Retriever Technology
1600 Lena St., STE D1
Santa Fe, NM 87505 www.retrievertech.com

PHASE I PROJECT NARRATIVE

Principal Investigator: Andrew H. Bartlett

Project Title

A Multi-Discipline Approach to Digitizing Historic Seismograms

Funding Opportunity Number: DE-FOA-0001619

CFDA Number: 81.049

Topic 4b, Technology to Facilitate Monitoring for Nuclear Explosions: Other

Table of Contents

Identification and Significance of the Problem or Opportunity

Retriever Technology, under previous DOE contracts, has developed SKATE (Seismic Kit for Automatic Trace Extraction), an advanced tool for digitizing historic analog seismograms. An sophisticated prototype version is available now, at **seismo.redfish.com**, for any researcher to access and use, at no cost. With SKATE, we have developed accurate feature recognition algorithms, and implemented disambiguating methods that automatically assign active traces to their correct position in a downloadable final time series. Because there is a pressing need to digitize the vast and unique stores of seismic information, our work to date has focussed on the automated and mass digitization of tens of thousands of seismograms. To that end, we have digitized, up to the point of editing, over 30,000 seismograms at a cost of less than $0.02 cents per image, thereby demonstrating the low cost capabilities of SKATE and pointing the way towards the mass digitization of many historic repositories, particularly the WWSSN film chip archives.

We recognize that in order for mass digitization to be effective, the software tools need to output data that requires minimum user input and editing. We also recognize the need to have a digitization tool that can be utilized to investigate *any* seismogram of interest. Therefore, our work in this Phase I research will satisfy both of these requirements by: a) further increasing the accuracy of all of our feature identification, segmentation, and connection algorithms, and b) opening up our web-based digitization software to allow our EC2 instances to digitize any inputted seismogram. These inputted seismograms could be chosen from our online repository of over 150,000 WWSSN short period and long period seismograms, or they can be user inputted images from their own collection.

A successful Phase I effort will increase the accuracy of our existing algorithms; will allow for any seismogram to be digitized; will publish all code and share the code to allow for a significant amount of the process to be run using open source and free software on a desktop computer; and will create a repository where the software can persist nearly indefinitely, with mechanisms in place to continuously share, develop and deploy advances in the software.

Technical Approach

A key to our technical approach is the exclusive utilization of online, web-based tools, and the new capability of our software to digitize user-supplied seismograms. The advantages of our online tool are numerous, and include:

- No software purchase or installation is required to operate SKATE.
- The use of Amazon EC2 instances allow for computational power that is not available in typical client computers.
- The use of open source software and publicly shared code simplifies use and maintenance.

Our philosophy is to use all open source software as is possible, and to publish, share, and allow for any qualified user to modify the program. To that end, all of our code has been or will be published and is available to any user at a public GitHub account, located at https://github.com/retrievertech. GitHub is a web-based Git repository hosting service. It offers all of the distributed version control and source code management functionality of Git, with access control and several collaboration features such as bug tracking, feature requests, task management, and wikis for every project[1]. In other words, our public repository can be browsed and downloaded by anyone, and modified by qualified users. We are managing this account and will actively solicit qualified users who will be able to discuss, manage, create repositories, submit contributions to others' repositories, and review changes to code. These features, combined with the ability for SKATE to operate at a nominal yearly fee will allow the code to persist nearly indefinitely, and to be subject to continuous user-inputted improvements.

Our technical approach begins with fundamental image analysis techniques, and ends with the development of a web-deployable software package. The steps in between combine micro (pixel level) and macro (segment and larger levels) approaches that utilize the significant amount of image information in the seismogram with many advanced types of analysis techniques. These include utilizing basic seismograph response to create connectivity models for trace gaps, using predictive curve fitting techniques to model trace paths, incorporating image properties to match and connect likely segments, and bounding likely paths with global energy constraints. Our management of the project will emphasize regular meetings and interactions among team members, and leverages an experienced advisory team to track progress and interactions.

We have focused our previous work on mass digitization, and have the tools in place to allow for the automated digitization of large archives. To demonstrate that, we have already digitized over 30,000 WWSSN long period

[1] https://en.wikipedia.org/wiki/GitHub

seismograms, and have those results available today for any user to immediately access, edit, and download the final time series data. What this Phase I proposal now includes is the enhancement of SKATE to allow users to input their own seismograms, either all-day records or partial records that contain regions of interest. This new functionality will greatly increase the utility of SKATE by transforming it into a standard tool that can be used by any researcher to digitize any seismogram. By emphasizing usability, we aim to insure that the DOE investment in this Phase I research will result in a the creation of a permanent tool that will meet the needs of scientific research, at minimum cost, and with the ability to continuously evolve.

