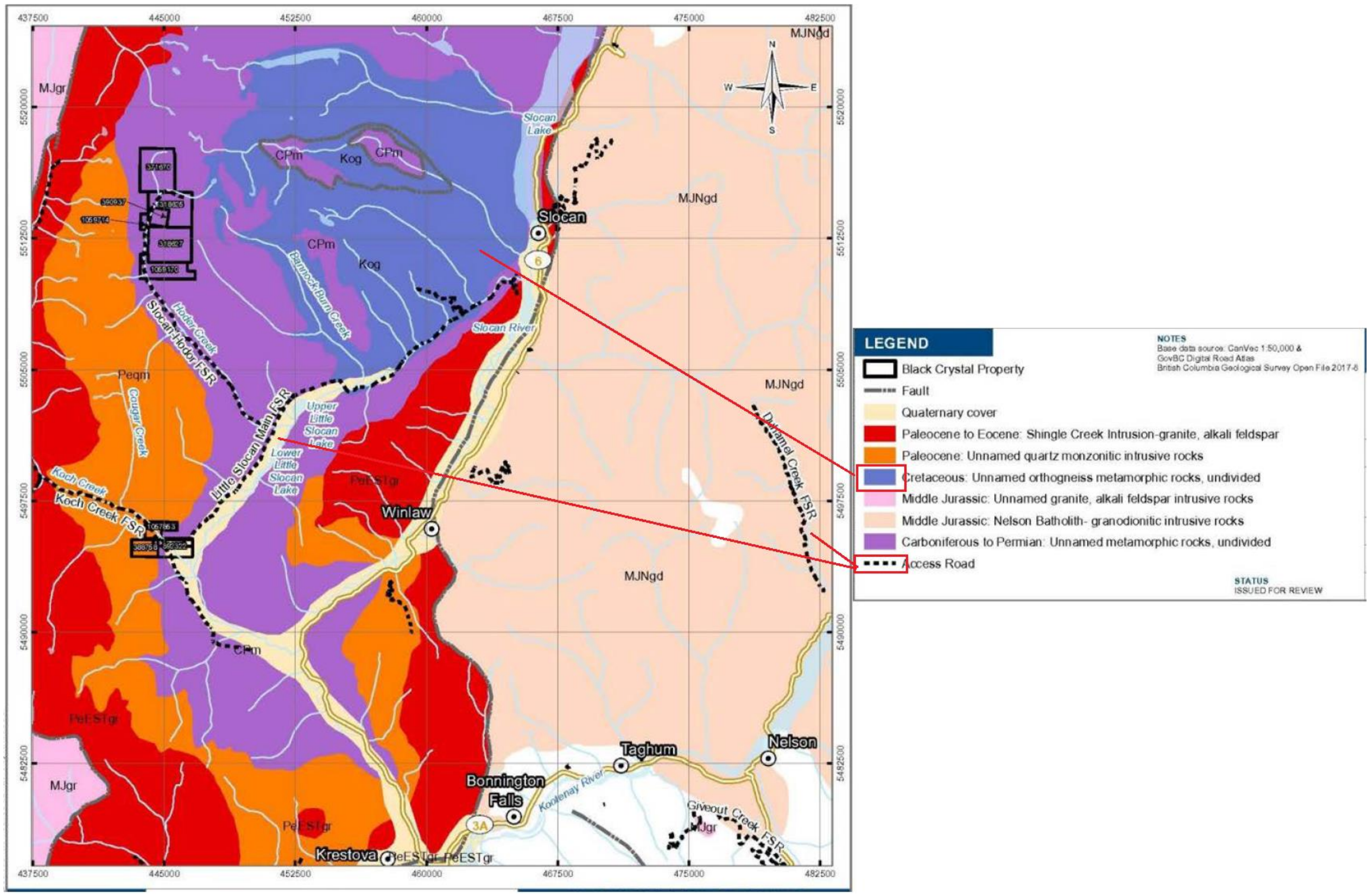
The Map Feature Extraction Challenge

Challenge Overview

This challenge seeks to find innovative approaches to **automatically** extract features in scanned maps or raster images, based on the symbols present in the legend. Currently, the United States Geological Survey (USGS) manually performs this task for almost every map in a mineral resource assessment (for more information on mineral resource assessments, please visit the Background page of the competition website: https://CriticalMinerals.darpa.mil/Background.

