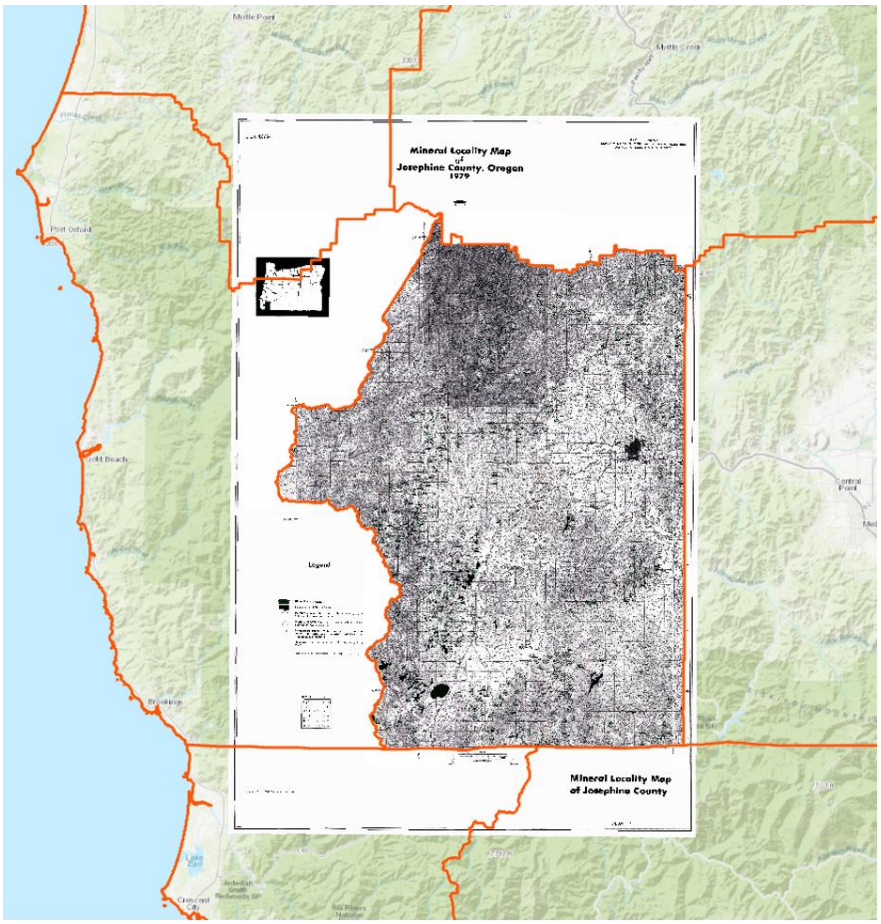
The Map Georeferencing Challenge

(current as of September 12, 2022)

Challenge Overview

This challenge seeks to find innovative approaches to **automatically** assign geolocation information using scanned and historical image data (i.e., with no explicit coordinate information associated with the map images). Currently, the United States Geological Survey (USGS) manually performs this task for almost every map in a mineral assessment (for more information on mineral 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., ArcGIS), it is still primarily done manually by recognizing coordinate ticks and grid intersections, or looking for toponyms, environmental, geological or political features on the map, and then relating these back to a base map. To completely automate this process, it is anticipated that sophisticated text and feature recognition and matching will be required, using techniques from the fields of computer vision and machine learning. Participants may use any information on the geological or topographic maps for this task without limitation.



The image above shows a raster map correctly georeferenced and then placed on top of a base or reference map. As a result, both maps have their geological and political features spatially aligned.

Challenge Details

The goal of this challenge is to develop automated approaches for georeferencing raster data (e.g., scanned and historical maps) with the ability to produce highly accurate geospatial positioning information. Solutions should preferably be open source, although candidates can use proprietary services (services with restricted licensing) provided they are clearly documented, and open-source alternatives are identified where possible. Solutions should demonstrate value to USGS objectives of automating precision georeferencing of map images. We reserve the right to disqualify solutions that don't meet USGS objectives.