Anticipated Public Benefits

There are numerous public benefits that our digitization software provides. Nuclear non-proliferation efforts require data from a large number of sources, including seismic data. For example, there exists a backlog of information obtained from the Peaceful Nuclear Explosion (PNE) program that can provide unique insights into current studies looking for evidence of clandestine nuclear explosions.[2] Numerous studies on the need to digitize these waveforms have been made. "From a long-term perspective, it is important to archive high-quality signals from past explosions, and to make them usable for diverse research at present and in the future, because they are unique and are so much better than earthquake signals for certain types of study of the interior of planet Earth. This is a project that must be maintained long-term, and augmented to archive high quality signals from all energetic explosions especially from those at yield levels that will not be achieved again unless nuclear explosions resume on a large scale[3]." Further,"the technical challenge in monitoring treaty compliance into the future includes the need to monitor for nuclear explosions in all possible types of geological environment[4]," in which, "the digitized seismograms can be used not only for developing new methods for explosion monitoring, but to calibrate stations of the CTBTO International Monitoring System...and to investigate the ways in which nuclear explosions have had an influence on the medium in which they were conducted[5]."

Other types of investigative and predictive seismological studies would also benefit from the availability of this data, such as improved climate change studies based on the background noise present in all seismograms[6]. As well, there are many types of paper records for a variety of geophysical applications, such as seismic oil exploration, or borehole geophysical records, which our underlying code could be modified to address. Successful completion of This Phase I (and beyond) research will result in a flexible and unique analysis tool for historic seismograms that will be funded by nominal user fees (reflecting the low cost of maintaining the SKATE website) and continuing efforts, as evidenced by ongoing USGS funding of our website, to successfully solicit funding and interest for this important software tool.

Technical Objectives

Retriever Technology has developed and deployed SKATE, allowing any user to access, without purchasing or installing any software, to edit as necessary, and to download time series data from over 30,000 digitized images. We encourage readers to visit the website at seismo.redfish.com in order to explore and understand its features and functionality. Additionally, the reader is directed to a report from previous work that details the current status of the software, and is available at https://docs.google.com/document/d/1lQ7IlSvtQ2fEqEMWkhJy-qF8vqdiNqxAUbexDB2Jjrs/edit?usp=sharing. In order to further develop its capabilities and ensure its long term usefulness, our technical objectives will focus on four key areas. While these tasks focus on WWSSN seismograms, most of the work is generalized to any analog seismogram type.

1) Allow SKATE to process user-supplied images. Currently, SKATE supplies time series data only for those 30,000 images which we have processed on our EC2 instance. This reflected the successful beginning of achieving our goals of automating and completing the production of useful data for as many historical seismograms in our archive as possible. However, there is demand for better digitization tools for individual seismograms of interest.

---

[2] E.g.: P. G. Richards, et al., A Plan for Location Calibration of IMS Stations in and near Kazakhstan, *Geophysics and Non-proliferation problems*, 2, National Nuclear Centre of the Republic of Kazakhstan, June 2001.

[3] "A digital seismogram archive of nuclear explosion signals, recorded at the Borovoye Geophysical Observatory, Kazakhstan, from 1966 to 1996," An, et al., GeoResJ, Volume 6, June 2015, Pages 141–163.

[4] http://www.ldeo.columbia.edu/res/pi/Monitoring/Data/Brv_arch_ex/EOS92_Richards&Kim.pdf

[5] http://www.ldeo.columbia.edu/~richards/my_papers/AFRL-RV-PS-TR-2015-0089.pdf

[6] "Seismic noise analysis as a tool for studying climate change: A proposal to digitize historical seismograms from pre-cursor GSN stations," R.C. Aster.

By allowing users to upload and process any seismogram, we will greatly increase the usefulness and visibility of SKATE, helping to provide a permanent site where the significant efforts already put into SKATE can persist for ongoing digitization and research needs.

2) Our current feature detection tools will be further developed. We have implemented many algorithms to identify trace features and intersections, while rejecting unwanted (i.e. noise) features. There are many areas where these algorithms can be refined and improved within the scope of this Phase I work, resulting in more accurate trace detection, better trace assignments, and greatly reducing editing requirements. Also, timing marks will be explicitly identified and this information will be used in assignment algorithms.
3) Implement new connection algorithms. We have in place a Euclidian Distance method for properly assigning traces in the presence of significant seismic activity. We will add several new and powerful algorithms that together will produce a much more accurate result, allowing for significantly less user interaction and editing.
4) Ensure the persistence of SKATE. As a web based tool, SKATE requires nominal yearly fees to operate. A small fraction of this Phase I funding (approximately 6%) will be utilized to continue to host SKATE for five years, including its ability to process any user supplied seismogram. The public hosting of all the code and ancillary documentation on our Github repository at https://github.com/retrievertech creates the ability for the larger community to participate and improve this software in an ongoing fashion. The image processing code is written in Python, and is referred to in this document as the Pipeline.The Python code for the Pipeline, most of which is able to run on a desktop computer[7], will be made available for download and operation on any user's desktop computer, allowing for additional analysis and code modification capabilities.

These tasks are carefully chosen to allow the most important improvements in SKATE to be implemented within the time and budgetary constraints of this Phase I proposal. Our overarching goal is to have a complete and useful tool that outputs usable and accurate data from any whole or partial seismogram.

