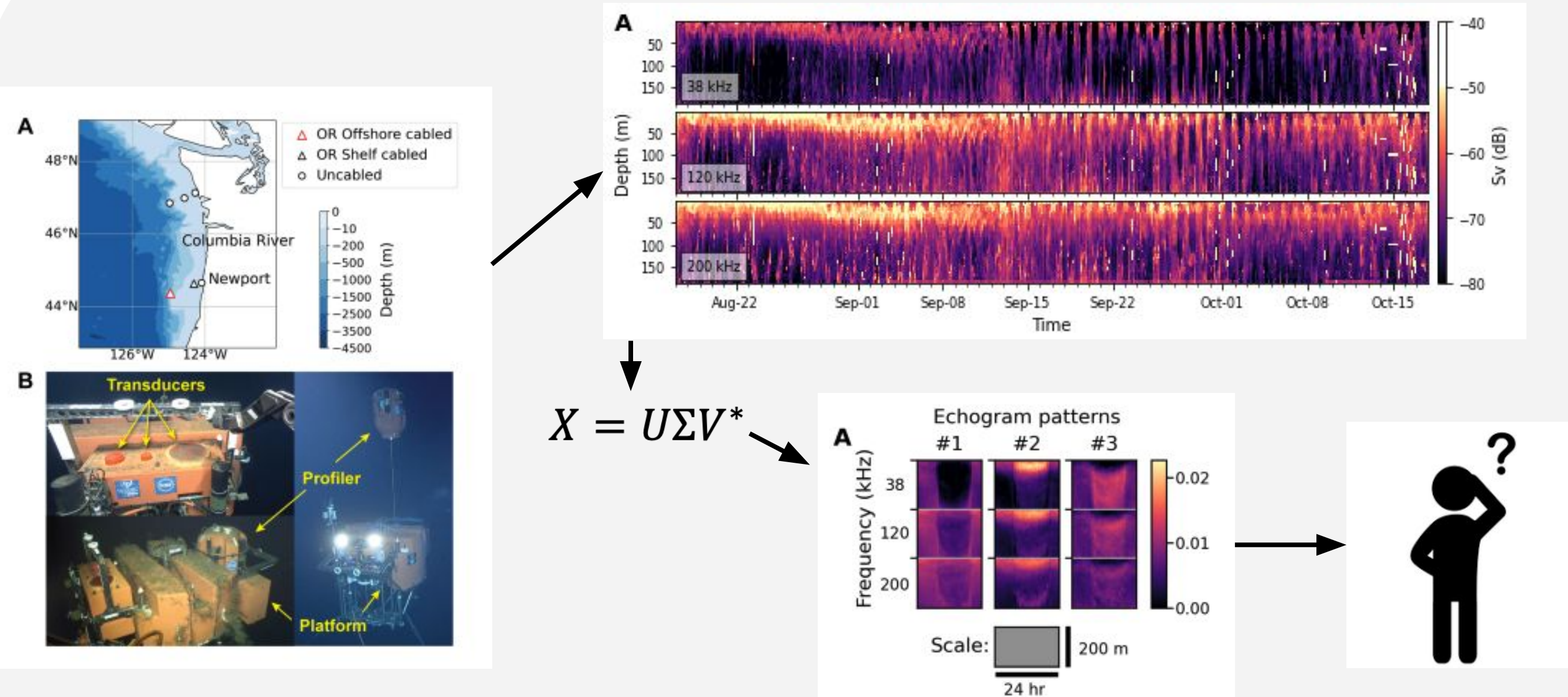


A Tool for building Intuition about Principal Component Analysis

Xinran Wang, Eason Zhang, Ruican Zhong, and Sophia Jannetty

UW CSE512: Data Visualization – Spring '22

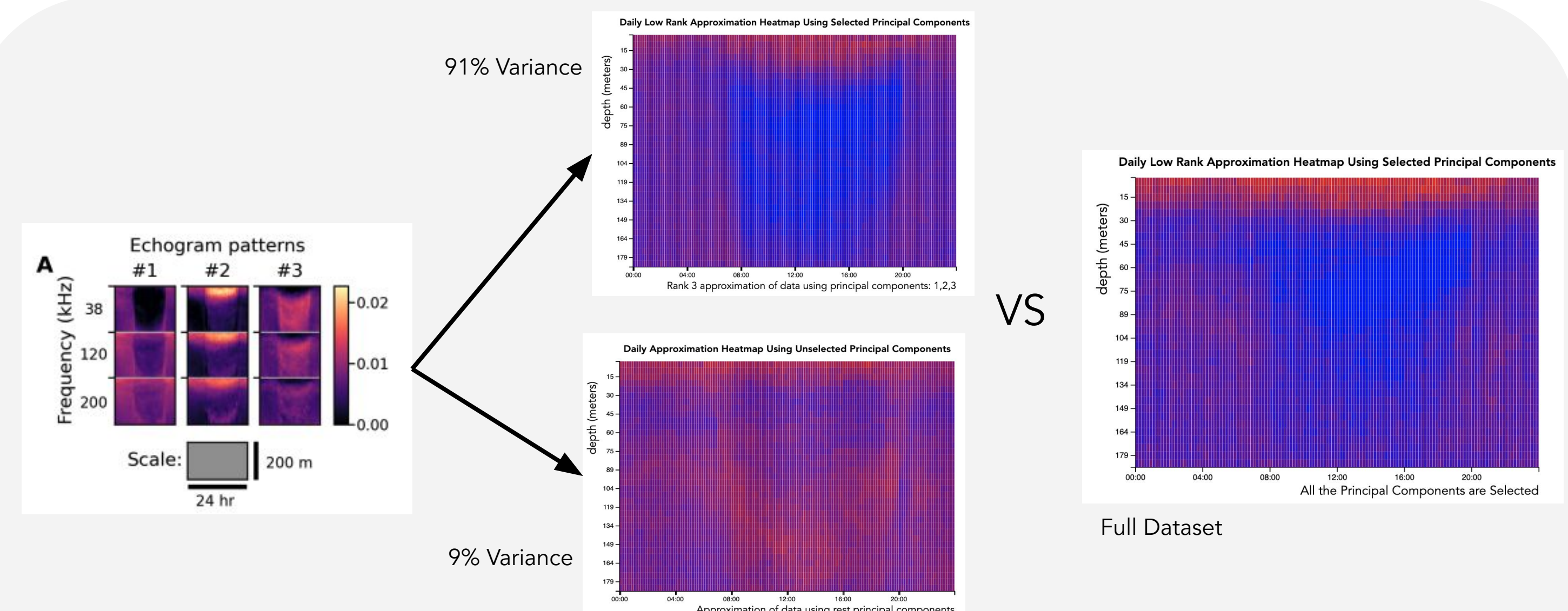
Motivation



Lee, Wu-Jung, and Valentina Staneva. 2020. "Compact Representation of Temporal Processes in Echosounder Time Series via Matrix Decomposition." The Journal of the Acoustical Society of America 148 (6). United States: 3429–42. doi:10.1121/10.0002670.

Principal component analysis is a powerful dimensionality reduction technique that is useful for extracting temporal patterns in ocean sonar time series data. Current visualization methods visualize components in isolation or low-rank approximations of datasets. This makes it difficult for researchers to get a full picture of what is happening in the dataset and makes it difficult for students to understand the meaning of individual components.