The process of manually georeferencing images of maps involves identifying features from the image that can be used as reference points and determining their spatial coordinates. On map images, the reference points could be latitude-longitude ticks or lines, but some maps either lack coordinate marks, or the coordinate system is not clearly defined. In these cases, features from base layers on the map can also be used to identify reference points. Common base layers on geologic maps that can be used for georeferencing include topographic lines, roads, government boundaries, stream/shoreline shapes, etc. (reference: https://pro.arcgis.com/en/pro-app/2.8/help/data/imagery/overview-of-georeferencing.htm).

For this challenge you will be given an initial set of training data that will consist of map images and corresponding comma-separated values (CSV) files, which contain a set of randomly selected points on the map and their respective latitude-longitude values, correctly georeferenced in the North America Datum of 1983 (NAD 83) coordinate system.

At the same time, a validation data set will be released which will include a separate set of map images and corresponding CSV files with randomly selected map points, but *without* the corresponding latitude-longitude values.

For the final evaluation, you will be given a new set of map images and associated CSV files in the exact same format. You will be expected to provide a *single, concatenated* CSV file with the predicted latitude-longitude values for each of the map points on each image in the evaluation set, in the NAD83 coordinate system and in decimal degrees format.

Solutions to this challenge require both coarse- and fine- scale georeferencing. To facilitate coarse-scale georeferencing and focus on the fine-scale georeferencing problem faced by USGS researchers, we have released clue files for each map in the validation set. A clue will be released for the evaluation map set as well. The clue consists of a randomly selected latitude and longitude coordinate pair rounded to the nearest 10th of a degree that plots within the map image. The addition of the clue focuses the challenge on the part that will bring the most benefit to the USGS: accurately locating, scaling, and georeferencing the map so that the locations of map features are within an acceptable error compared to manual georeferencing done by a human. For example, USGS publication standards require points on a 1:24,000 scale map to be accurate to within 40 ft or 12.2 meters (https://pubs.usgs.gov/fs/1999/0171/report.pdf).

All map images in the training, validation and evaluation sets will be located in North America. Please see the **Validation and Evaluation Details** section of this document for detailed submission instructions.