While there are tools to assist with this process (e.g., GIS software), it is still primarily done manually by tracing the important points, lines, and polygons in a map to generate vector features and relating these back to their respective legend descriptions. Where automated solutions do exist, they are typically tailored for a specific map feature and use simple approaches that do not work in more challenging cases (for example, identifying polygon features using simple color matching). This does not sufficiently address the wide range of maps that USGS must contend with. To develop a *generic* approach to this challenge, it is anticipated that machine learning techniques, including computer vision, will be required.



The map above contains polygon and line features that can be extracted from the map. These features include subtle difficulties, such as disjoint polygon features and line features and basemap features that divide and cross over multiple polygons.

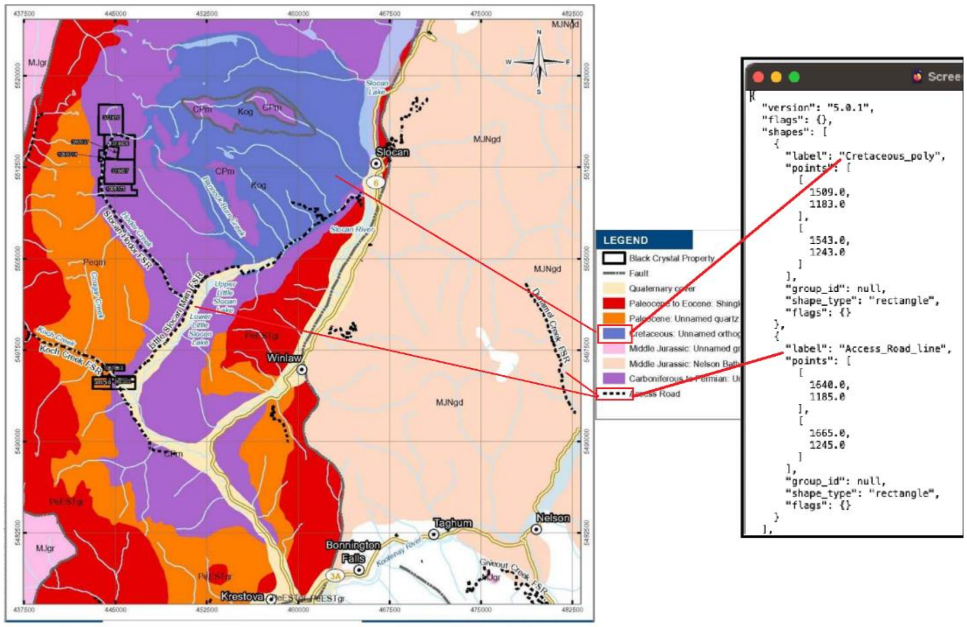
Challenge Details

The goal of this challenge is to develop methods capable of detecting and extracting specified features of a map by using the map legend as reference. Vectorization is the process of detecting the features from a raster or projected scanned map and then converting them into a spatial data format, such as shapefiles or feature classes. For maps of interest to USGS, this process can take weeks just for a single map. Applying this manual process to thousands of scanned/raster maps, each with dozens of distinct features, is impossible.

Research has been done on classifying one or many pre-defined features on maps, but such models need a large number of annotated images, and a new model must be created for each new feature. Additionally, it is generally not possible to automatically create a universal identifier for a specific feature that reappears across different images. For example, the lines showing fault lines on a geologic map can be depicted on different maps using different symbols, colors, patterns, and line types. To solve this challenge at scale, solutions will need to generalize to identify and vectorize features based on varying legends symbols.

For this challenge, you will be given an initial training set of data that will consist of map images with a number of specific legend items annotated (and classified as either point, line, or polygon) and the map feature corresponding to each legend item specified uniquely with a separate binary raster file. The raster files contain a binary array of 1s and 0s, where the selected point, line, or polygon features are represented by 1s (a point is represented by a single pixel and the width of the lines is also a single pixel).

Performers will also be provided with a json file for each map image, which specifies the pixel coordinates of the legend symbol for each of the features of interest in that map.



```
{
  "version": "5.0.1",
  "flags": {},
  "shapes": [
    {
      "label": "Cretaceous_poly",
      "points": [
        [
          1509.0,
          1183.0
        ],
        [
          1543.0,
          1243.0
        ]
      ],
      "group_id": null,
      "shape_type": "rectangle",
      "flags": {}
    },
    {
      "label": "Access_Road_line",
      "points": [
        [
          1640.0,
          1185.0
        ],
        [
          1665.0,
          1245.0
        ]
      ],
      "group_id": null,
      "shape_type": "rectangle",
      "flags": {}
    }
  ],
```

The image above shows an example of how the legend feature bounding box coordinates will be represented in the json file for each basemap.

