Discovering Phytoplankton Trends in Oceans Using Data Science

Abhigyan Kaustubh[3], Elton Dias[3] and Tanmay Modak[3]
Advisors: Prof. Bill Howe[1], Dr. Sophie Clayton[2], Jeremy Hyrkas[1]
[1]UW CSE, [2]UW eScience Institute, [3]UW Information School

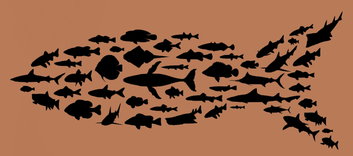
The Problem

- There is lack of data on the community structure of small phytoplankton over a large geographical area at high spatial resolution.
- SeaFlow provides us with an opportunity to capture, store and analyze information regarding microbial populations
- Most phytoplankton cannot be taximonically identified using SeaFlow Data, hence we explore newer methods using Data Science

Why study Phytoplankton?

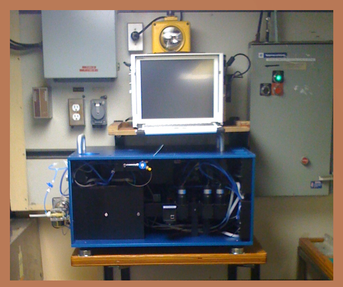
O_2

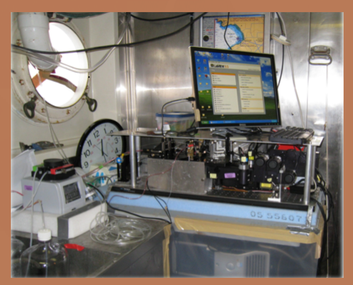
Phytoplankton produce half of the world's oxygen though photosynthesis



Phytoplankton are the base of the marine food chain

What is SeaFlow?



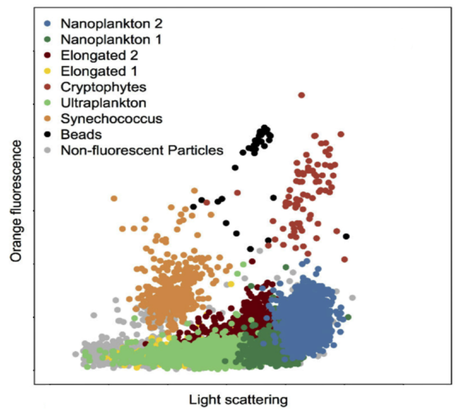


SeaFlow is a flow cytometer that is designed to study populations of microbial organisms in 3 minute intervals. The instrument collects information about the size and pigment content of an individual cell and counts several thousands cells every second in real-time by utilizing light scattering and autofluorescence properties of individual cells to discriminate and quantify different cell populations

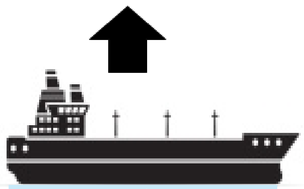
Reference

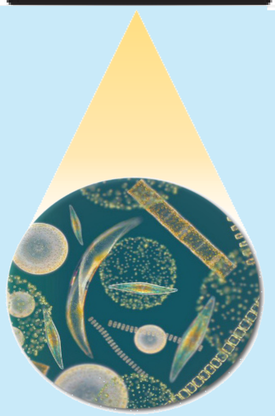
Jarred E Swalwell, Francois Ribalet, E. Virginia Armbrust, SeaFlow: A novel underway flow-cytometer for continuous observations of phytoplankton in the ocean, ASLO, USA.

Myria
Distributed Big Data management system



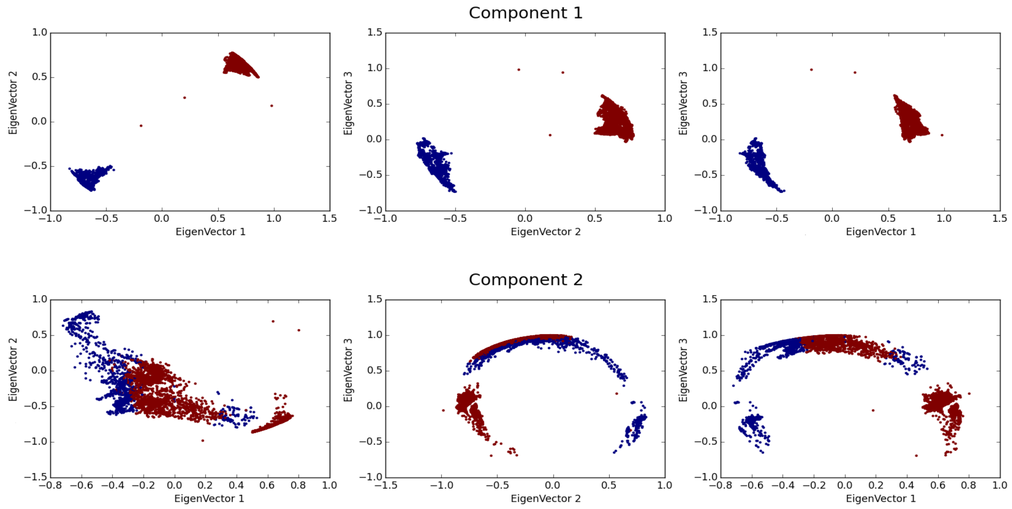
Visualization of SeaFlow data





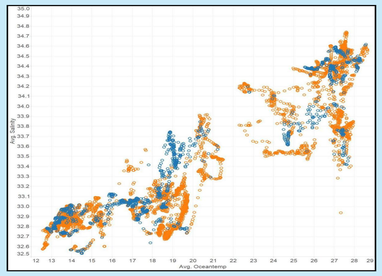
Analysis of Principal Components with Clustering

- Dimensionality Reduction using PCA to identify features that explain maximum variance
- Silhoutte Score computation of the Top Principal Component to determine best value of K
- KMeans Clustering to further discern various groups of data points and formulate relationships
- Introduced evironment variables to the clustered groups as evidence to explain observation.

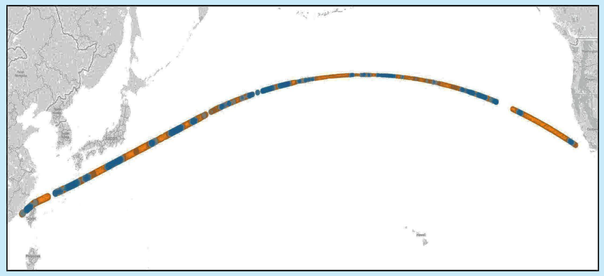


Discoveries

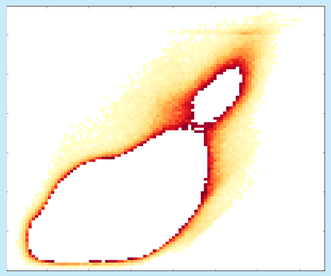
- Though there is no clear distribution , there is a stark difference in community structure visible across 2 clusters
- The first cluster indicates a community of smaller cells, whereas the second cluster indicates a community of larger cells
- Areas in the ocean showing high variance in temperature and salinity show high variance in the community structure



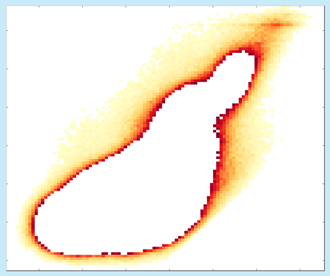
Temperature vs Salinity Plot



Cruise path tracing plankton groups.



Forward Scatter vs Chlorophyll
First Cluster



Forward Scatter vs Chlorophyll
Second Cluster