Technical Objective 1. Add user operation of any inputted seismogram.

- Allow users to run SKATE on existing seismograms in our archive. This opens up analysis to all of the 150,000+ scanned full resolution images that are stored on our website. This includes all short period and long period seismograms[8].
- Allow users to input any seismogram or partial that they upload to the site.
- Add parameter adjustments for select features to allow for on the fly tuning of algorithms.

Technical Objective 2. Pipeline improvement.

Technical Objective 2a. Improve ROI and meanline detection. Improve noise rejection / feature detection.

- Refine existing methods, and add new methods to increase the accuracy of Region of Interest (ROI) and meanline detection.
- Continue to develop algorithms for spurious feature (noise) removal. Automatically identify 'Dogtags'-the stamp on most WWSSN seismograms with handwritten station information written in- to eliminate false segment identification. Utilize ROI information to eliminate ROI boundaries as misidentified segments.
- Improve segment identification, especially at intersections, segment ends, and low intensity paths, allowing for more complete segments in these regions.

[7] Except the editing tools, which are web-based. Nevertheless, running the image analysis tools on a desktop allows for rapid understanding and modification of the code.

[8] Previous comments suggested that SKATE was unable to digitize short period waveforms. Our decision to focus on long period waveforms in previous work reflected our belief that these were of greater interest to researchers. The features in SP seismograms have been shown to be equally amenable to digitization by SKATE, and will be available for analysis as outlined in this Technical Objective.

Technical Objective 2b.Map timing marks in SP and LP WWSSN seismograms.

- Create location maps for timing marks based on known SP and LP timing mark features and locations.

Technical Objective 2c. Create continuous and corrected time series data.

- In the current implementation of SKATE, time series data is outputted as a multi-column .csv (and JSON for some of the data). We will implement code to output this as a single, continuous time series, with distortion correction.

Technical Objective 3. Develop and implement connection algorithms.

Technical Objective 3a. Geometric segment assignment.

- Improve our existing geometric segment assignment. Eliminate overlapping in time segments that are assigned to the same mean line. Add gray scale information to infer ownership of the trace by comparing to nearby assigned traces. Implement multiple runs of the algorithms to increase accuracy of assignments.

Technical Objective 3b. Bayesian methods.

- Utilize Bayesian methods to identify likely trace paths on either side of trace crossing intersections.

Technical Objective 3c. Agent based MTSP.

- Utilize MTSP (Multiple Traveling Salesmen Problem) agent-based methods to create possible continuous trace paths across the entire seismogram. Running this multiple times will reinforce best paths.

Technical Objective 3d. Multi criteria decision making.

- Combine the three connection algorithms of Work Plan 3 results into a single Multi Criteria Decision Making matrix.

Technical Objective 4. Long term operation of SKATE.

Technical Objective 4a. SKATE persistence.

- Keep the application running for 5 years. Move un-digitized image data into Glacier storage for cost savings.

Technical Objective 4b. Publish the Python Pipeline.

- Publish the Python pipeline for desktop utilization. Publish all relevant documentation to allow this, as well as all other documentation relevant to SKATE operation, including algorithms investigated but not yet utilized.

Work Plan

At the completion of this Phase I work, SKATE's status will be a highly functional software package operating at a Technical Readiness Level 7[9]. In particular, by opening up the program to user inputted images and modifying our website to run the Pipeline on any image, we will have created a complete process that will be available for day to day research purposes. Management supervision of the project is tight, as the PI has been integral to the development of SKATE from its inception. As well, much of the ongoing work will either be done by the original personnel, or will be developed on already researched, but not yet implemented, work.

[9] https://cfwebprod.sandia.gov/cfdocs/CompResearch/docs/TRL-Guidance-final.pdf. The online version is by definition in an operational environment.

Work Plan 1. Add user operation of any inputted seismogram

SKATE currently processes an image only by command of an administrator. By allowing any user to spin up an EC2 instance and process a seismogram, we will help enable SKATE to make the transition from a research level software to an operational package. Moreover, allowing users to input their own seismograms expands its usefulness to a much wider range of seismogram types and end users.

- Modify website to allow for operation of the Pipeline upon user command[10].
- Allow users to upload and digitize their own images. We will specify the formats that are allowable and create rules, such as boundary buffers as necessary in order to prevent out of bounds and other errors.
- There are numerous hard wired parameters in SKATE. Allow the user to pull up an 'advanced' screen to change these parameters in order to improve results. Feature size discriminators, gaussian filter sizes,and other known key parameters in the process will be made available.
- We will add functionality to the Python code. Because the Python code can be run on a desktop, add segment .csv output and other useful and easily implemented features.

Work Plan 2 Pipeline improvement

Work Plan 2a Improve ROI and meanline detection. Improve noise rejection / feature detection

Improvement and addition of noise removal algorithms will greatly reduce user editing, as nearly all editing requirements to date are the removal of spurious features. These noise features are most typically associated either with common and reproducible features, such as the border of the ROI or the 'dogtag' stamps, or with small false segments associated with variations in background intensities.