In addition to the training data set, you will also be given a validation data set, which will include a separate set of map images and corresponding json files. This challenge is focused on the vectorization step so performers will be provided with the pixel coordinates of the legend symbols for the features of interest for all maps in both validation and evaluation datasets, but *without* the corresponding binary raster files.

For the final evaluation, you will be given a new set of map images and corresponding json files. You will be expected to provide a zip file with predicted raster files for each legend feature included in the json file that goes with each map. Please see the **Validation and Evaluation Details** section of this document for detailed submission instructions.

Challenge Dataset

The maps in the dataset contain three types of features:

- Points, represented by any number of different symbols
- Lines, which can be distinguished by attached symbols, bolding, dashing, color, etc.
- Polygons, which are generally distinguished by color and/or patterns.

While various map types will be included in the challenge, USGS has particular interest in extracting point features from topographical maps and polygon, line, and point features from geological maps, so these will be heavily represented.

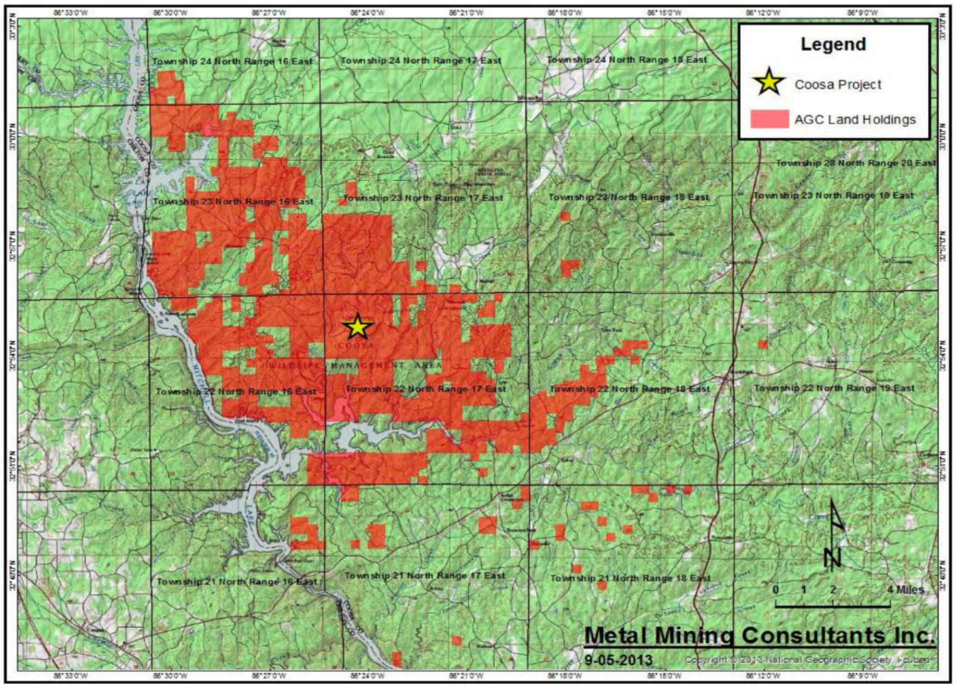
For line features in geological maps, USGS is only interested in “fault line” features so these will be the only line-type features annotated in these maps. Performers should be aware that in many cases, a continuous fault line on a map may alternate between being dotted or solid, but should be vectorized as a single continuous line.

Upon successful registration, participants will receive instructions for accessing the following data:

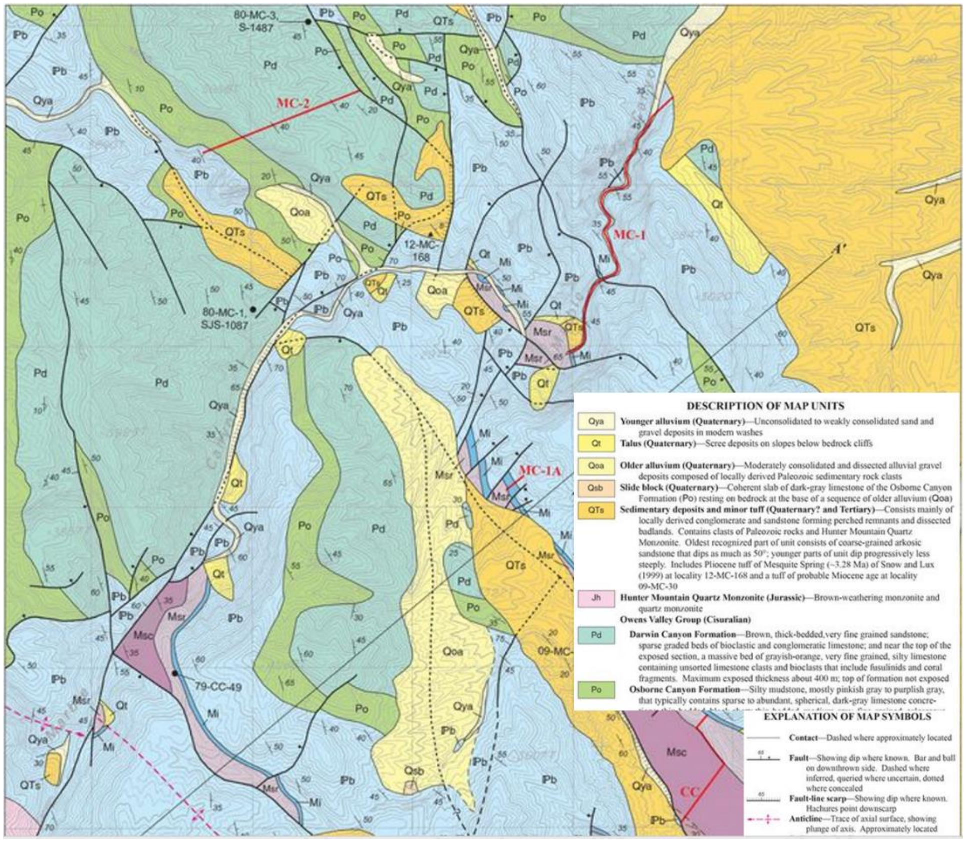
1. Training.tar.gz containing:
    a. A set of <basemap_name>.tif files
    b. One or more <basemap_name>_<feature_name>.tif binary raster files per <basemap_name>.tif file
    c. A set of corresponding <basemap_name>.json files which contain
        i. A list of the legend feature names (corresponding to the feature name part of the binary raster filename)
        ii. The corresponding bounding box coordinates for the legend entry for these features on the image
        iii. A classification of the feature as either point, line, or polygon.
2. Validation.tar.gz containing:
    a. A set of <basemap_name>.tif files
    b. A set of corresponding <basemap_name>.json files with the same properties as above

Examples of Variations in Scanned Maps

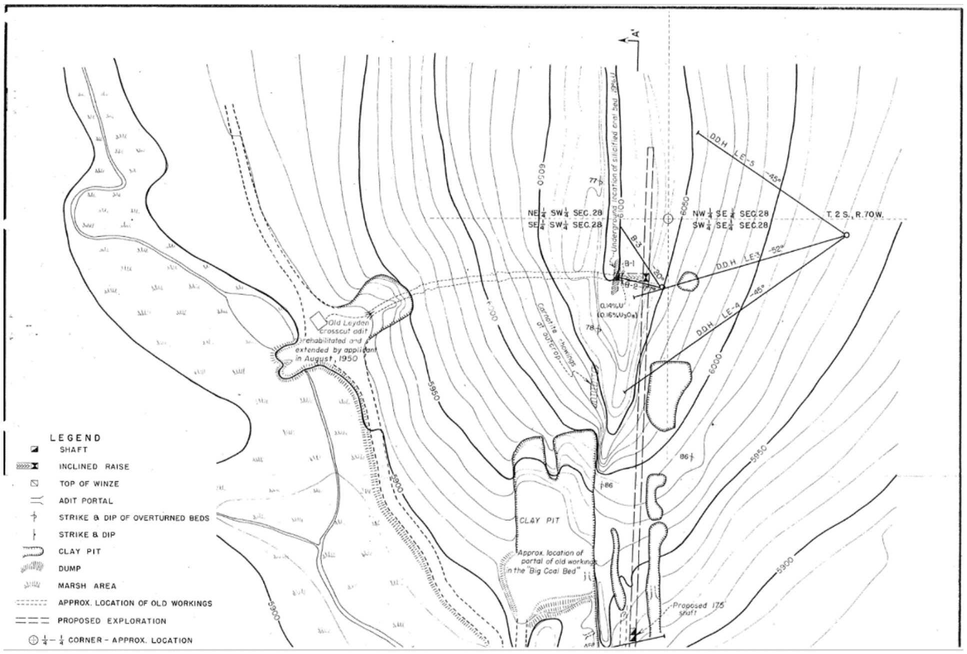
Here we provide examples of the types of maps that can be found in the training, validation and evaluation sets. Specifically, these maps demonstrate the variation in point, line and polygon features that a successful solution will need to extract. Note that there are many features on the basemap that are not part of the extraction task; additionally, not every legend feature has been identified for extraction for this task. For the sake of this challenge, your model should consider any features that are *not* explicitly identified in the json file as part of the basemap; you will not submit raster files for these. Also note that, in some cases, the features to be extracted may overlap, meaning that one or several pixels may be part of multiple features. For example, point markers, line features, grid lines, contour lines, and text that appear in a colored (polygon) region should be ignored for the purposes of extracting the polygon.


