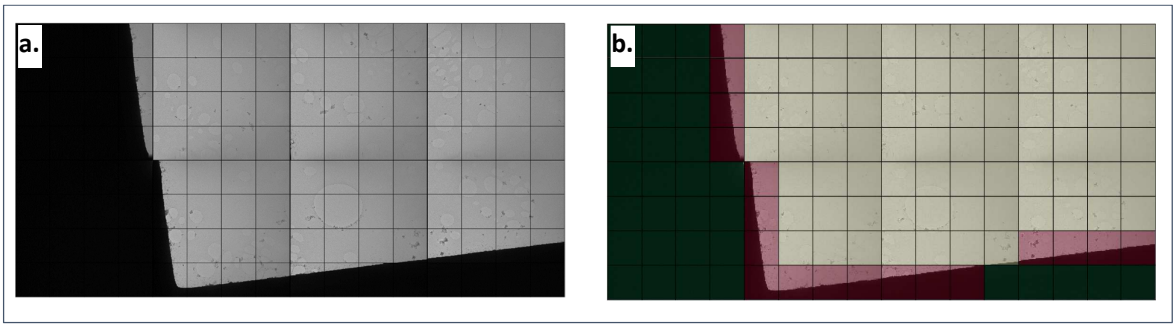


Figure 42 Results of the full pipe line on data set 2, which contains black regions. **a.)** The original data set, consisting of 4 x 2 images and image sliced into 4x4 patches. **b.)** The full pipeline is not able to identify the patches with particles, but does a good job in detecting the black patches and patches with a border.

12.11 METRICS & HYPER-PARAMETER TUNING (AUG 2018)

Goal:
Optimize results by tuning hyper-parameters: number of clusters, number of patches, number of PCA components.

About the metric:

- As discussed in paragraph 12.7, measuring the performance of unsupervised learning by a score is not without issues. If there is no ground-truth available, one has to use an *intrinsic score metric,* which evaluate how well the clusters are defined.
- For most of these metrics, when optimizing the score, one should plot the score vs. the hyper-parameter being optimized and look for the 'elbow' in the graph.
- Many such metrics exist; after a short evaluation, I used the **Silhouette Coefficient**.

The experiment:

1. a. Run the full pipeline on data sets 1 multiple times, varying the number of clusters for each run
   b. Repeat 1a. for different number of patches
2. a. Run the full pipeline on data sets 1 multiple times, varying the number of patches for each run
   b. Repeat 2a. for different number of clusters
3. Run the full pipeline on data set 1 multiple times, varying the number of *PCA Components*

Results 1 - Number of clusters:

- The curves of the Silhouette score vs the number of clusters are not very smooth, see Figure 43
- This makes it hard to identify the 'elbow'; depending on number of patches it could be 3, 4 or 5.
- Visual inspection (see paragraph 12.7) indicated the real optimum is at 2 or 3 clusters
- Hence, the applicability of this optimization is limited.

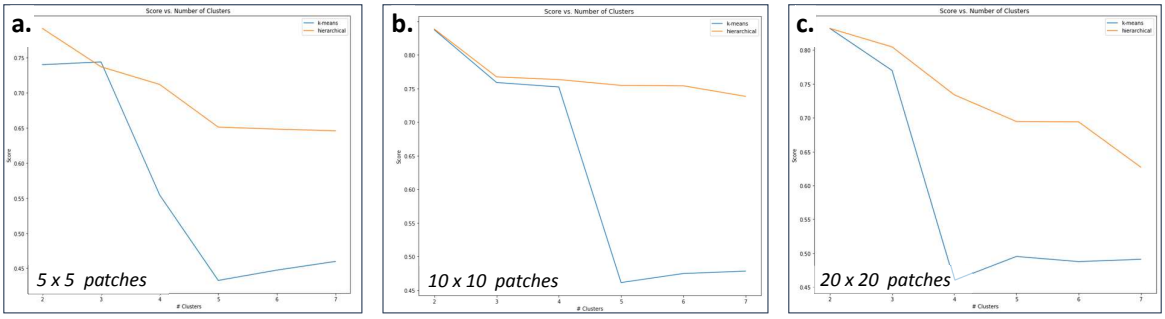


Figure 43 Optimizing the 'number of clusters' on data set 1 for *k-means* and *hierarchical clustering ;* the graphs show the *Silhouette* score vs the number of clusters for varying number of patches **a.)** 5x5 patches; **b.)** 10x10 patches; **c.)** 20x20 patches.

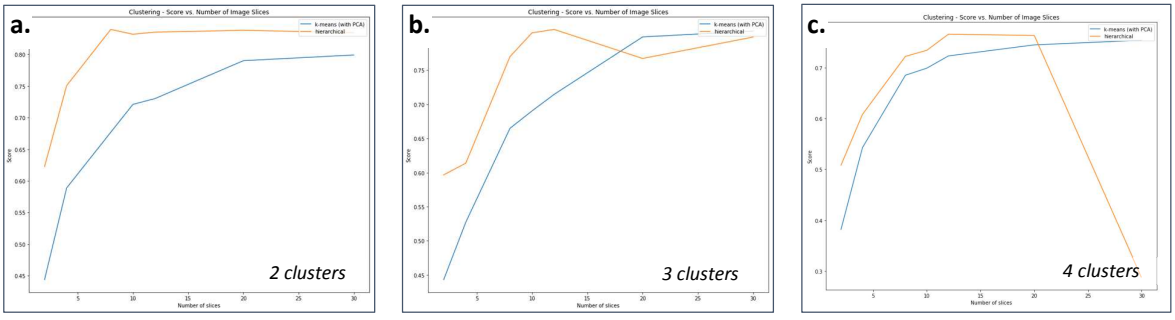


Figure 44 Optimizing the 'number of patches on data set 1 for *k-means* and *hierarchical clustering ;* the graphs show the *Silhouette* score vs the number of patches for varying number of clusters **a.)** 2 clusters; **b.)** 3 clusters; **c.)** 4 clusters.