Improvements in the detection of the meanline will also be done, as meanlines are a key component to eventual segment assignment[11].

ROI. The ROI is identified by a Hough transform procedure, resulting in a linear border. We will build on this initial identification by searching for large changes in intensity near the ROI borders, starting with Canny methods, and create an ROI boundary that follows the exact contour of the ROI. This information will be used to eliminate false segments that are often associated with the ROI boundary.

Dogtags. Dogtags can be identified fairly readily by the typical size and morphology of the box that surrounds them and the typical location in the upper left (or other corner) of the seismogram. Generalized Hough algorithms will be used to compare against area, location and typical histogram data for the dogtags to ensure proper identification and removal.

Meanline detection. While the current technique works well, better discriminators and filters will be incorporated into the existing Hough methods for meanline detection. We will add additional morphological identifiers for quiescent segments (which are used to locate the meanlines) to better tune the algorithms. Average spacing will be used to look for likely locations. Other techniques will be added as deemed effective.

Small spurious noise. We have used *a priori* methods to identify timing marks, by looking for regular features within a typical size range. For WWSSN images we will add a tool that inputs the known timing mark spacings of Work Plan 2b and timing mark known sizes in order to categorize these. This information will be combined with our existing timing mark ID techniques to further eliminate spurious small noise features that otherwise can be confused as timing marks. Other spurious features will be eliminated in the connection algorithms of Work Plan 3 by deleting orphaned segments that have not been assigned to any meanline.

[10] Note that this will create a nominal cost burden since each seismogram requires only $0.02 to process. All other editing uses a fixed fee EC2 instance that adds no additional costs to our budget.

[11] See Work Plan 3.

Work Plan 2b. Map timing marks in SP and LP WWSSN seismograms.

Timing marks are implicitly found in, e.g., the Euclidian connection algorithm as in Work Plan 3a. We have also investigated other methods, such as those that can be found in a shared document at https://drive.google.com/file/d/0B8dC-QzX8dcjbG9uSllzZ2E1OTA/view?usp=sharing. We will create a map that will:

- Use the size and location discriminator as discussed in Work Plans 2a and 3a to create a list of timing mark locations.
- Use the methods discussed in the linked document to identify additional timing mark locations.
- Create a geometric model of locations and use this, along with distance and size filters, to explicitly identify and index timing marks for later use.

These data will be used to: 1) Segment the image into foreground and noise data (i.e. it will greatly assist with feature identification and noise reduction), and 2) Assist the assignment algorithms of Work Plan 3 by simplifying meanline assignment of high probability timing marks.

Also, with timing marks identified, temporal locations anywhere in the seismogram can also be mapped, and the marks can be explicitly included in the time series once the marks are identified. This 'moving' of timing mark data will not be done in this Phase I work (though the linked document above shows it is readily done in image space), but the data will now be in place to be addressed in Phase II or other future funding. Note that, e.g., frequency analysis in data space is a far more robust method than working in image space.

Work Plan 2c Create continuous and corrected time series data

Time series data has been available since the completion of previous work[12]. We will develop this further to output a single time series, rather than the current multi-column format that outputs separate data for each meanline (hourly or 15 minute segments for WWSSN data). We note that amplitude is relative to its global position on the original image. We will localize the data by using positional data from all segments per line (i.e. per hour for LP seismograms) and create a non-linear meanline/zero energy line to output the data as absolute with respect to its meanline. This is a straightforward task that will use the JSON data described in Work Plan 3. We believe that a second order curve fit will suffice for meanline fit, which appears to account for image distortions resulting from the original photographic process.

Each meanline will then be normalized to a single zero energy value and all hourly (for WWSSN LP) lines will be outputted as a single file. We will create a framework[13] for more complete temporal positioning of the lines, which need to account for gaps at the beginnings and ends of lines,as well as other effects, and will need to incorporate timing mark information for proper positioning. The tasks for this latter effort are understood, but will not be completed in the limited time of this Phase I.

Work Plan 3 Develop and implement connection algorithms

JSON data on all segments is already created by our Python scripts. This information will be organized in a form universal to the various connection algorithms to allow all relevant image data to be combined and analyzed by our connection algorithms. This JSON metadata includes segment start and end location, gray scale information, both local and averaged for each segment. Such well organized metadata opens the door for data analysis beyond that which is within the scope of this Phase I work.

To assign segments to their proper position within a final time series, we utilize two complementary techniques. The first is to assign a segment to its meanline. Meanlines are the zero-energy lines about which the seismic traces oscillate. So for a WWSSN long period image, there are typically 24 meanlines in the image; one per hour, traversing the width of the seismogram. A simple way of looking at the meanlines' role in the segment assignment solution is to note that if all trace segments are properly assigned to a meanline, then the analysis is complete. The second technique is to connect a segment to the segments in its neighborhood. In this case, the use of priors in the Bayesian technique in Work Plan 3b will create likely connections based on the behaviors of neighboring segments. The MTSP technique of Work Plan 3c is somewhat of a hybrid: while connecting segments to neighboring segments, the final result also requires that the

[12] Previous reviewer comments mistakenly stated that time series data wasn't available.