Legend
Coosa Project
AGC Land Holdings
Township 24 North Range 16 East
Township 24 North Range 17 East
Township 24 North Range 18 East
Township 23 North Range 20 East
Township 23 North Range 16 East
Township 23 North Range 17 East
Township 23 North Range 18 East
Township 23 North Range 19 East
COOSA
WILDLIFE MANAGEMENT AREA
Township 22 North Range 16 East
Township 22 North Range 17 East
Township 22 North Range 18 East
Township 22 North Range 19 East
Township 21 North Range 16 East
Township 21 North Range 17 East
Township 21 North Range 18 East
N
0 1 2 4 Miles
Metal Mining Consultants Inc.
9-05-2013
Copyright © 2013 National Geographic Society, i-cubed
86°33'0"W
86°30'0"W
86°27'0"W
86°24'0"W
86°21'0"W
86°18'0"W
86°15'0"W
86°12'0"W
86°9'0"W

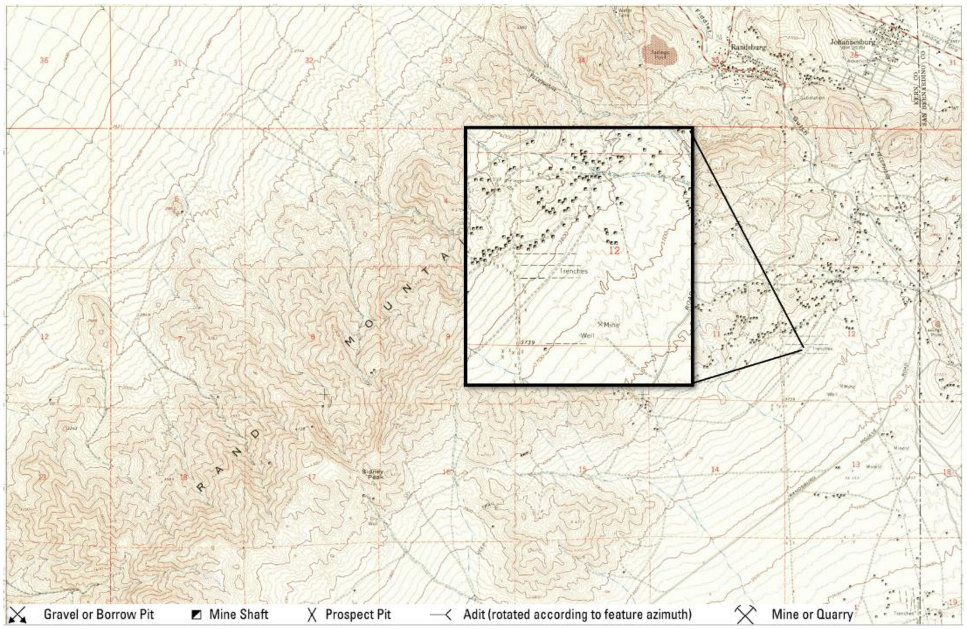
The above map has a point feature (star) and a discontinuous polygon feature.