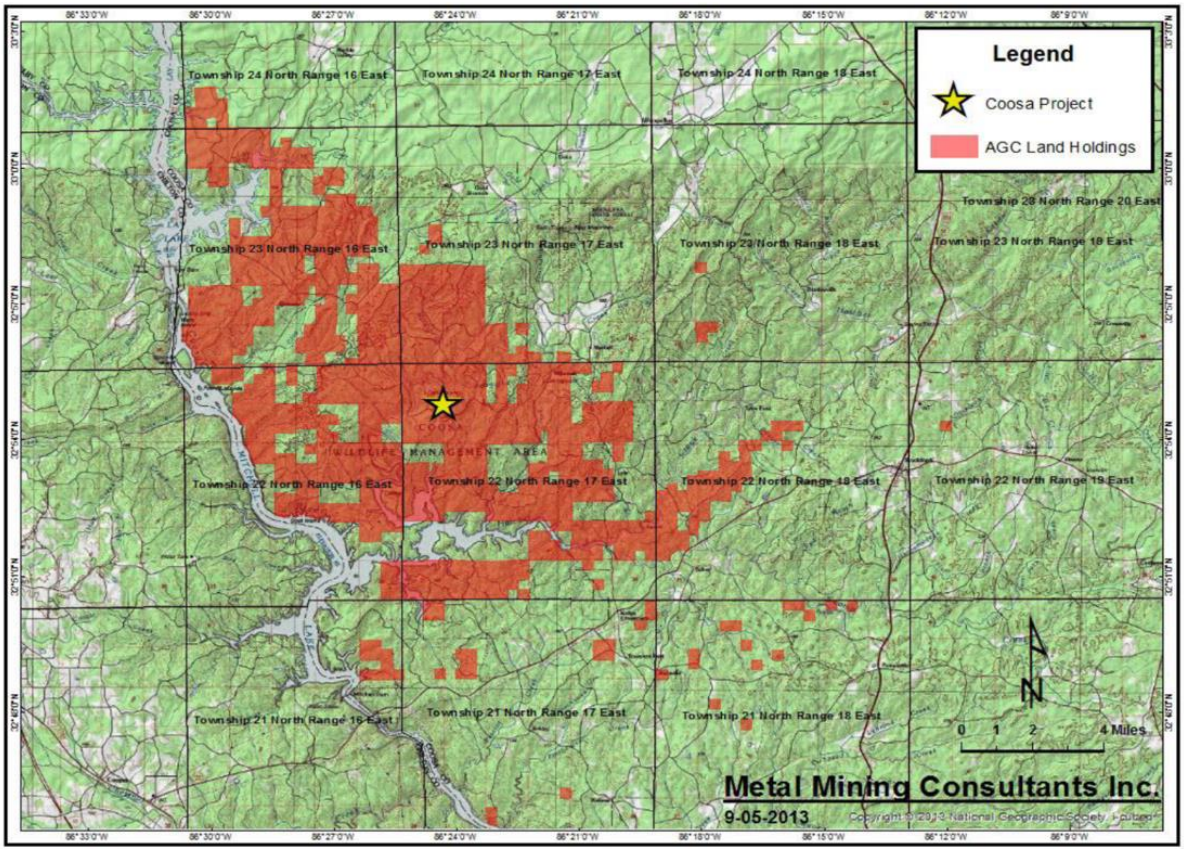
Challenge Dataset

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

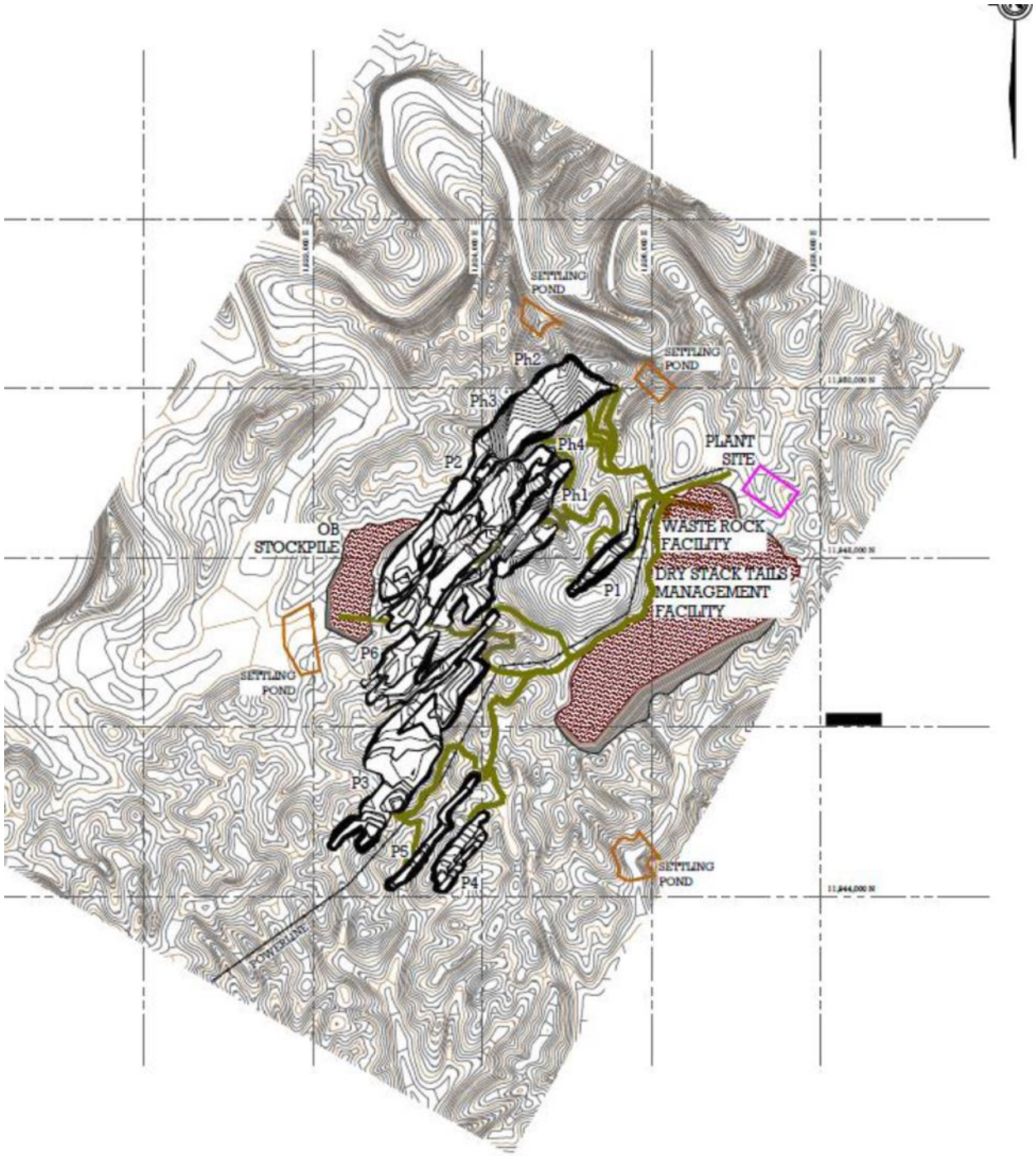
1. Training.tar.gz
    a. Contains a set of 724 <raster_id>.tif files
    b. Contains a set of 724 <raster_id>.csv files
    c. The csv files contain the image coordinate values for a set of points in the corresponding map image file.
    d. The top left corner of the image will be considered (0, 0) for the pixel location (row, col).
    e. Headers are raster_id, row (x), col (y), NAD83_x, NAD83_y – **DO NOT CHANGE THESE HEADERS OR THE VALUES FOR "raster_id"**
    f. All ground truth will be in NAD 83 and in the decimal degrees format.
2. Validation.tar.gz
    a. Contains a set of 305 <raster_id>.tif files
    b. Contains a set of 305 <raster_id>.csv files, formatted as above but with "0,0" values for the latitude-longitude coordinate fields. Your solution should populate these fields with your predicted coordinates.
    c. Again, DO NOT CHANGE ANY HEADERS OR VALUES FOR "raster_id"
3. Validation_clues.tar.gz
    a. Contains a set of 305 <raster_id>.csv files which contain a randomly selected latitude and longitude coordinate pair, rounded to the nearest 10^{th} of a degree, that plots within the map image.
    b. The clue files should NOT be a part of your validation submission.