[13] Indexing of each line will allow it to be manipulated according to any additional positioning information, in either the X or Y directions.

chosen paths end at the same meanline at which they started, and that the integral of their energy equals zero. The combination of all these techniques in a multicriteria decision matrix will output time series data that is significantly reduced in errors compared to our current techniques. The goal is to reduce editing of the solution to a minimum.

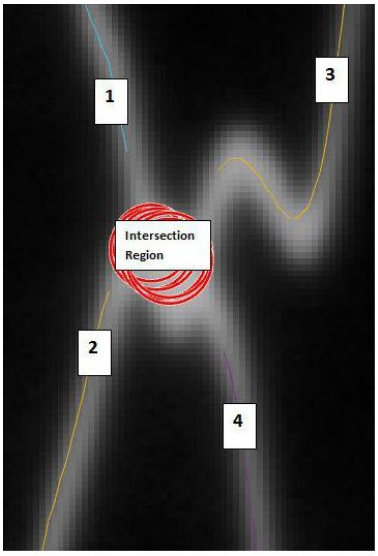
Note that all of these techniques have already been developed to a significant degree, such that their incorporation into SKATE fits well within the time constraints of this Phase I work. Efficiently incorporating the entirety of a seismogram, rather than just selected ROIs, will be one challenge of this work.

Work Plan 3a Geometric segment assignment

We will use the JSON data to greatly increase the accuracy of geometric connection algorithms. We currently are running a single set of connection algorithms based on Euclidian distance metrics. Those segments within a certain distance of a meanline are assigned to that meanline. The code then iterates and expands its search for unassigned segments until all segments are assigned. The work plan is as follows.

- Because assignment is somewhat order dependent, we will implement multiple runs with different start points and time directions, and add up results to identify best choices, as appropriate.
- We will use the timing mark map of Work Plan 2b, and combine this with the techniques already in place that implicitly identify timing marks, adding additional information for segment assignment. This also provides for some noise rejection by eliminating spurious features that do not meet the criteria for small features as timing marks AND have no open domain in the timeseries into which they can fit. This 'elimination of the negatives' is a very useful outcome of all assignment algorithms.
- We will improve the iterative assignment by adding additional rules. We will look at the size of the empty domains when judging which meanline the segment should be assigned to. Some overlap is currently allowed because the code currently looks only at the mean x-value of a segment; new rules will require all x-values to be unique. Additionally, the average and individual values of the gray scale for segments will be compared to neighboring, already-assigned segments, to check for goodness of assignment. A scheme based on (at least) standard deviation of intensities will be used as a filter.

Work Plan 3b Bayesian methods



We will use probabilistic methods to determine the likelihood of any particular segment connection coming out of an intersection or gap as being correct. The connection information will be used in the multicriteria decision matrix of Work Plan 3d. To date we have already investigated one specific intersection type, as represented in Figure 1, using real data. We will incorporate this work, using the abovementioned JSON information, to get connection information for all intersections. The method works as follows. In the 4-segment case that we have used to develop this technique, two segments come into an intersection, and two exit it. We compare the various connection possibilities: 1 to 3; 2 to 4; 1 to 4; and 2 to 3, using Bayesian methods and a fitting model for the segments.

The Bayesian method first computes the ratio of the probabilities of the first possible set of connections in the time forward direction, to the second, alternate, set of connections in the time forward direction. This forward ratio is represented as (where P is probability): P(1 >3 AND 2> 4)/P(1>4 AND 2>3)

Fig. 1 Typical intersection investigated with Bayesian methods.

The same is done in the time backwards direction, computing the backwards ratio as:

P(3>1 AND 4>2)/P(4>1 AND 3>2)

We used a Kalman filter to model the segments. The advantage of this probabilistic method is that any number of filters and fits can be tried, in order to best suit seismic data[14]. An extrapolation of the modeled path returns a distribution on the predicted points in the path. Typically this is Gaussian, but can be altered depending on requirements. From this distribution a probability for each path is determined.

Using this information, Figure 2 shows a plot of odds ratios, using a log scale as is convention. What is desired for good connections are either high ratios both forward & backwards or low ratios both forward & backwards. This corresponds to wanting things in the upper right quadrant or lower left quadrant, and indicates either the connection in the numerator or denominator, respectively, is most likely.

Figure 2 shows a plot obtained using data from 57 connections, using data from SKATE. There are lines on the plots at x=+1/1 and y=+1/1 to better identify the quadrants. This corresponds to odds ratios > 2 or < 0.5 which is somewhat arbitrary but on inspection represents a good boundary value[15]. The cross-shaped region enclosed by these boundaries represent 'bad' regions, meaning a failure for either forward or backward paths to be likely.

The plot shows all the intersections that were tested, with most residing and clustered in the upper right quadrant, indicating not only a good fit, but also validating the approach[16]. In other words, if a significant number of intersections resided in the 1 < (x,y) < 1 'bad' region, that would have represented an inability of the method to discriminate between the choices of paths.