MC-2
MC-1
MC-1A
CC
A'
80-MC-3, S-1487
80-MC-1, SJS-1087
12-MC-168
09-MC-
79-CC-49
Qya
Qt
Qoa
Qsb
QTs
Msr
Msc
Mi
Pd
Po
Pb
DESCRIPTION OF MAP UNITS
Qya Younger alluvium (Quaternary)—Unconsolidated to weakly consolidated sand and gravel deposits in modern washes
Qt Talus (Quaternary)—Scree deposits on slopes below bedrock cliffs
Qoa Older alluvium (Quaternary)—Moderately consolidated and dissected alluvial gravel deposits composed of locally derived Paleozoic sedimentary rock clasts
Qsb Slide block (Quaternary)—Coherent slab of dark-gray limestone of the Osborne Canyon Formation (Po) resting on bedrock at the base of a sequence of older alluvium (Qoa)
QTs Sedimentary deposits and minor tuff (Quaternary? and Tertiary)—Consists mainly of locally derived conglomerate and sandstone forming perched remnants and dissected badlands. Contains clasts of Paleozoic rocks and Hunter Mountain Quartz Monzonite. Oldest recognized part of unit consists of coarse-grained arkosic sandstone that dips as much as 50°; younger parts of unit dip progressively less steeply. Includes Pliocene tuff of Mesquite Spring (~3.28 Ma) of Snow and Lux (1999) at locality 12-MC-168 and a tuff of probable Miocene age at locality 09-MC-30
Jh Hunter Mountain Quartz Monzonite (Jurassic)—Brown-weathering monzonite and quartz monzonite
Owens Valley Group (Cisuralian)
Pd Darwin Canyon Formation—Brown, thick-bedded,very fine grained sandstone; sparse graded beds of bioclastic and conglomeratic limestone; and near the top of the exposed section, a massive bed of grayish-orange, very fine grained, silty limestone containing unsorted limestone clasts and bioclasts that include fusulinids and coral fragments. Maximum exposed thickness about 400 m; top of formation not exposed
Po Osborne Canyon Formation—Silty mudstone, mostly pinkish gray to purplish gray, that typically contains sparse to abundant, spherical, dark-gray limestone concre-
EXPLANATION OF MAP SYMBOLS
Contact—Dashed where approximately located
Fault—Showing dip where known. Bar and ball on downthrown side. Dashed where inferred, queried where uncertain, dotted where concealed
Fault-line scarp—Showing dip where known. Hachures point downscarp
Anticline—Trace of axial surface, showing plunge of axis. Approximately located

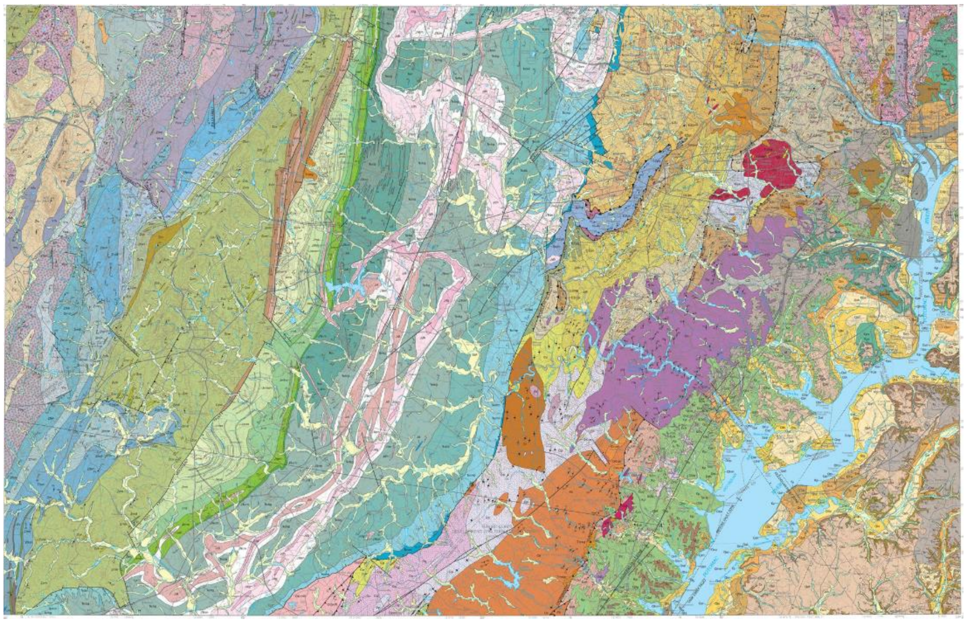
The above geologic map contains multiple polygon and line features.