Examples Of Variations in Scanned Maps

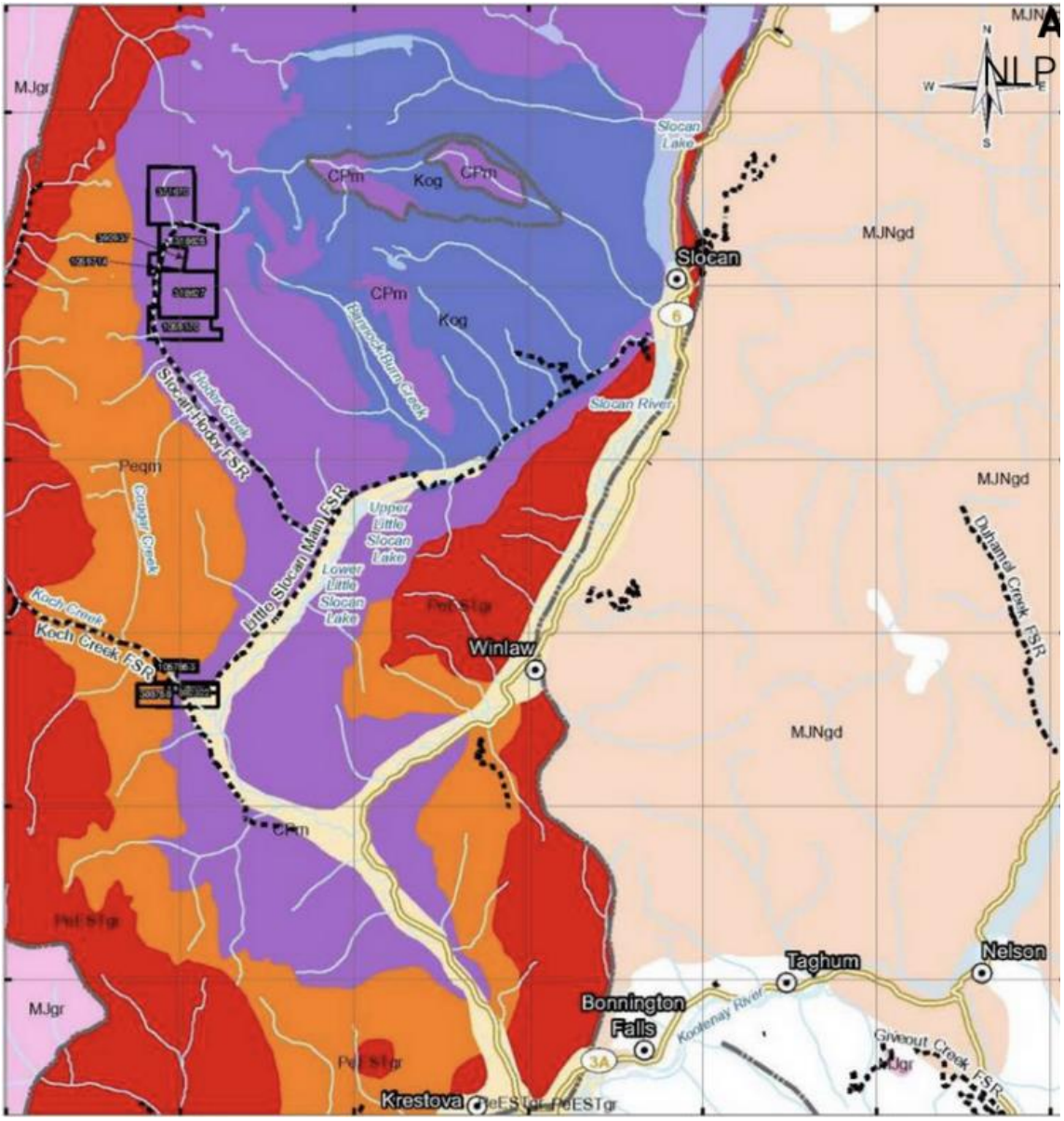
Below are examples of the variations in the map images that you can expect to find in the Training, Validation and Evaluation sets.