Results 2 - Number of patches:

- The curves of the Silhouette score vs the number of patches are shown in Figure 44
- The 'elbow' in the plots is either at 10 or 20 patches, depending on number of clusters
- *Hierarchical clustering* can break down for large number of patches (= large number of data points), which did happen in Figure 44c . The algorithm is known not to scale up very well.
- It is debatable if this way of optimizing the number of patches is valid: when increasing the number of patches per image, the total number of data points increases (for each patch, image statistics are extracted), while the patches themselves become smaller (resulting in other statistics).
- In other words: **when changing the number of patches, the data is modified** that is fed into the clustering algorithm.
- Hence, this optimization is of limited value, which is confirmed by inspecting the heatmaps: the hierarchical clustering becomes *worse* for higher number of patches (while its score increases)

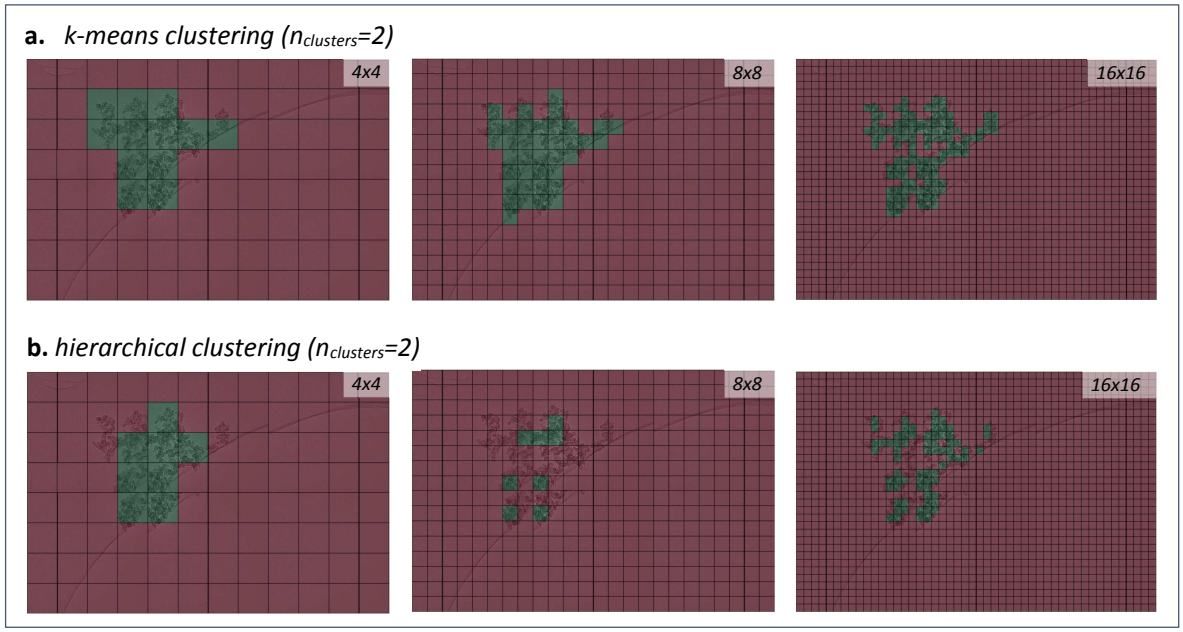


Figure 45 Visually showing the difference between *k-means* **(a)** and *hierarchical clustering* **(b)** when optimizing the number of patches. For the *hierarchical clustering*, the visuals for 8x8 and 16x16 patches per image like worse than the 4x4, although their clustering score in Figure 44 is higher.

Results 3 - Number of PCA Components (*k-means* only):

- The plot in Figure 46a show the *Silhouette* score goes down for higher number of components! This can be explained by how the score is calculated. The score is evaluated on the data that goes into the clustering algorithm, which is the *PCA* transformed data. If the number of *PCA* components is limited, the dimensionality of the data is reduced. Hence, comparing the scores is comparing data of different dimensionality! Apparently for this data set, clustering is strongest in the lowest dimensions, which are the direction in which there is the highest variability (*PCA* orders the components based on variance).
- Based on the graph, one would also expect that even a single *PCA* component is sufficient to separate this data set in two clusters. This is indeed the case, as shown visually in Figure 46b. This is not really surprising: for this data set, we observed in paragraph 12.2 that a single (hand-picked) statistic could already separate the data to a large extend.

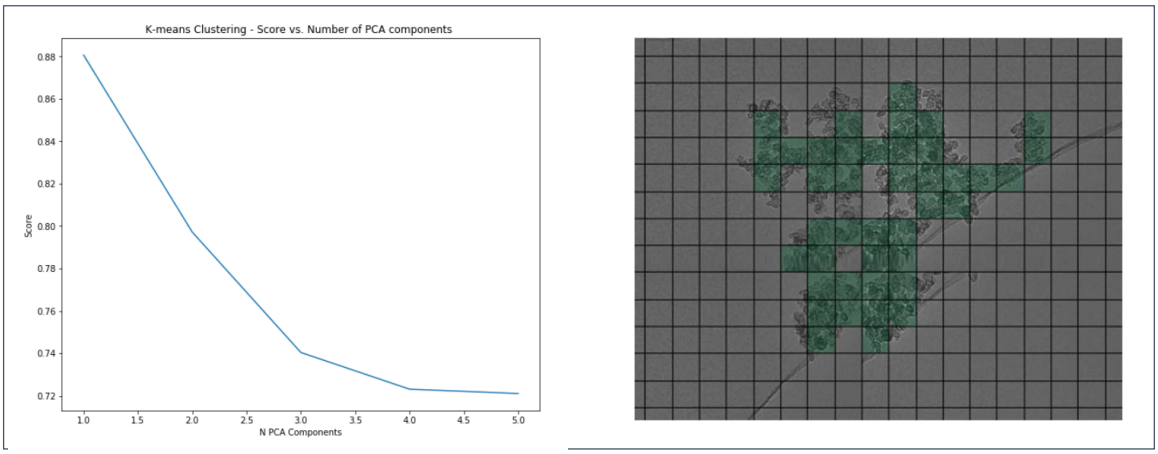


Figure 46 Tuning number of PCA components. In this set, even 1 component is sufficient. This is however an 'easy' set where a single hand-picked statistic (i.e. 'standard deviation') was already close to identify the crystals.

Lessons Learned:

- Evaluating scores of unsupervised clustering methods without a 'ground-truth' is far from trivial.
- One should be careful to rely on these techniques and jump to wrong conclusions.
- Visualization and human inspection remain key for validation.

12.12 Variant: Full Pipeline with 'sliding window' (Sept 2018)

Goal:

Evaluate if 'sliding windows' variant of the ML pipeline results in a finer localization.

About the sliding window variant:

In this variant, the patches from the fixed grid are replaced by patches that are extracted at each pixel and its neighborhood. It is as if a window with the size of a patch, slides over the image pixel-by-pixel.

Results:

- The sliding window was applied on single images only, as this is computationally expensive.
- As visible in Figure 47, this approach is able to pinpoint the micro-crystals accurately.
- On images with black border – which were difficult with the prior technique – particles are also detected. However, some boundaries are erroneously also classified as particle (Figure 47c).
- This approach results in a pixel-by-pixel cluster assignment, which can be plotted as a heatmap, but also as individual images per cluster (as shown in Figure 48.)
- The number of clusters used with this technique is more critical (illustrated in Figure 48).