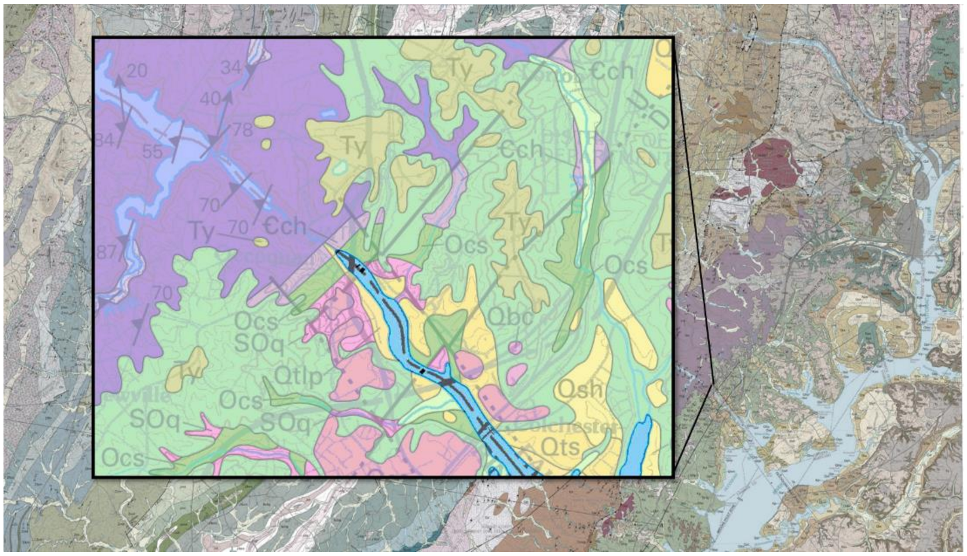
The above mineral site map contains point and line symbols that may be difficult to distinguish.



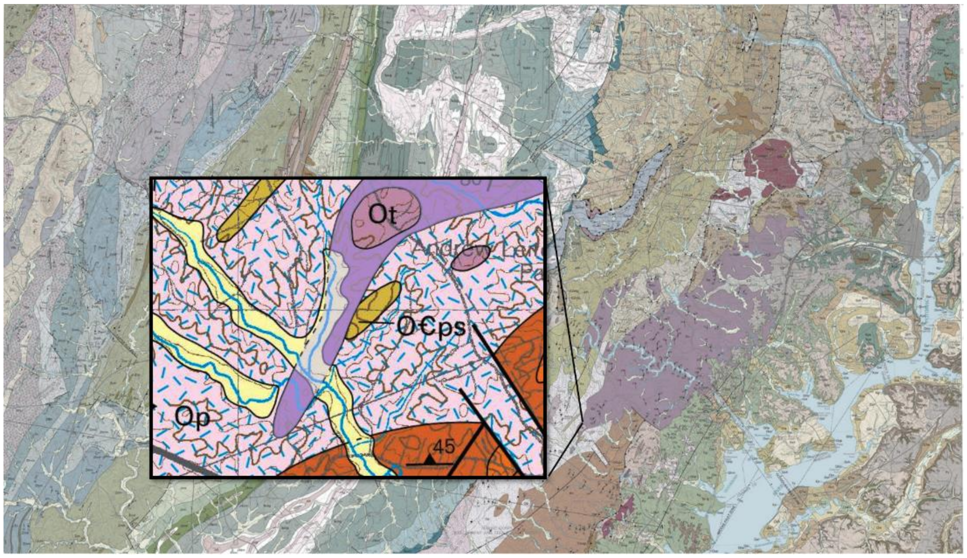
The above topographic map demonstrates the potential density of points and overlapping features.



The above geologic map demonstrates the range of complexity of polygon extraction. This map also demonstrates a particular challenge with water features. In practice, small water features would likely be vectorized as part of the larger polygon, whereas a major water feature (such as the Potomoc River on the right) would not.



For example, in the section of the map above, some water features are included within bedrock polygons (dark purple), but other water features are excluded. Surficial deposit (i.e. sand in river channels) features pose a similar challenge, as do other versions of overlapping features.



In the section of the map above, the purple bedrock polygon extends across yellow alluvium () polygon and blue stream. Additionally, there is a polygon denoted by a multi-colored pattern, rather than a single solid color.

These kinds of nuance are difficult to define with hard rules (see standards) but the training data provides good examples of how these features are accounted for in practice, with manual feature extraction.

Challenge Baseline

A simple baseline was developed using the Python OpenCV library to detect colors and symbols from the specified legend features and then detect these same colors on the pixels of the map image; however, this is perhaps the most obvious starting point. The examples and edge cases above illustrate why this approach is not sufficient for many of the maps and legend features that the USGS encounters in a typical mineral assessment.

The baseline notebook can be accessed here: https://CriticalMinerals.darpa.mil/Files/AI4CMA_Challenge_2_Baseline.ipynb.