The above example follows a common practice of adding latitude-longitude on the edges of the map and providing grid lines as well. In this example, the grid lines are faint and there is another more prominent grid that may confuse a computer vision-based model.



In the above example, the latitude-longitude are provided right next to the intersecting grids, which is also a common practice.



In the above example, the map has a high number of more clearly defined geological features as well as toponyms that can be used for referencing and alignment.

Examples Of Edge Cases

The Evaluation set will contain some challenging edge cases, such as those in the examples below.

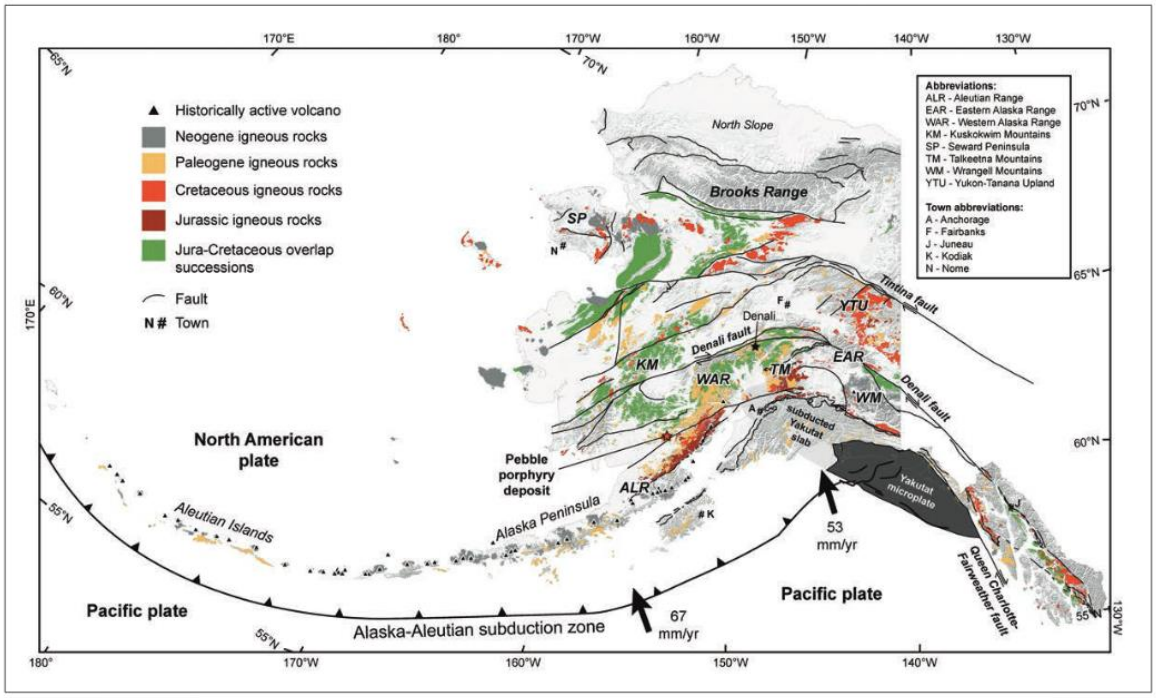
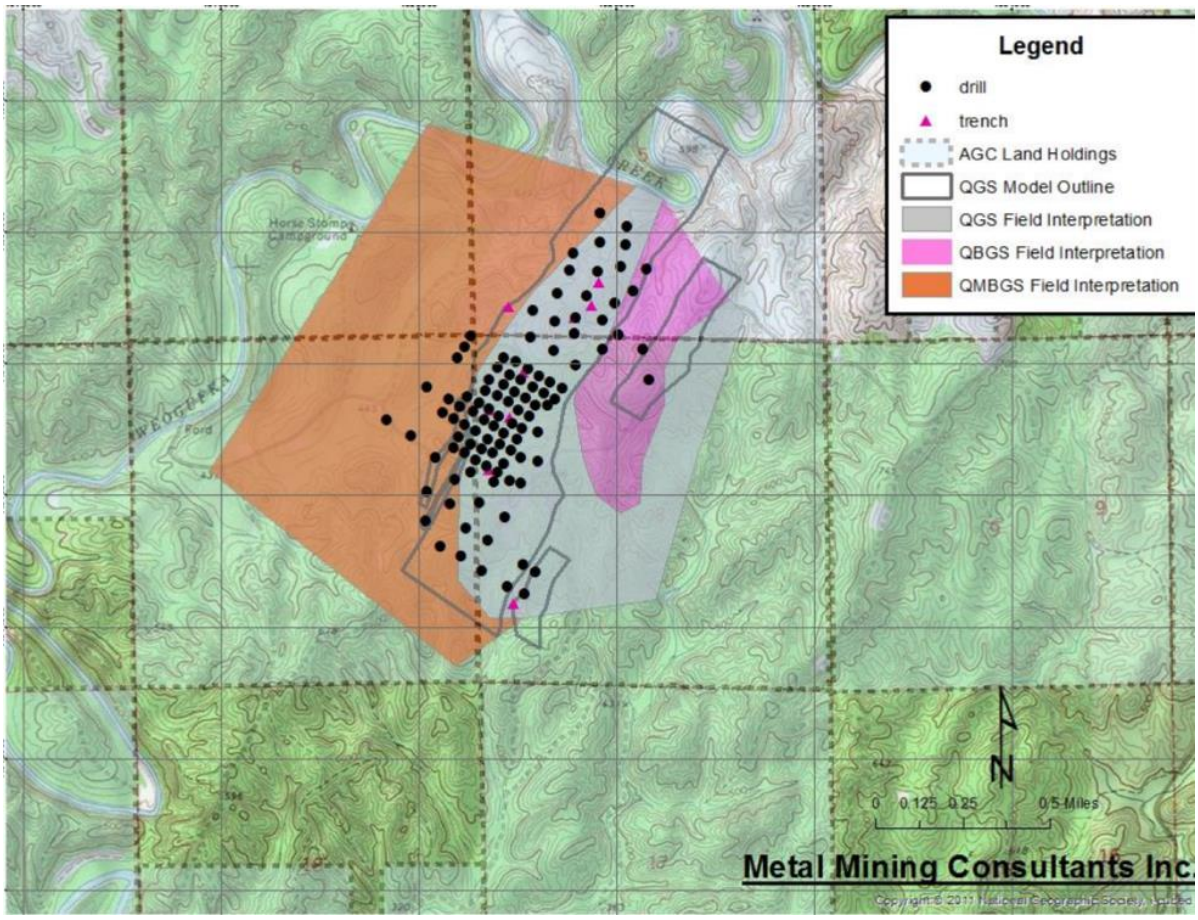


Figure 1. Statewide geologic framework and modern structural setting of Alaska. Plutonic belts and overlap successions are derived from Wilson, Hults, Mull, and Karl (2015).