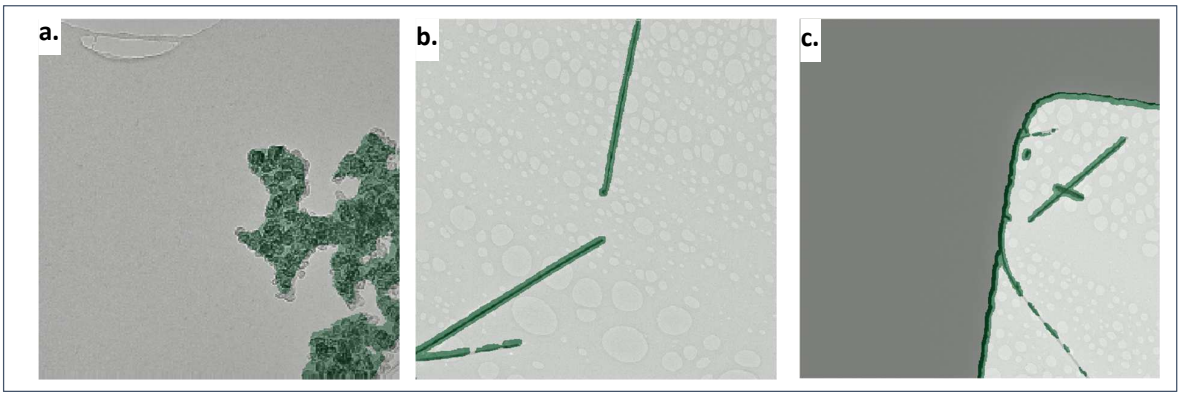


Figure 47 Result of 'sliding window' variant on single images from data set 1 **(a) ,** data set 3b **(b)** and dataset 3a **(c).** On all images, the micro-crystals are located accurately, but in **(c)** also the edge is picked up in the same cluster as the particles.

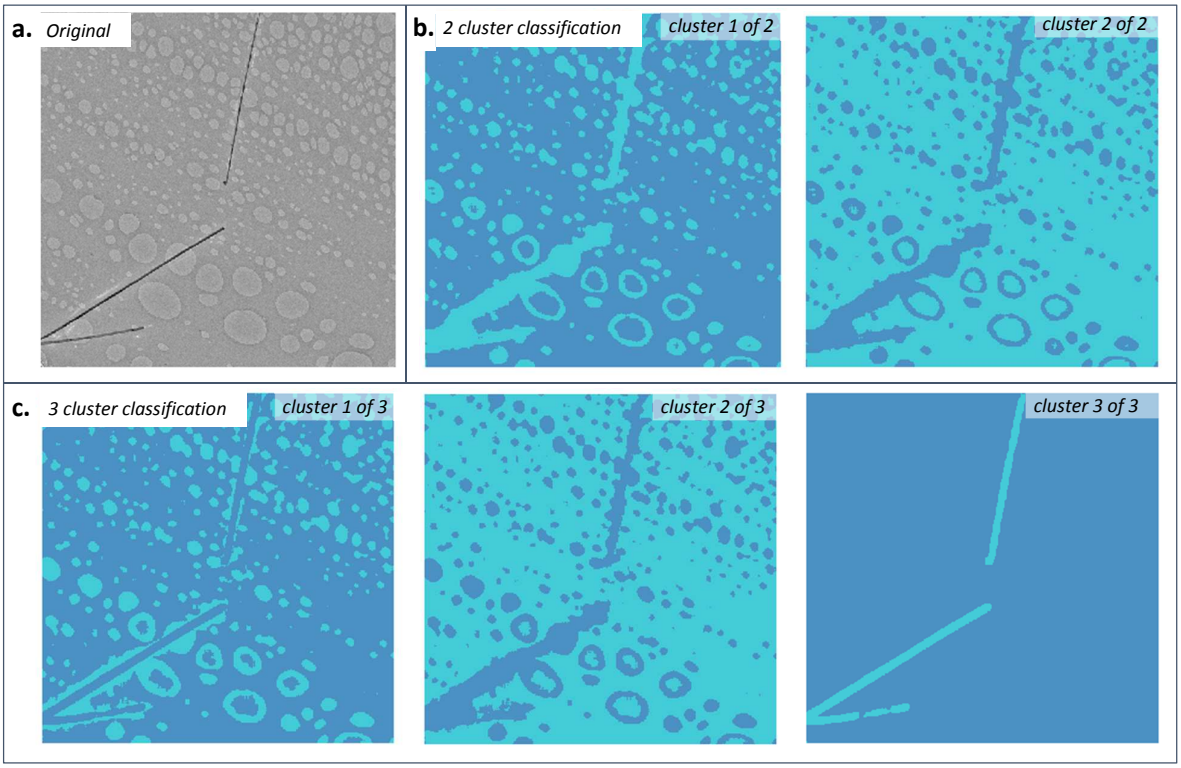


Figure 48 Result of 'sliding window' variant on data set 3b. For this particular image **(a)**, with 2 clusters **(a)** the rod like crystals are not separated from the circular features in the background, but with 3 cluster **(b)** they are isolated into cluster 2.

Lessons Learned:

- The 'sliding window' allows localization of crystal with high granularity
- This comes at a cost of extra computation, making it less suitable for large image sets
- It was better able to find particles in images with black areas. This could be an indication that the non-detection with the other method has to do with the granularity.
- The method is more sensitive to the parametrization, like the number of clusters to search for.

12.13 ANOTHER ATTEMPT AT THE DIFFICULT SETS (OCT 2018)

Goal:

Revisit the data sets that contain black areas, using the new insights and applying the full pipeline.

The experiment:

1. Run the full pipeline on data sets 2 and dataset 3a with smaller patch size and varying number of clusters
2. Try a two-step approach on these data sets, first filtering out the black regions, followed by a search in remaining area
3. Try the pipeline on the difficult data set 4, in which the particles are hardly visible

Results

- On the two data sets with black regions (dataset 2 and dataset 3a), using 4 clusters and small patch size, the particles could be found; see Figure 49 and Figure 51
- Alternatively, a two-step approach was used to first filter out the black and partial black regions using 3 clusters, followed by a search for 2 clusters in remaining area; see Figure 50 and Figure 52
- With both techniques, there are some false positives close the black region. This may be circumvented by a smarter filtering (e.g. also excluding the first patches next to the border)
- On data set 4, which has black regions and particles that are hardly distinguishable from the background, limited success was achieved, but only when carefully selecting the image region.
- With this 'sliding window' approach, results are better, but only after tweaking the patch size.
- On such set, the technique is not yet reliable, but it shows that in principle it is possible.