Design Goals



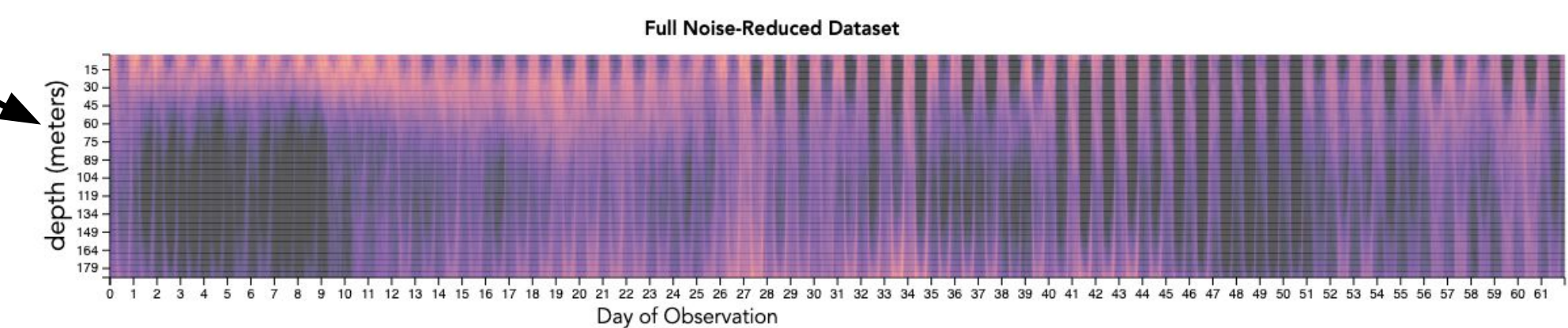
We wanted to create a visualization that helps *students* understand how principal components combine to make low rank approximations of data, and how low rank approximations constructed from principal components that capture most variance in the data more closely resemble the full dataset than approximations from principal components capturing less variance. We also wanted to let *ocean researchers* explore the principal components of daily data in Wu-Jung and Staneva's echosounder water column data.

Approach

The top graph displays the noise-reduced echosounder water column data over 62 days, but it is an overwhelming amount of data to explore

PCA: Low Rank Approximation of Ocean Sonar Dataset with Selected Principal Components

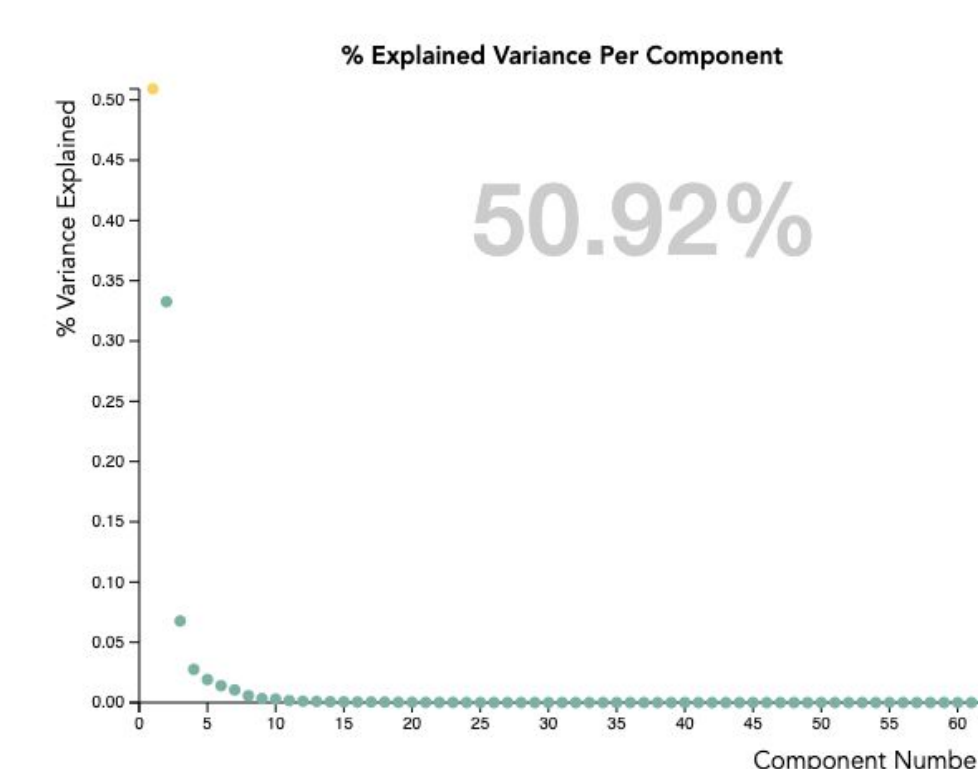
This echogram depicts noise-reduced echosounder water column data collected over 62 days. It can be challenging to identify specific patterns in a dataset this large, and especially difficult to understand what is happening over the course of one day.