The above example does not contain any visible grid lines or ticks. The grid is implicit in the boundary tick marks.



This map contains geological features but no toponyms. The latitude-longitude grid are the fine lines which are obscured by the quadrangle lines and the colored overlay.

Base Maps

Base maps are necessary for georeferencing scanned and historical maps, especially those which do not contain any tick or coordinate information. Below are a few resources that you can use to find base maps to assist with the challenge task:

1. https://www.usgs.gov/faqs/there-api-accessing-national-map-data

   The National Map (TNM) has one API (TNMAccess) that provides access to all TNM downloadable products that are available through The National Map Download Client. Developers can use HTTP GET and POST requests to directly access products, or use the light graphical interface on the API page to generate queries to send to the API.

2. https://ngmdb.usgs.gov/topoview/

   TopoView shows the many and varied topographic maps of each of the areas through history. This can be particularly useful for historical purposes, such as finding the names of natural and cultural features that have changed over time.

3. https://www.arcgis.com/home/item.html?id=99cd5fbd98934028802b4f797c4b1732

This layer presents detailed USGS topographic maps for the United States at multiple scales. The map includes the National Park Service Natural Earth physical map at 1.24km per pixel for the world at small scales, i-cubed eTOPO 1:250,000-scale maps for the contiguous United States at medium scales, and National Geographic TOPO! 1:100,000 and 1:24,000-scale maps (1:250,000 and 1:63,000 in Alaska) for the United States at large scales. The TOPO! maps are seamless, scanned images of USGS paper topographic maps.

4. https://usgs.maps.arcgis.com/home/item.html?id=f84ec1bf1de0452d8e7c9e5f8fac4ca5

   The PLSS is the basis for federal land ownership. This data includes township, range, section (first Division), and intersected. There are four layers loaded that are scale dependent with scale dependent labels. At the smallest scales, the state boundaries appear, and as the user zooms in, townships, then section, then PLSS Intersected boundaries appear. ArcGIS REST

5. https://jakob.schwalb-willmann.de/basemaps/#supported-services-and-maps

   basemaps is a lightweight R package to download and cache spatial basemaps from open sources such as OpenStreetMap, Carto, Mapbox, and others. This package aims to ease the use of basemaps in different contexts by providing a function interface as minimalist as possible.

Challenge Baseline

To provide informational baselines for this challenge, several open-source automated solutions were investigated. Using these as a reference, a custom baseline was developed to test evaluation protocols and generate scores against the validation set for the leaderboard. This baseline represents just one of several possible approaches and does not address the full complexity of the problem. The baseline notebook can be accessed here: https://CriticalMinerals.darpa.mil/Files/AI4CMA_Challenge_1_Baseline.ipynb.

Additional Tools and References

The list below is not exhaustive, and many of these tools are semi-automatic.

- QuadG+ is a state of the art publicly available software that automatically converts an image from a scanner's coordinate system [(x, y) pixels] to a known spatial reference system (SRS). However, it cannot handle all of the maps that USGS encounters during assessments.
- NotAQuad: A tick based, semi-automated tool.
- MapKurator: https://github.com/kartta-labs/Georectifier
- Content-based Image Retrieval for Map Georeferencing by Jonas Luft (http://meatboy.jonasluft.de/data/ICC21_full_paper_submission.pdf)
- ArcGIS: has an automatic capability that uses Red-Green-Blue (RGB) bands to sync maps. https://developers.arcgis.com/python/guide/geo-referencing-and-digitization-of-scanned-maps/.
- PyTesseract: Provides optical character recognition (OCR) capabilities.(https://github.com/madmaze/pytesseract)

Pre-training Notice

It is fair game to pre-train your network/model with anything that is available online.

Examples:

The David Rumsey Map Collection , the Library of Congress, New York Public Library, MapWarper, and McMaster University Library are large map datasets.