Additional Tools and References

- From a physical scanned map to a digital elevation model using the legend and kriging by Carlos Rus et. al. (2005)
- Information extraction from topographic map using colour and shape analysis by NIKAM GITANJALI GANPATRAO 2014 (https://www.ias.ac.in/article/fulltext/sadh/039/05/1095-1117)

Pre-training Notice

It is fair game and even expected that you will pre-train your network/model with anything that is available online. However, if you are selected as a finalist, you will be required to describe any datasets you used for pre-training and provide them for the assessment process.

Validation and Evaluation Details

Validation

During the validation windows, you may submit your predictions for the maps and features in the validation set. These submissions will be used to update the Leaderboard each Friday during the challenge and provide feedback to participants on their model performance.

You will be allowed to make multiple submissions within the validation windows. See the Key Dates section on the competition website, https://CriticalMinerals.darpa.mil/The-Competition. However, you will only receive a score for your last successful submission received before the validation window closes.

For each validation submission, you must submit one binary raster image for each legend feature provided in the <basemap_name>.json file for a particular map image. The binary raster file should follow the following naming convention: <basemap_name>_<feature_name>.tif. Please make sure that the binary raster files should have 3 channels, with the same pixel values in each channel, exactly like the raster files in the training and validation datasets.

Your submission should be zipped and must be named as <team_name>.zip. If your team's name has spaces, you must replace these with an underscore (“_”) in your file name.

An example submission file is available on Amazon S3, alongside the training and validation data. You must upload the zip file to a submission upload portal before the close of each validation window. Instructions for the upload portal will be provided to you.

Evaluation

For the final evaluation, performers will be provided with an entirely new set of map images that are a representative mix of attributes found in the training and validation sets.

Performers will then have 24 hours to submit their results via a submission upload portal. Instructions for the upload portal will be provided to you. Failure to submit within the 24-hour timeframe will disqualify the submission from the evaluation.

You will be allowed to make only one submission during the 24-hour final evaluation window. You will not be contacted to correct errors in your submission so strict adherence to the challenge format is important.

If your submission is among the top five best scores, you will also be asked to submit a technical package which includes your code, a brief description of your technical approach, and reference links to any external training sets you used, alongside your results in accordance with the Competition Rules available at https://criticalminerals.darpa.mil/Files/AI4CMA_Competition_Rules.pdf. Failure to meet these criteria will result in a disqualification.

The final ranking of performers will be posted on the website leaderboard within one week of the submission deadline and the top THREE (3) qualified winners will be contacted by email. In the unlikely event of a tie, the submission that was received first will be awarded first place, and the other submission will be awarded second place, etc. Fourth and fifth place will receive an honorable mention on the competition website.

Submission Format

Your results must be submitted in a zip file with the following contents and formats

1. You must submit one binary raster image for each legend feature provided in the <basemap_name>.json file for a particular map image. The binary raster file should have the following naming convention: <basemap_name>_<feature_name>.tif. Please make sure that the binary raster files should have 3 channels, with the same pixel values in each channel, exactly like the raster files in the training and validation datasets.
2. Your raster files should be zipped into <team_name>.zip.
3. You will be given a link to a submission portal to upload the zip file. The submission portal will first verify your email address, so you must use the same email address as you used to register for the challenge.

Evaluation Metrics

For each legend feature, to find the corresponding true pixel for every predicted pixel, we will find the closest pixel pairs in the predicted and the ground truth binary raster. In the case of point and line features, the distance between the closest pixels will be considered, with a distance cutoff. For polygon features, only the overlapping pixel pairs will be considered, hence a distance of zero only.

For each feature, we will use the distances between the predicted and true pixel pairs to generate an f-score. In the case of polygons, each pixel pair will be weighted based on its difficulty of prediction by a color matching baseline.

Across all maps, the f-scores for each feature type (i.e., point, line and polygon) will be used to calculate a per-feature median score. The final score will be the average of the three median scores with a double weightage given to the polygon median. These averages will be used for ranking teams.

Note that only an unknown subset of the maps and/or map features from the evaluation and validation set will be scored; this is to mitigate the risk of gaming. Also note that a score of zero will be given for any missing features, so it is up to performers to confirm that their submission is complete.

The metric code is provided in a Python notebook, which can be accessed here: https://CriticalMinerals.darpa.mil/Files/AI4CMA_Challenge_2_Metric.ipynb.

Questions

If you have any questions, please send an email to ai4cma-questions@mitre.org