Principal component analysis is a technique that identifies orthogonal vectors through a dataset that capture the most of that data's variance. We can visualize principal components in isolation to identify patterns in the data and we can combine principal components (by summing them after scaling them to their eigenvalues or singular values) to construct low-rank approximations of the dataset. This can help identify which patterns are contributing to which elements in the full dataset and can be used to reduce the storage size of the dataset (by eliminating the contributions of principal components that capture very little variance).

Below we have run principal component analysis on data from day 9 to assess what is happening in depicts the amount of variance captured by each principal component. Select a principal component by clicking a point in the variance chart on the left. The echogram on the right will depict a low-rank approximation of the data using the unselected principal components for comparison. Select multiple principal components to see a low-rank approximation of the data using all selected principal components and see the full dataset visualized in the echogram on the left. Select multiple principal components to see a low-rank approximation of the data using all selected principal components and see the full dataset visualized in the echogram on the left. From there you can click points to deselect them and see what patterns are removed from the overall data when you remove the contributions from specific principal components. Click the "Clear All" button to deselect all principal components. Explore data from all three frequencies in the dataset!

Select principal components to view the low rank approximation



Points in the variance chart can be selected. The sum of the explain variance of the selected components will be displayed in the chart. The heatmap on the left shows the low-rank approximation generated with the selected PCA components, while the one on the right shows the approximation generated with the rest of the PCA components

Press here to select all:

Select All

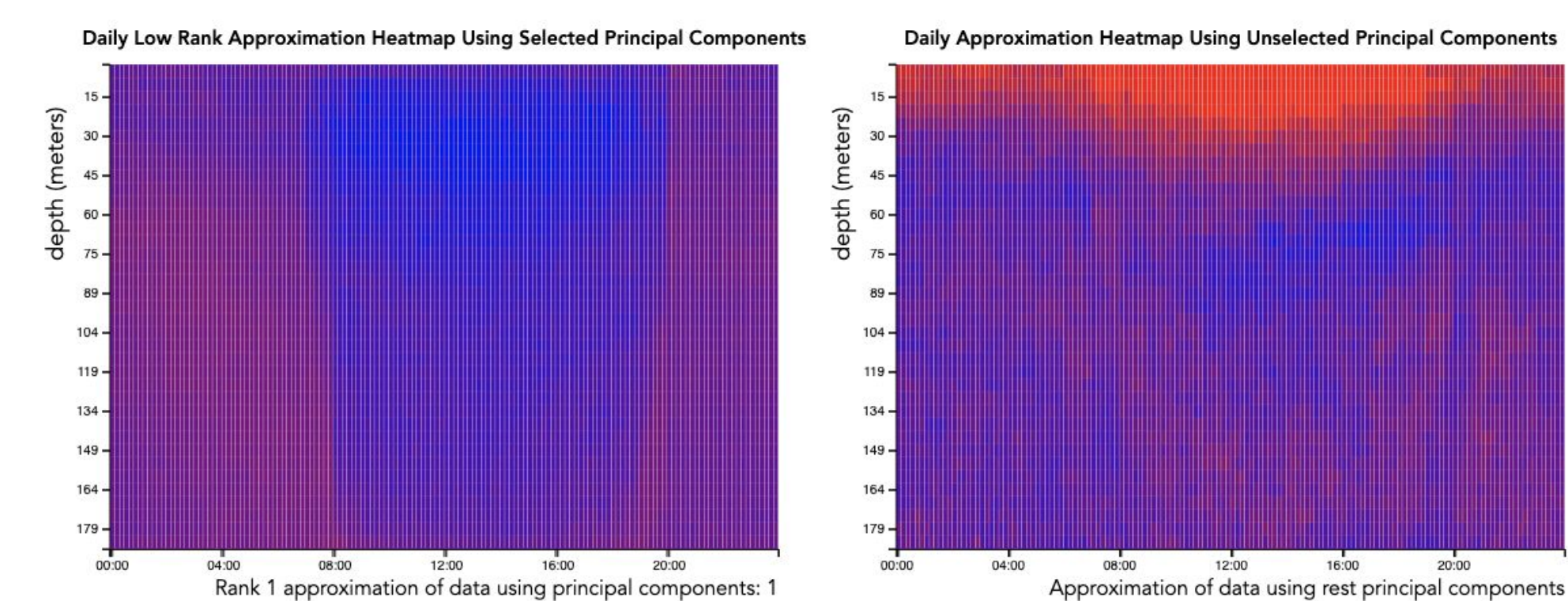
Press here to deselect all:

Clear All

Select your preferred frequency from the drop-down list:

38khz

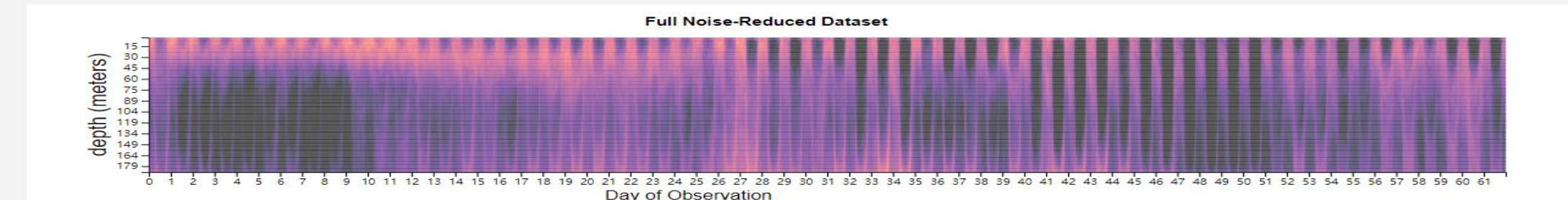
Select All and Clear All buttons allow users to quickly select or deselect many components. Frequency dropdown allows users to select which frequency data they are visualizing



Try selecting principal components that capture very little variance. How do the low-rank approximations capturing little variance compare to low-rank approximations capturing lots of variance?

Evaluation/Feedback

PCA: Low Rank Approximation of Ocean Sonar Dataset with Selected Principal Components



Principal component analysis is a technique that identifies orthogonal vectors through a dataset that capture the most of that data's variance. We can visualize principal components in isolation to identify patterns in the data and we can combine principal components (by summing them after scaling them to their eigenvalues or singular values) to construct low-rank approximations of the dataset. This can help identify which patterns are contributing to which elements in the full dataset and can be used to reduce the storage size of the dataset (by eliminating the contributions of principal components that capture very little variance).

Below we have run principal component analysis on the data with the data matrix reshaped so that each column contains all pings at all depths recorded that day (the matrix visualized above has one column per ping, our reshaped data matrix stacked all pings for each day so each column is a day). Thus the principal components capture patterns of what happens in the water column over the course of one day. The scatter plot depicts the amount of variance captured by each principal component. Select a principal component by clicking a point to see it visualized in the echogram on the left. The echogram on the right will depict a low-rank approximation of the data using the unselected principal components to show what data are missing from the visualization on the left. Select multiple principal components to see a low-rank approximation of the data using all selected principal components and see the full dataset visualized in the echogram on the left. From there you can click points to deselect them and see what patterns are removed from the overall data when you remove the contributions from specific principal components. Click the "Clear All" button to deselect all principal components. Explore data from all three frequencies in the dataset!

Press here to select all:

Select All

Press here to deselect all:

Clear All

Select your preferred frequency from the drop-down list:

38khz

These data were provided to us by Wu-Jung Lee and Valentina Staneva, were published in The Journal of the Acoustical Society of America in their 2020 paper "Compact Representation of Temporal Processes in Echosounder Time Series via Matrix Decomposition," and can be found in this repo.

In response to feedback from Dr. Staneva we made the following changes:

1. Included "ocean sonar" or "echosounder water column" in the title to be more explicit about what is used in the scientific community
2. Added a static visualization displaying all data across the 62 days to provide an overview
3. Added more explanations to clarify what data is being visualized
4. Implemented the "Select All" and "Clear All" buttons to ensure easy selection. Specifically, the "Select All" button provides the option to quickly select and visualize the entire dataset
5. Included dropdown menu to select and visualize different frequencies
6. Added references to the ooi-echo-matrix-decomposition repository

Thanks to Valentina Staneva Ph.D and Wu-Jung Lee Ph.D for supporting this project!

Link to the Interactive Visualization:
<https://cse512-22sp.pages.cs.washington.edu/Sonar-Principal-Components/>