Lessons Learned:

- Having results on multiple data sets now, shows the technique is really viable
- For larger sets, performance starts to become an issue, but mostly for the visualization

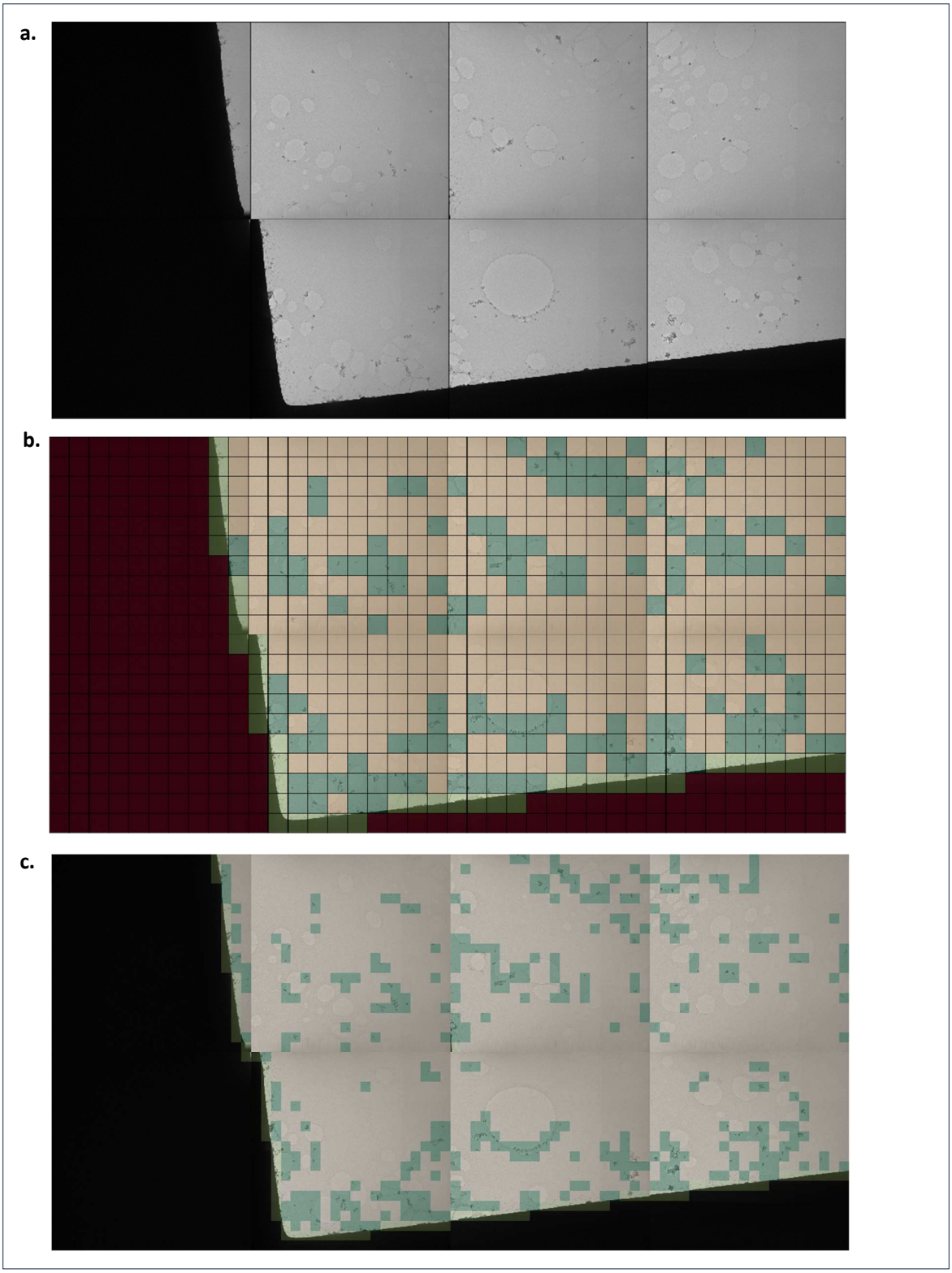


Figure 49 Results of the full pipeline, searching for **4 clusters** in data set 2. **a.)** the original 4 x 2 images; **b.)** using a granularity of 10x10 patches per image; **c.)** using a higher granularity of 20x20 patches per image (and slightly different visualization).