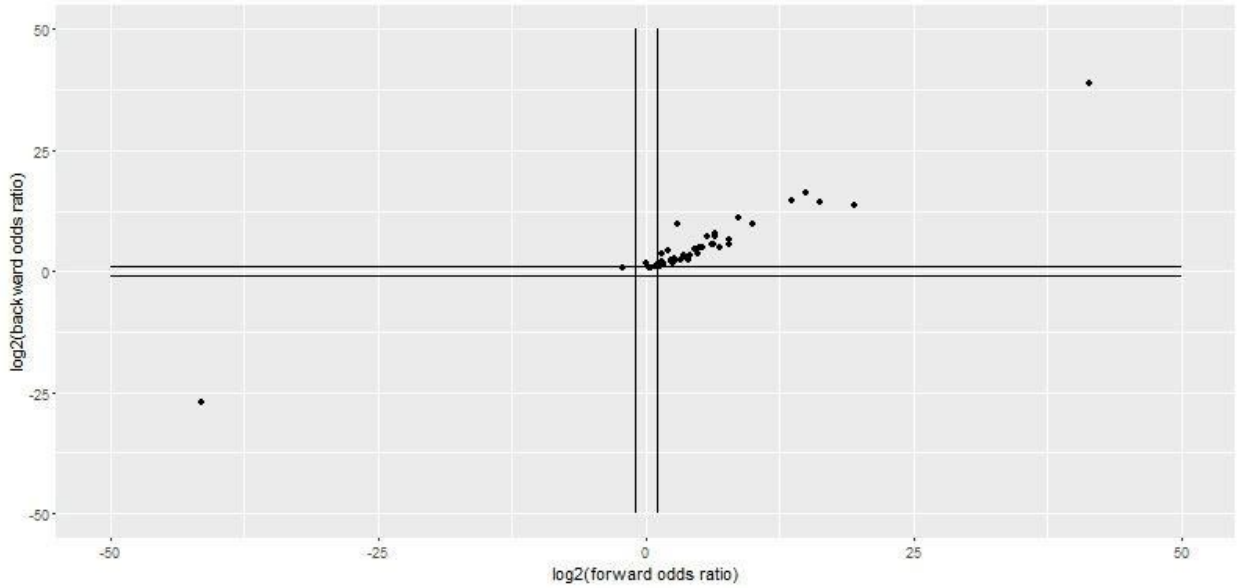


Fig. 2 Probability ratios for possible connections.

This Phase I work will use this approach on at least the subset of questionable intersections, if not all intersections and gaps. Computationally it is very fast, particularly compared to connection methods that rely on calculations in image space. This again highlights the advantage of working in data space vs. image space; it is significantly faster and therefore much less expensive. We will generalize the code to handle N-segment intersections (where N will typically be a maximum of 6). We will investigate the weighting of the path choices based on the odds ratio and use this value in the multicriteria selection method as discussed in Work Plan 3d.

[14] For example, we tried financial filters and had less success because, unlike seismic data, financial data does not revert to the mean over time.

[15] When intersection type didn't fit the idealized model, which occasionally happened for various reasons, the method invariably failed. Note that this wasn't a failure of the method, only of the assumed intersection type that we were modeling.

[16] We would expect to see some in the lower left as well, but believe that that the ordering of the segment data as outputted by SKATE skewed results towards Quadrant I.

Work Plan 3c MTSP connection methods

MTSP (Multiple Traveling Salesman Problem) represents all segments in a seismogram as nodes on a graph that need to be connected. We will use an 'ant algorithm' on the graph to cast this as a multiple traveling salesman problem with the number of salesmen equal to the number of meanlines. We will represent each segment as a node with a domain over which it exists. An advantage of this technique is that it doesn't require any intersection information; it works with all segments regardless of their start and end points.

The nodes/segments have internal properties of period, amplitude and phase, gray scale and more, which we already obtain and store for use as needed. One important constraint to leverage is that two nodes with overlapping time domains are not connectable and can not be on the same trace line. Another constraint is the requirement that the total integral of the path with respect to its mean, or zero-energy line, is equal to zero.

The nodes are randomly connected to other nodes within some given time and amplitude distance threshold (limiting the search field), creating a graph of multiple paths from one side of the seismogram to the other. Any path can only be traversed once, and can only be claimed by a single agent (ant) at a time. A solution/tour is judged first on whether all agents make it across the whole seismogram which will immediately eliminate many bad or impossible tours and which will reinforce certain edges/routes. Next, tours will be judged and their paths reinforced based on shortest distance traveled. "Distance" can include not just pixel distance but also closeness of fit in, e.g., fourier space, intensity values, and other metrics. By running the algorithm multiple times, those most likely paths will become reinforced, and unlikely or impossible paths such as those that end up associated with a different mean line will be eliminated. Figure 3 shows the results of one such set of multiple tours, with potential paths shown in white. (No tours are on the bottom half because the seismogram is quiescent there.) One advantage of this technique is computational speed, as simple matrix calculations drive the path; we can run this hundreds or thousands of times very quickly. We also anticipate that this will help to eliminate noise by not requiring every segment (meaning a segment could be noise and not real data) to have an assignment.

