# CSP 571 Data Preparation and Analysis Project Proposal & Outline

**Group Member:**

1. Elisha Maria Krista Orlina - A20520119
2. Jayanth Pulugula - A20514528
3. Tarun Sai Yakkala - A20511629
4. Venkata Sai Neeraj Boggarapu - A20523045

**Project Title -** Assessing Long-Term Trends and Environmental Impacts of Sea Ice Melting in the Arctic Region

# Problem definition and planning:

## Description

This project aims to assess the long-term trends and environmental impacts of sea ice melting in the Arctic region using a variety of machine learning and statistical techniques. We will use data from the National Snow and Ice Data Center (NSIDC) to identify trends in sea ice melting over time and to investigate the environmental impacts of sea ice melting on Arctic ecosystems, marine life, and human communities.

## Research Goal:

The research goal of this project is to develop a comprehensive understanding of the long-term trends and environmental impacts of sea ice melting in the Arctic region. This information is essential for developing effective strategies to mitigate the negative impacts of sea ice melting.

## Specific Questions:

The following are the specific questions that this project will address:

- What are the long-term trends in sea ice melting in the Arctic region?
- How have the rates of sea ice melting changed over time?
- What are the environmental impacts of sea ice melting?
- How can we use machine learning and statistical techniques to improve our
- understanding of sea ice melting and its environmental impacts?

## Data source and Reference Data

## Proposed Methodology

We will use the following methodology to address the specific questions of this project:

1. Data preparation: We will clean the data from NSIDC and remove any outliers. We will also transform the data into a format that is suitable for our analysis.
2. Data exploration: We will use exploratory data analysis (EDA) techniques to identify trends, patterns, and anomalies in the data. We will also use EDA to identify the most important features for predicting sea ice melting.
3. Model development: We will develop a variety of machine learning and statistical models to predict sea ice melting. We will use a variety of model selection techniques to choose the best model for our data.
4. Model evaluation: We will evaluate the performance of our models on a held-out test set. We will use a variety of evaluation metrics, such as accuracy, precision, recall, and F1 score, to assess the performance of our models.
5. Environmental impact analysis: We will use our models to predict the environmental impacts of sea ice melting on Arctic ecosystems, marine life, and human communities. We will use a variety of data sources, such as scientific papers, government reports, and news articles, to gather information on the environmental impacts of sea ice melting.
6. Communication: We will communicate our findings to others by writing a report, giving a presentation, or creating an interactive visualization.

## Metrics for Measuring Analysis Results:

We will use the following metrics to measure the results of our analysis:

- Accuracy: This metric measures the proportion of predictions that are correct.
- Precision: This metric measures the proportion of positive predictions that are correct.
- Recall: This metric measures the proportion of actual positive cases that are predicted correctly.
- F1 score: This metric is a harmonic mean of precision and recall.
- Correlation coefficient: This metric measures the strength of the relationship between two variables.
- R-squared: This metric measures the proportion of the variation in one variable that can be explained by another variable.
- Mean squared error: This metric measures the average of the squared differences between the predicted values and the actual values.

## Project Outline

## Literature Review and Related Work:

We will review the existing research on sea ice melting and its environmental impacts. We will focus on the following areas:

- Trends in sea ice melting
- Environmental impacts of sea ice melting
- Mitigation strategies for sea ice melting
- Machine learning and statistical techniques for sea ice prediction

## Data Sources and Reference Data:

The dataset we have considered for this project has 24 csv files out of which 12 are for the northern part of the icecaps and 12 are for the southern part of the ice caps. So each dataset contains ice cap related data for each month. So the first task we have to do with this data is to combine and prepare the data. In this case we have combined the data for the north part of the dataset and the south part of the dataset separately. After combining the dataset if there are any missing values in the extent feature column and area feature column we have imputed it with average of the area there are some missing values in the dataset in order to overcome this we have chosen to impute the missing values with the average of all values of the same month.

We will use the following data sources for our analysis:

- Sea Ice Index, Version 3 from the National Snow and Ice Data Center
- Scientific papers on sea ice melting and its environmental impacts
- Government reports on sea ice melting and its environmental impacts
- News articles on sea ice melting and its environmental impacts

## Data Processing and Pipeline:

We will perform the following data processing steps:

- Cleaning: We will remove any errors or inconsistencies from the data.
- Imputing: We will fill in any missing values in the data.
- Transformation: We will transform the data into a format that is suitable for our analysis.
- Feature engineering: We will create new features from the existing data that may be more informative for predicting sea ice melting.
- Outlier detection: We will identify and remove any outliers from the data.

## Data Stylized Facts:

We will perform the following data analysis tasks to identify stylized facts:

- Distributional analysis: We will examine the distribution of the data to identify any patterns or anomalies.
- Clustering: We will group similar observations together using clustering algorithms.
- Dimensionality reduction: We will reduce the dimensionality of the data using dimensionality reduction techniques.
- Feature selection

## Model selection:

More specifically apart from using SVM, Linear regression, random forest and gradient boosting machines (GBMs) we are also using ARIMA.

(ARIMA stands for Autoregressive Integrated Moving Average. It is a statistical model that is used to analyze and forecast time series data. ARIMA models are widely recognized and effective in predicting time series data like sales, prices, or meteorological patterns. ARIMA, is used to comprehensively capture data patterns, trends, and seasonal variations by utilizing past values, differences, and residuals. As the data we are working on is on ice caps dataset we are using ARIMA model to forecast the data.)

## Software Packages and Applications:

- R Language for data analysis and visualization.
- R Studio as the integrated development environment for data analysis.

## Libraries:

The libraries we will be using to forecast the data in this project are as follows:

tidyverse
dplyr
lubridate
zoo
ggplot2
tseries
forecast
fpp
vars
TSA

## References:

Moon, T. A., Overeem, I., Druckenmiller, M., Holland, M., Huntington, H., Kling, G., et al. (2019). The expanding footprint of rapid Arctic change. Earth's Future, 7, 212–218. https://doi.org/10.1029/2018EF001088

Steig, Eric J., David P. Schneider, Scott D. Rutherford, Michael E. Mann, Josefino Comiso, and Drew T. Shindell. "Warming of the Antarctic ice-sheet surface since the 1957 International Geophysical Year." Nature 457, no. 7228 (2009): 459-462.

Turner, J., Colwell, S.R., Marshall, G.J., Lachlan-Cope, T.A., Carleton, A.M., Jones, P.D., Lagun, V., Reid, P.A. and Iagovkina, S. (2005), Antarctic climate change during the last 50 years. Int. J. Climatol., 25: 279-294. https://doi.org/10.1002/joc.1130