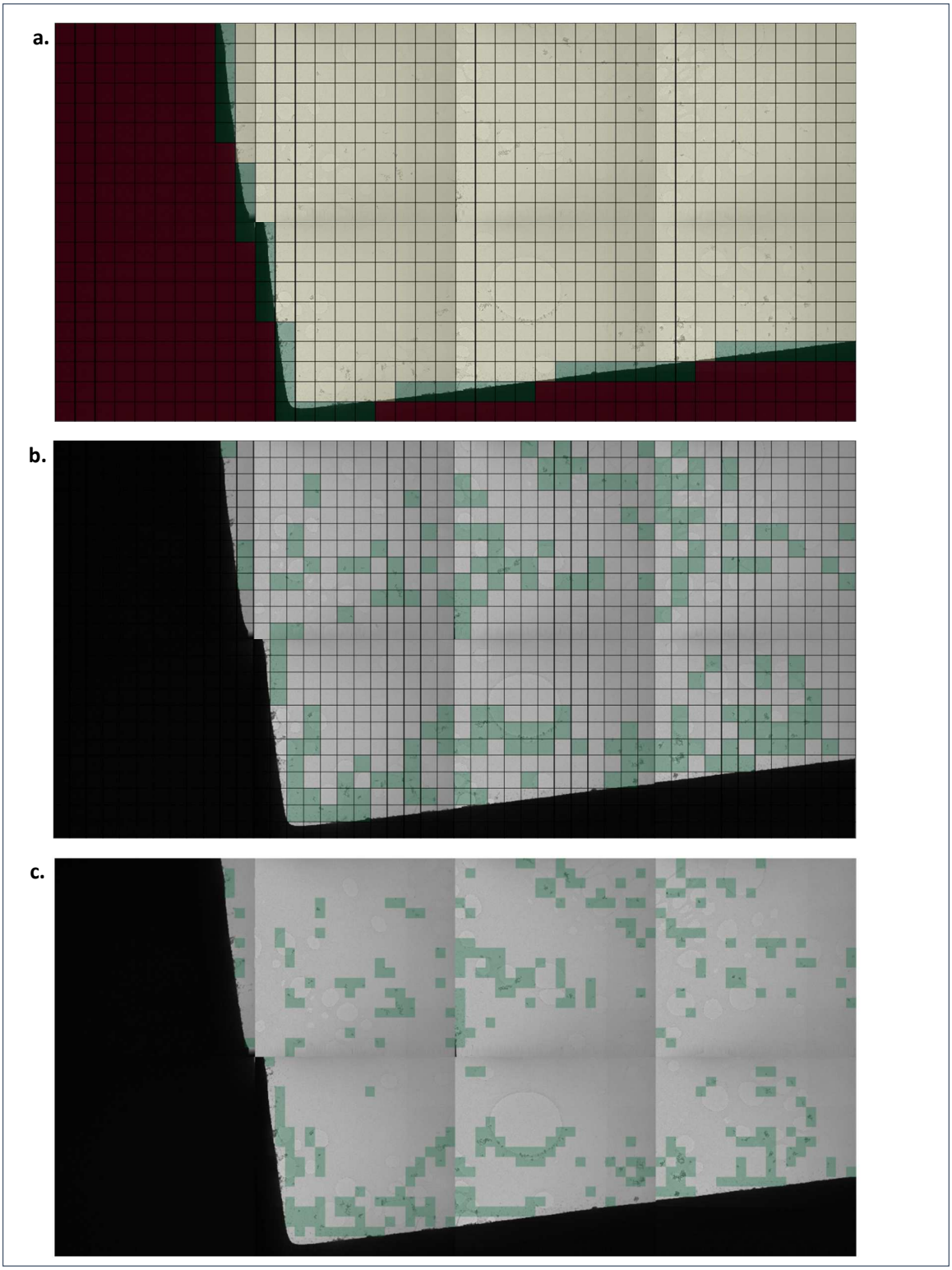


Figure 50 Results of the 2-step approach on data set 2. **a.)** In the first step, searching for 3 clusters to separate the areas with clusters (yellow) from the black regions; **b.)** after filtering out the black regions, using a granularity of 12x12 patches per image **c.)** using a higher granularity of 20x20 patches per image after the filtering (and slightly different visualization).

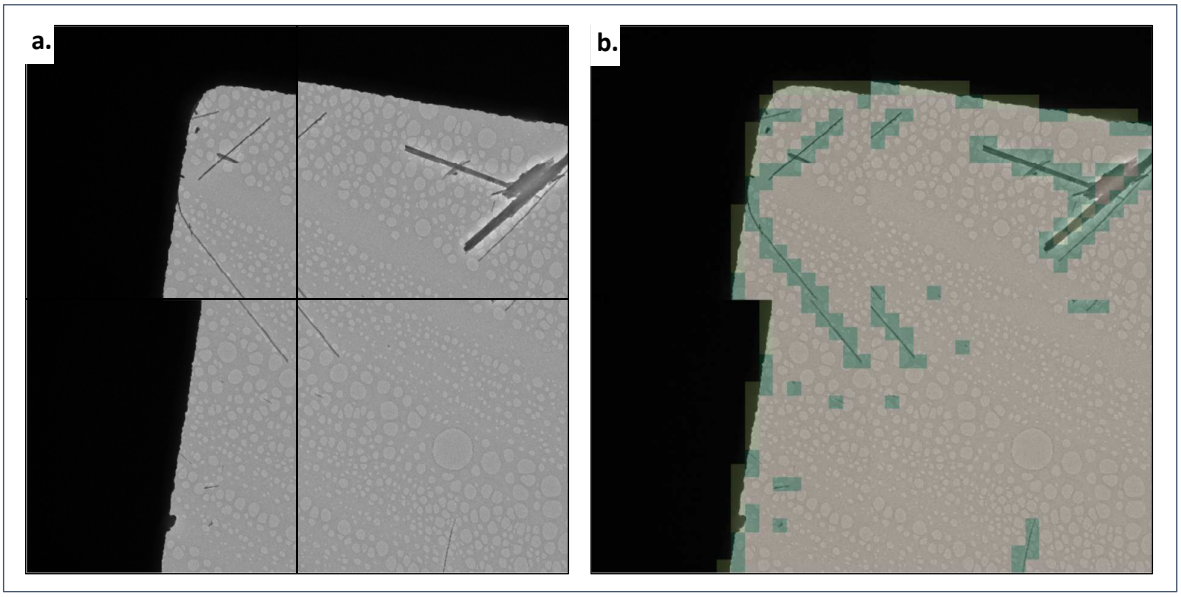


Figure 51 Results of the full pipeline, searching for **4 clusters** in data set 3a. **a.)** the original 2 x 2 images; **b.)** using a granularity of 20x20 patches per image;

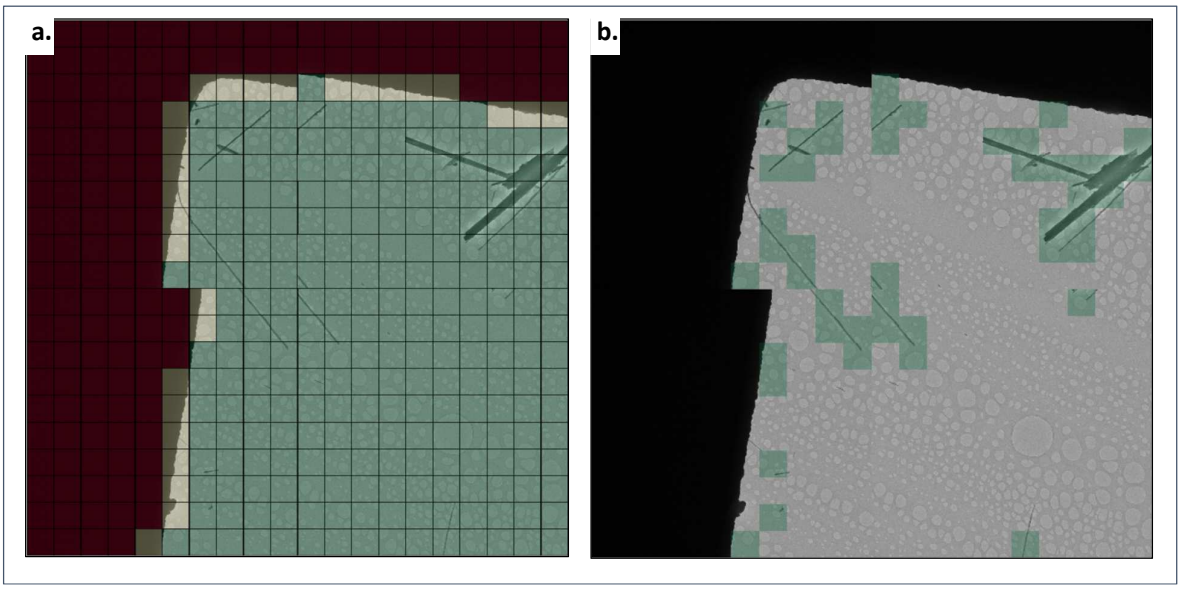


Figure 52 Results of the 2-step approach on data set 3a. **a.)** In the first step, searching for 3 clusters to separate the areas with clusters (green) from the (partial) black regions (dark-red, yellow); **b.)** after filtering out the black regions, using a granularity of 10x10 patches per image finds the elongated, rod-like particles, but also false positives near the edge.