Validation and Evaluation Details

Validation

During the validation windows, you may submit latitude-longitude predictions for the maps 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 a *single, concatenated* CSV file with the predicted latitude-longitude values for each of the map points on each image in the validation set, in the NAD83 coordinate system and in decimal degrees format. You must submit latitude-longitude values *for every map point* in the CSV file; even a single empty submission will negatively affect your score.

Your submission file must be named as <your_team_name>.csv. If your team 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 email the CSV file to **ai4cma-submissions@mitre.org** before the close of each validation window.

Evaluation

For the final evaluation, performers will be provided with an entirely new image set that is representative of the variety that USGS encounters in the wild. As noted in the examples above, the evaluation set will contain images at different difficulty levels, including some potentially more difficult than those in the training and validation datasets.

Performers will then have 24 hours from the release of the evaluation set to submit their results via email to **ai4cma-submissions@mitre.org**. 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; therefore, strict adherence to the challenge format is important.

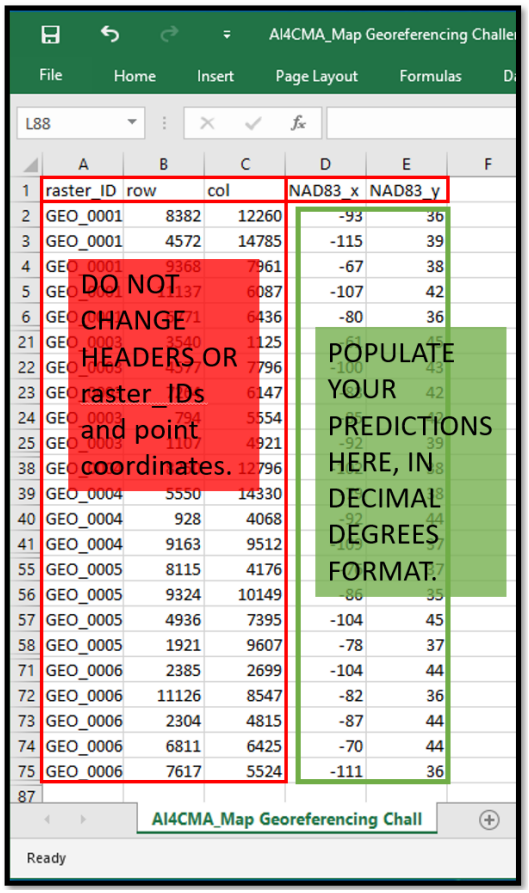
If your submission is among the top five best scores, you will also be asked to submit a technical package which includes your code, a brief description of your technical approach and reference links to any external training sets you used, alongside your results in accordance with the Competition Rules (https://criticalminerals.darpa.mil/Files/AI4CMA_Competition_Rules.pdf). Refusal to provide these items may result in a disqualification. 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 as a single concatenated csv file as described and shown below:

1. **The csv file must be named as <your_team_name>.csv**. If your team name has spaces, you must replace these with an underscore ("_") in your file name.
2. The submission file should follow the same format as the provided sample csv file. You will have to populate the NAD83_x and NAD83_y fields with the predicted latitude-longitude for each map point, for each map in the evaluation set.
3. **DO NOT change any raster_ids** as these are unique and will be used to evaluate your submission against our ground truth.
4. Remember that predictions for each reference point must be in the NAD 83 coordinate system and in the decimal degrees format.



Evaluation Metrics

Root Mean Square Error (RMSE) will be the evaluation metric for ranking submissions. RMSE is the standard deviation of the residuals (prediction errors) and measures the difference between the submitted latitude-longitude coordinates for each map point and the ground truth coordinates for each point. For this challenge, the residuals will be measured in kilometers (as geodesic distance, see https://en.wikipedia.org/wiki/Geodesic), which takes into account the curvature of the earth and is thus more accurate for map points at higher latitudes. An RMSE score will be computed for each map in the evaluation set, and the ranking will be determined by *the median of all RMSE scores.* Again, it should be emphasized that latitude-longitude coordinates must be submitted for *all of the map points*; failure to do so will lead to zeros being entered for that point and a losing RMSE score for that map.

Questions

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