All likely connections will be assigned a weighting value, whose value will be determined in part by the outcome of the MTSP tour itself[17]. These results will be used for the overall path selection criteria in Work Plan 3d.

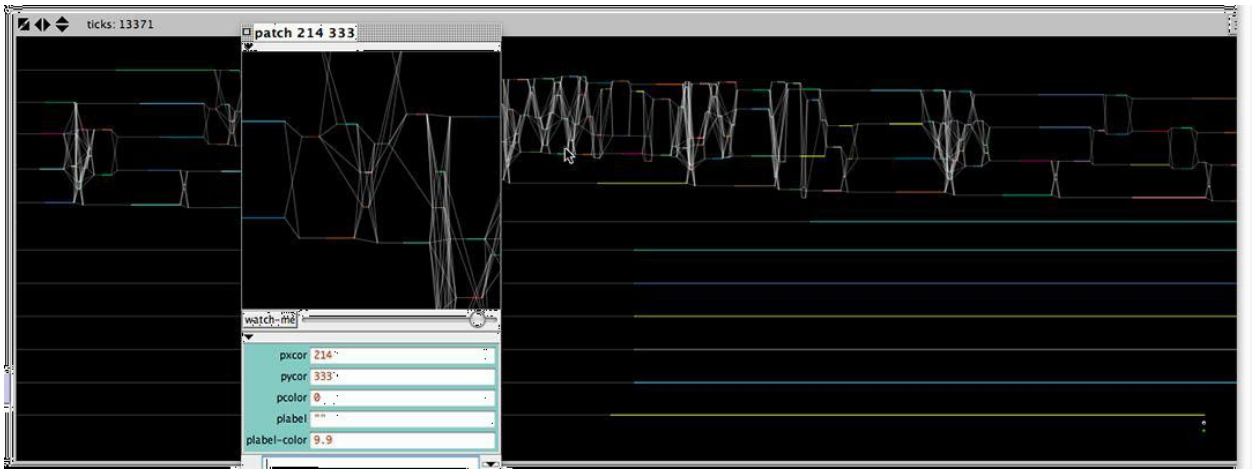


Fig. 3 Possible paths an agent can take traversing a seismogram. All possible paths are shown.

Work Plan 3d Multi criteria decision making

All of the connection metrics result in a list of possible paths for all segments that then need to be combined into a single combined solution. A multi criteria decision matrix will be constructed that collates all N segments into a single N x N matrix.

Figure 4 is an idealization of this matrix. The rows will be all segments, 1...N, and represent the end of a segment that needs a connection forward in time. The columns are again all segments, 1… N, that represent the beginnings of segments that need a connection backwards in time. Each possible connection will have a score assigned to it based on a

[17] Initially all tours will have equal weights until enough outcomes give information for weighting schemes.

summation of scores. We will investigate weighting of connection scores, but will start with equal weighting until methods suggested by the data become available.

The table has '**X**' values for those impossible connections, i.e., those connecting backwards in time. The locations listed as 'not allowed' represent some value in time forward space where it is unlikely that a connection will be found. Once all possible segment connections are made we will assign all connected segments to the nearest mean line; domains for large numbers of connected segments will be easier to locate. Unconnected segments will occur; comparison with empty domains, comparison against noise values, and iteration of this solution with any combination of connection algorithms are several of the ways to check the solution.

We will output the data into a .csv or other similar type file, utilizing our current standard of columnar x-y data for each meanline. This output will then be used in the final time series as discussed in Work Plan 2c.

	1	2	3	4	5	6	7	8	9	...	N
1	X	Value 1a	Value 2a	Value 3a	Value 4a	Value 5a	Not allowed	Not allowed	Not allowed	Not allowed	Not allowed
2	X	X	Value 1b	Value 2b	Value 3b	Value 4b	Value 5b	Not allowed	Not allowed	Not allowed	Not allowed
3	X	X	X	Value 1c	Value 2c	Value 3c	Value 4c	Value 5c	Not allowed	Not allowed	Not allowed
4	X	X	X	X	Value 1d	Value 2d	Value 3d	Value 4d	Value 5d	Not allowed	Not allowed
5	X	X	X	X	X	Value 1e	Value 2e	Value 3e	Value 4e	...	Not allowed
etc	...	...	..	...	...	...	...	...	...	...	...

Fig. 4 Multi-criteria decision making metric. Values are sums of (possibly) weighted inputs.

Work Plan 4 Long term operation of SKATE

Work Plan 4a. SKATE persistence

SKATE's usefulness will be significantly increased by the work discussed in Work Plans 1-3. Our goal is at the completion of Phase I to have in place a useful tool for daily use that will gather users and relevance over time. Work Plan 4 will include changes our web site to reduce maintenance, and address code rot by freezing versions, and developing an EC2 image (as much as is possible) that will have correct versions and libraries to ensure long term viability and operation.