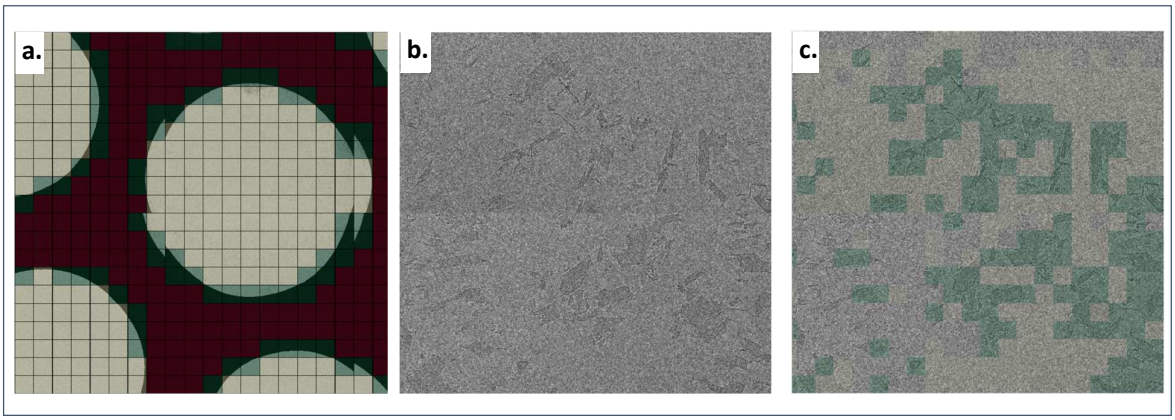


Figure 53 Results on data set 4, which only worked after extra preparation and tuning. **a.)** First, using 3 clusters to filter out black areas. **b.)** As only some part of the non-black areas contained visible micro-crystals, a manual cut-out was made; **c.)** using a search of 4 clusters on this cut-out, had limited success in identifying where the micro-crystals are.

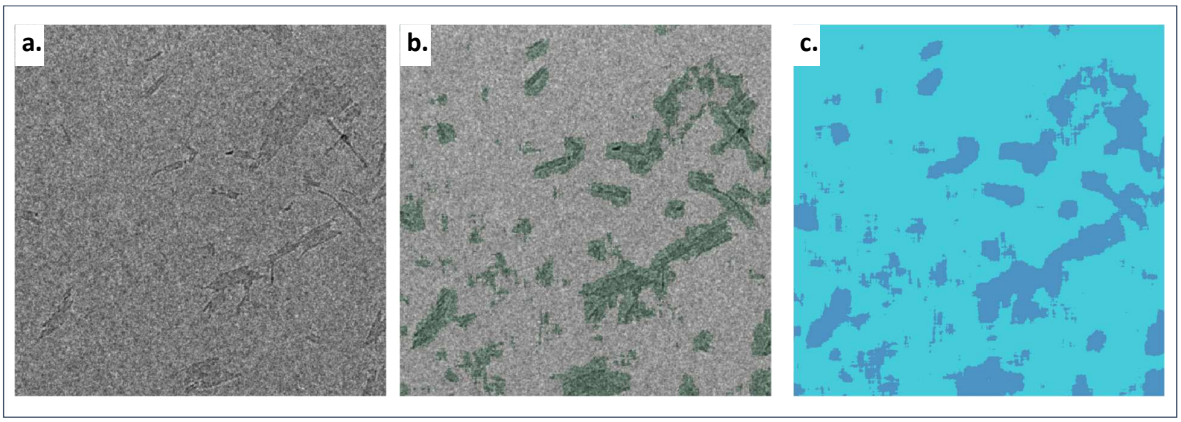


Figure 54 Results on data set 4 with the 'sliding window' approach, which only worked after extra preparation and tuning. **a.)** A cut-out of a non-black areas which contains visible micro-crystals **b.)** the result of the search for 2 clusters with the 'sliding window', but after carefully tuning the window-size **c.)** the cluster image containing the micro-crystals.

12.14 Wrapping up & Writing Report (Oct/Nov 2018)

I improved some of the scripts in order to re-run experiments with better visualization. The result of all those efforts are found in this report!

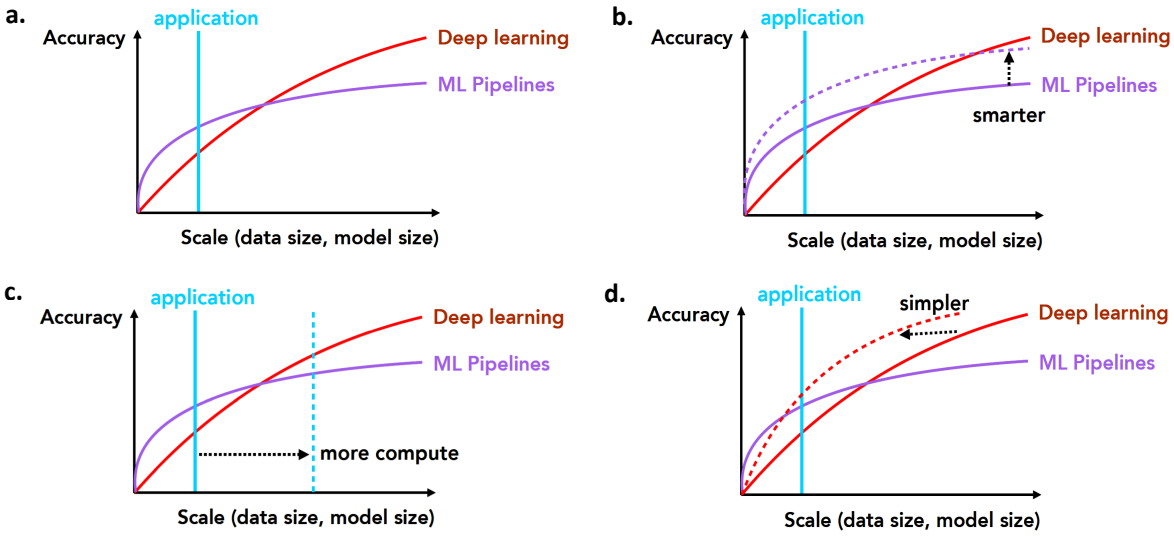
Lesson Learned: With the full pipeline in place and all insights gained, the micro-crystals can now be located in a much faster and smarter way!