The SKATE website currently has 154,654 scanned, full resolution WWSSN images in its archive. Of these, 17,915 have been processed up to the point of editing[18]. Almost 50% of the operational costs of SKATE are for standard S3 storage. We will significantly reduce costs by transferring all unprocessed and problematic seismograms to Amazon Glacier storage. These images will still be available to end users for on-demand processing as discussed in Work Plan 1, but under the download constraints imposed by Amazon.

[18] There are an additional 12,930 that have been processed but labeled by us a 'problematic.' These represent successful processing but have passed a self-imposed limit on the number of segments found. Large numbers of segments can result in very slow computer response and were therefore tagged as such. It is expected that improvements in the Pipeline will greatly reduce the number of so-called problematic segments.

Estimated monthly costs for this change in storage protocol, while preserving the EC2 instance to allow for processing user-supplied images, and the editing and downloading of existing and already processed data and images will be about $156.00. Therefore, maintaining the SKATE website beyond the termination of this particular proposal will allow users to continue to utilize this software, at a total 5 year budgeted cost of approximately $9360. Amazon credits will be purchased to ensure continuous operation. And, we will continue to solicit funding for website operations; currently, the USGS has funded almost one year's operation. We will continue to pursue other funding opportunities to ensure SKATE's persistence.

Work Plan 4b. Publish the Python Pipeline

By publishing all of our results on our GitHub public account, we will continue our efforts to create and advertise an open source user community where upgrades to SKATE can be made and incorporated into the existing software.

We will publish the Python code, with documentation, in order to allow for any user to run the pipeline code (but not the editing) on a desktop computer. This will broaden the user base of SKATE, allow users to modify and experiment with code development, and create a free and downloadable tool for many digitization needs. We have also published all or most of the code, both the pipeline as well as the web interface, on a public GitHub account, available at https://github.com/retrievertech. The GitHub repository will be broadened to include a large volume of relevant research material and unpublished results that will allow users to better understand and improve the code. Some of this work includes Shortest Path Algorithms (SPAs) such as Dijkstra's Algorithm as a method to extract traces from images of seismograms. SPA algorithms look to find the shortest path between two points on an image, given a cost function. Also to be made available is the Java code and results for clustering-based affinity propagation[19], which has shown especially good results in locating timing marks.

Work Plan 4 will complete the justification of the DOE's investment in this software by allowing a fully functional research tool to be made available for at minimum 5 years after the completion of this Phase I work.

Performance Schedule

The performance schedule as outlined below is based on a one year schedule, but that can be modified to shorter time periods if necessary without altering the order and levels of effort.

Effort	FTE Weeks	Task	Time Frame (months.)
9%	**6**	**Add user operation of any seismogram**	**1-3**
		Run EC2 on demand	
		Input external seismograms	
		Add parameter selection	
31%	**18**	**Pipeline Improvement**	**1-6**
		ROI/meanline/noise/segment ID improvement	
		Meanline improvement	
		Noise removal	
		Intersection detection	
		Timing mark mapping	

[19] http://www.icmla-conference.org/icmla07/FreyDueckScience07.pdf

		Size discriminators. Periodicity utilization	
		Time series improvements	
41%	**27**	**Develop connection algorithms**	**2-9**
		Euclidian code improvement	
		JSON data ordering	
		Bayesian methods	
		Link to JSON data	
		Port to JavaScript (currently in Java)	
		Add additional intersection type capability	
		Output to MCDM	
		Agent based MTSP	
		Link to JSON data	
		Port to JavaScript (currently in NetLogo)	
		Output to MCDM	
15%	**14**	**MCDM**	**7-12**
		Create indexed database	
		Output to segment assignment	
4%	**3**	**Long term SKATE operations**	**1-12**
		AWS tasks (added as the work progresses)	
		Publish and manage public repositories (added as the work progresses)	
100%	**68**	**Total FTE weeks**	

Equipment

There is no equipment purchase on this contract. It is a computationally intensive effort, and all participants have the necessary resources. Our previous work has provided us with high performance computing hardware and software packages, obviating the need for additional equipment purchases. Use of Amazon Web Services greatly streamlines and simplifies hardware requirements.

Facilities

Phase I work will be carried out at a single location. We have a 1300 ft^2 office in a professional office park. It is networked and set to accommodate up to >5 full time staffers. A conference area is available for regularly scheduled meetings. We have high speed internet connectivity and in-place teleconferencing capabilities. We have previously

created and utilized agreements with other office park residents to use that space for medium sized presentations (40+ attendees.)

Phase II Funding Commitment (Commercial Contribution) [OPTIONAL]

We have not secured commercial commitments, but are continuing to develop external funding sources. To date, we have implemented an agreement with the USGS that is currently paying for a significant fraction of the yearly operational costs of SKATE, and will continue to pursue the modest funding required to keep the application hosted and running. In the meantime, Retriever Technology will continue to cover any extra costs required to keep the website running.