LIST OF REFERENCES

1. Electron Microscope Market Size, Share & Trends; Grand View Research
   https://www.grandviewresearch.com/industry-analysis/electron-microscopes-market
2. Electron Microscopy and Sample Preparation Market; MarketsAndMarkets Research
   https://www.marketsandmarkets.com/Market-Reports/electron-microscopy-sample-preparation-market-47729204.html
3. MicroED opens a new era for biological structure determination, B. L. Nannenga & T. Goonen
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5656569/
4. Advances in Cryo-Electron Microscopy Will Improve the Global Protein Crystallization and Crystallography Market Through 2020, Business Wire
   https://www.businesswire.com/news/home/20160307005594/en/Advances-Cryo-Electron-Microscopy-Improve-Global-Protein-Crystallization
5. Crystal structure analysis of pharmaceuticals with electron diffraction, Dr. S. Nicolopoulos
   http://www.icdd.com/ppxrd/12/presentations/P11-Stavros-Nicolopoulos-ppxrd-12.pdf
6. Protein Crystallization and Crystallography Market - Global Opportunity Analysis and Industry Forecast 2017-2023, Allied Market Research
   https://www.alliedmarketresearch.com/protein-crystallization-and-crystallography-market
7. Crossing the Chasm, Marketing and Selling High-Tech Products to Mainstream Customers Geoffry A. More, 1991/1999/2014
8. The Lean Canvas by Ash Maurya, an adaption of Alex Osterwalder's Business Model Canvas
   https://leanstack.com/leancanvas
9. An Introduction to Statistical Learning, Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani; 7th printing (2013);
   https://www-bcf.usc.edu/~gareth/ISL/
10. Wikipedia on Feature Scaling
    https://en.wikipedia.org/wiki/Feature_scaling
11. Introduction to NMF, Keita Kurita
    http://mlexplained.com/2017/12/28/a-practical-introduction-to-nmf-nonnegative-matrix-factorization/
12. Scikit-learn - Machine Learning in Python.
    General: www.scikit-learn.org
    About clustering: https://scikit-learn.org/stable/modules/clustering.html
13. Textural Features for Image Classification, Haralick, R. M., Shanmugam, K. & Dinstein, I.
    IEEE Transactions on Systems, Man, and Cybernetics 3, 610–621,
14. Github Repository of results, source code and sample data of the work presented in this report
    Link: https://github.com/mpjanus/jadsproj

Appendix A ML PIPELINES VS DEEP LEARNING

A.1 MODEL ACCURACY VS SCALE



Courtesy of Olivier Bousquet

Figure 55 Different machine learning strategies, depending on application and amount data (courtesy of Olivier Bousquet). a.) Given a specific application, when amount of data or model size is limited, a ML pipeline generally performs better than a Deep learning network; b.) accuracy can by improving by using smarter algorithms in the ML pipeline. c.) if more data is available, a Deep learning network can be utilized to obtain higher accuracy; d.) alternatively, if not more data is available, a simplified Deep learning network may outperform a ML pipeline.

A.2 SELECTING MACHINE LEARNING ALGORITHMS

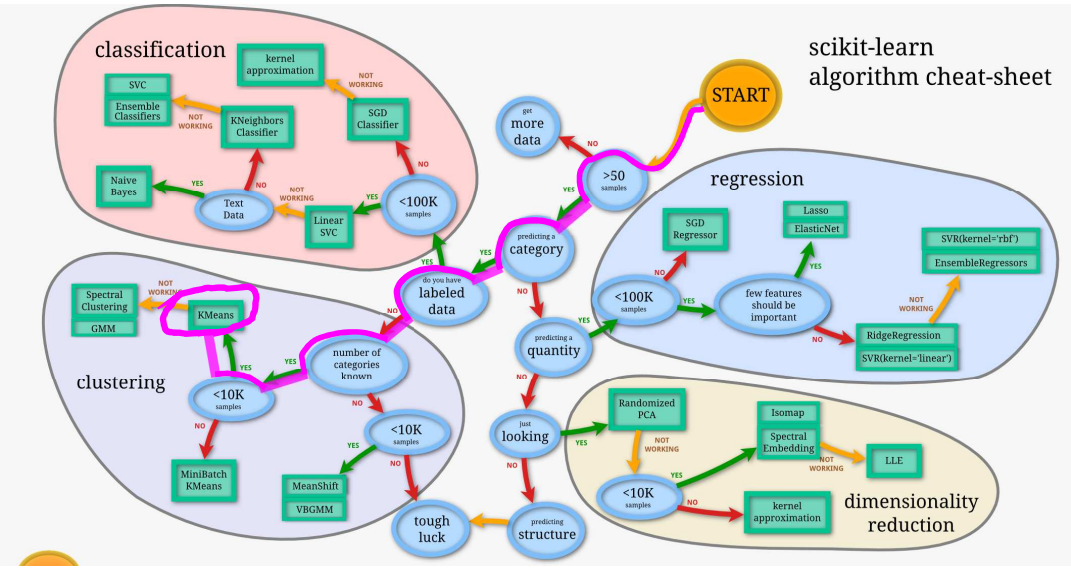


Figure 56 Guidance from Scikit-Learn on choosing a method for the problem at hand. The path that is applicable for the 'Micro-Crystal Locator' is highlighted in purple.

Appendix B THE DATA SETS

All data sets used in this work are based on images acquired with a transmission electron microscope (TEM) on non-confidential test specimens. Courtesy to Bart Buijsse of ThermoFisher.

B.1 DATA SET 1 (FLAKES OF NANOCRYSTALLINE GRAPHITE AT HIGH MAGNIFICATION)

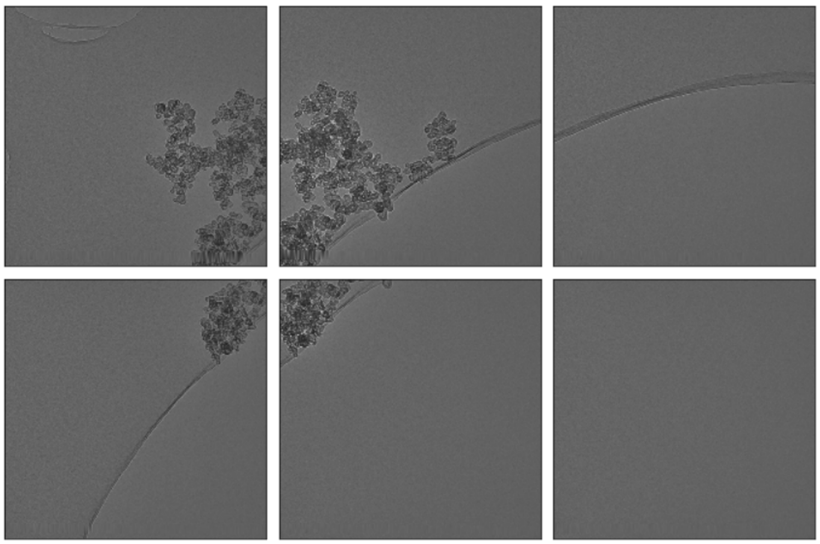


Figure 57 Data set 1, consisting of 3 x 2 images of 4k x 4k pixels. The images of the micro-crystals were acquired on a transmission electron microscope at ~ 30 000x magnification. These micro-crystal are flakes of nanocrystalline graphite.

B.2 DATA SET 2 (FLAKES OF NANOCRYSTALLINE GRAPHITE AT MEDIUM MAGNIFICATION)

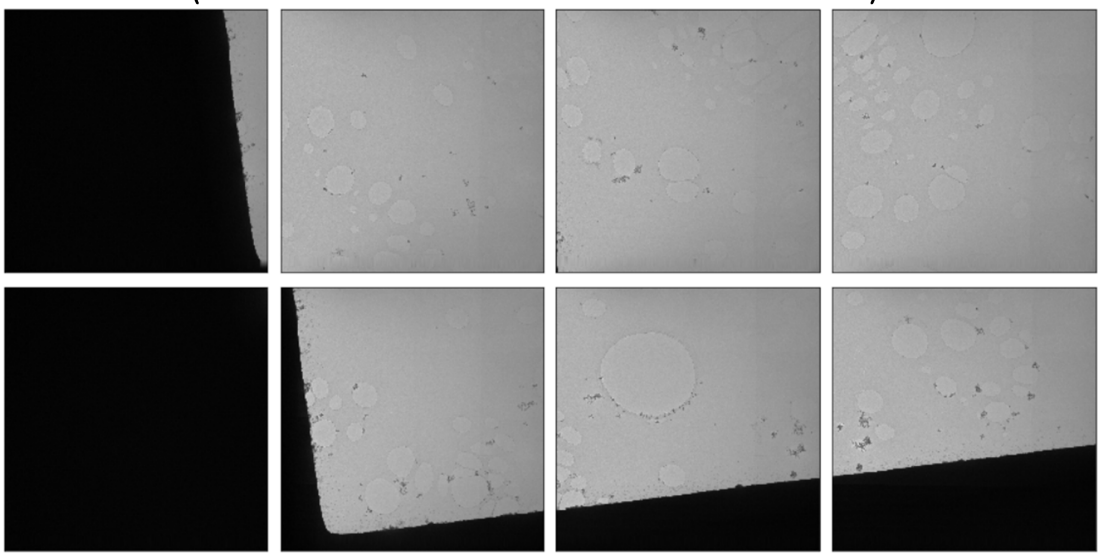


Figure 58 Data set 2, consisting of 4 x 2 images of 4k x 4k pixels . The image were acquired on a transmission electron microscope at ~ 3000x magnification. The micro-crystals are the small black dots and are of the same type as data set 1 (but at lower magnification). The black areas are 'grid bars' (see glossary).

B.3 DATA SET 3 (ASBESTOS FIBERS AT MEDIUM MAGNIFICATION)

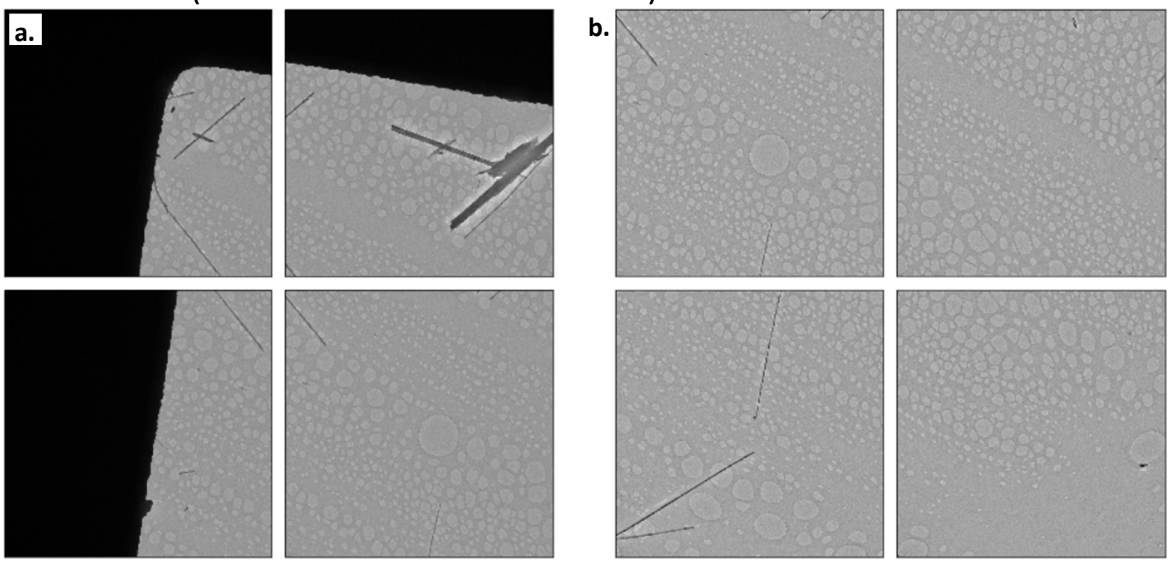


Figure 59 Data set 3 a and b, both consisting of 2 x 2 images of 2k x 2k pixel. Both sets are taken at ~ 3000x magnification, but at different locations on the specimen. The micro-crystals in these images are asbestos fibers. The black areas in a. are 'grid bars' (see glossary)

B.4 DATA SET 4 (POLYMERS AT LOW MAGNIFICATION)

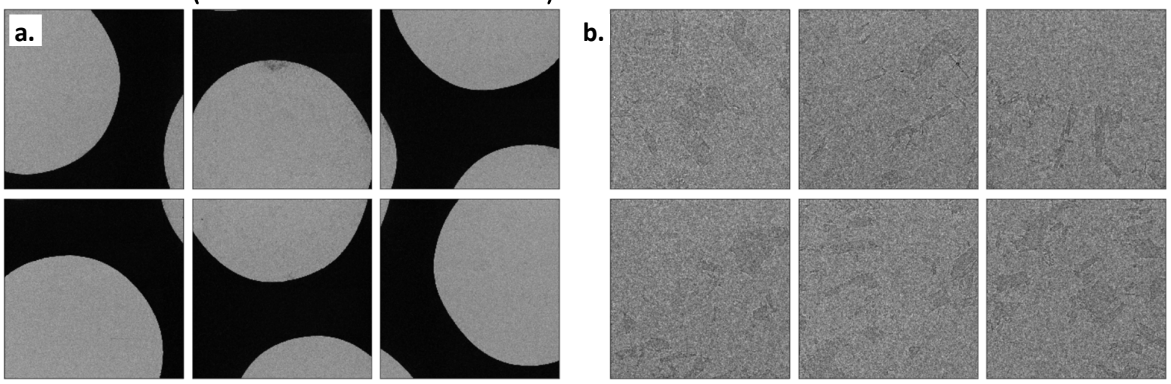


Figure 60 Data set 4, consisting of 3 x 2 images of 2k x 2k pixels. Both a. and b. are taken at ~ 150x magnification. b are cut-out images from a., away from the black 'circular grid'. The micro-crystals in these images are of a polymer type and hard to image (barely visible)