# CSSI Element: Comprehensive Time Series Data Analytics for the Prediction of Solar Flares and Eruptions

PI: Rafal A. Angryk (Computer Science), Co-PIs: Manolis K. Georgoulis (Heliophysics)
Institution: Georgia State University

Award #: 1931555

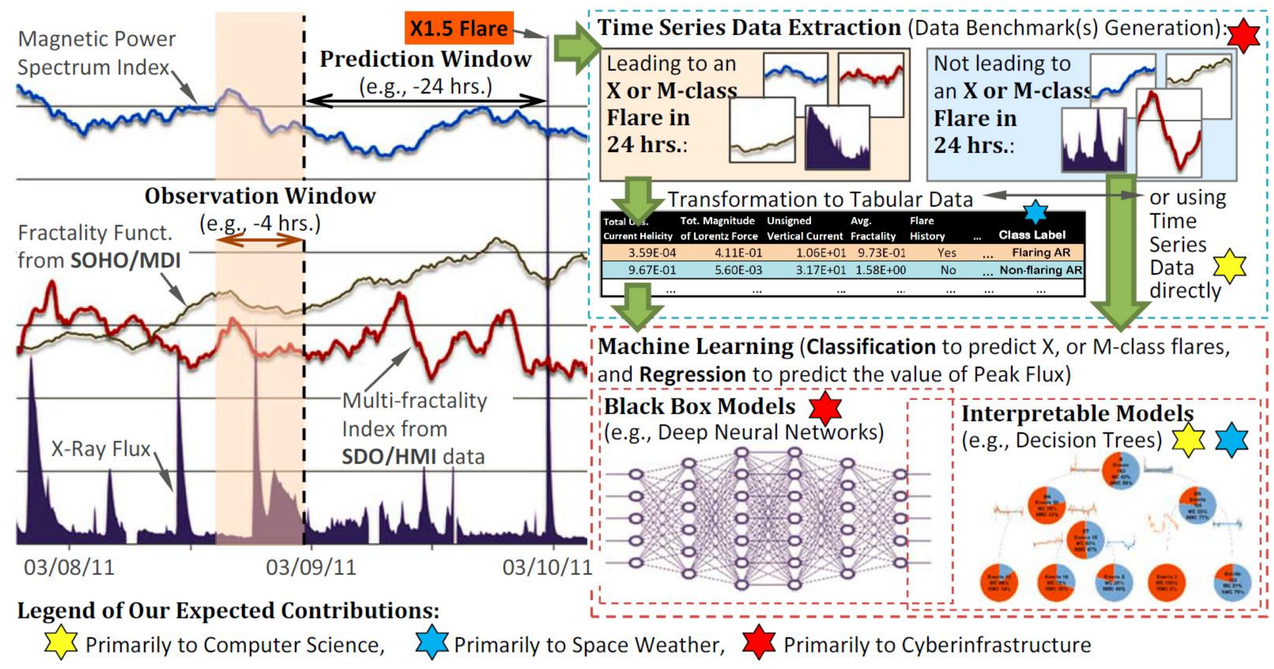
## Abstract

We report on progress made by our interdisciplinary Data Mining Lab at Georgia State University on this recently funded (October 1, 2019) project. We present brief overview of our project and focus on the first two phases of our research: (1) Data & Metadata Acquisition, and (2) Generation of Data Sets for Benchmarking.

## Motivation

Solar flares, along with accompanying solar eruptions, have the potential to disrupt the technology we rely on, such as GPS, radars, high-frequency radio communications, communication satellites (cell phones and Internet), and electricity distribution networks.
Our objectives are:
(1) to improve understanding of the time-dependent physical behavior of solar active regions to the point that we can predict whether, when and how strongly they will flare; and
(2) in doing this, to perform comparative, reproducible, and data-driven prediction of solar magnetic eruptions.

## Overview of Our Project

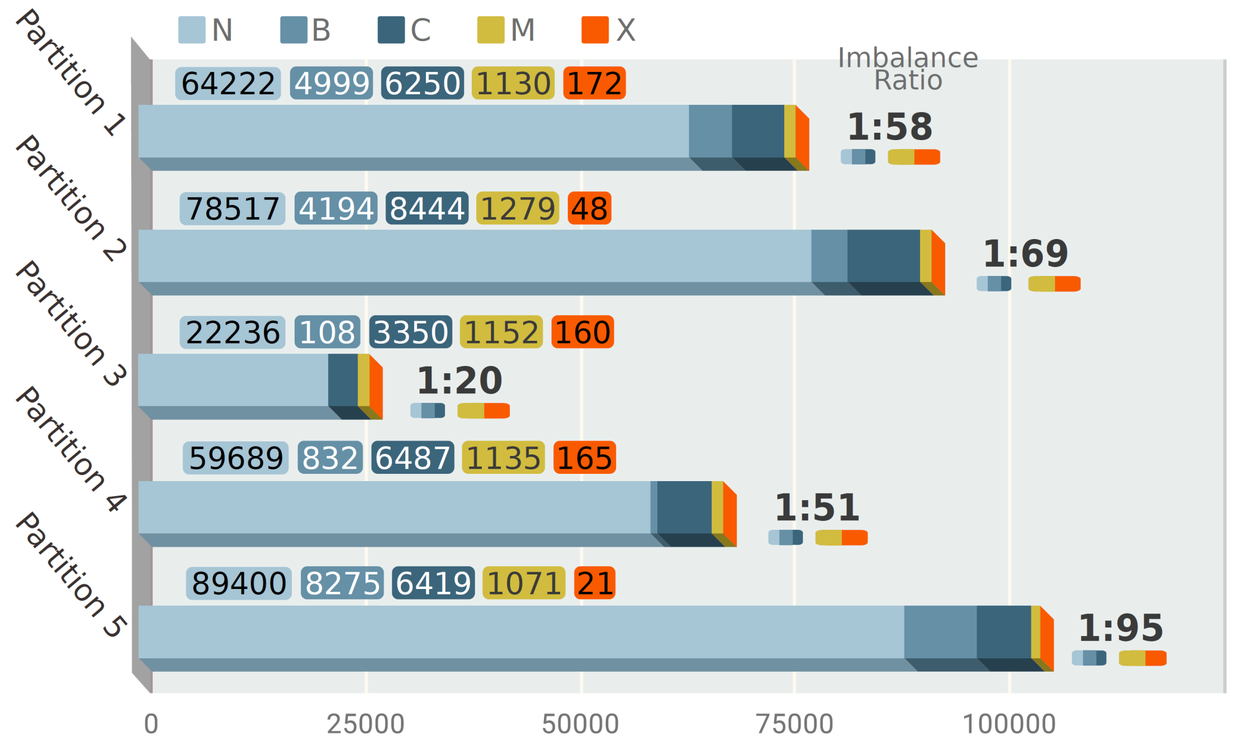



## SWAN-SF Dataset

To conduct our project we started from cleaning and integrating data from multiple solar data repositories. Our Space Weather ANalytics for Solar Flares (SWAN-SF) Dataset: (1) provides benchmarking ML models for flare prediction; (2) uses multivariate metadata time series from solar photospheric magnetograms of the Space weather HMI Active Region Patch (SHARP) series; (3) covers 4,098 MVTS data instances; (4) spans over 8 years (May 2010 – August 2018); (5) includes 51 parameters; and (6) integrates 13,641 flare reports.

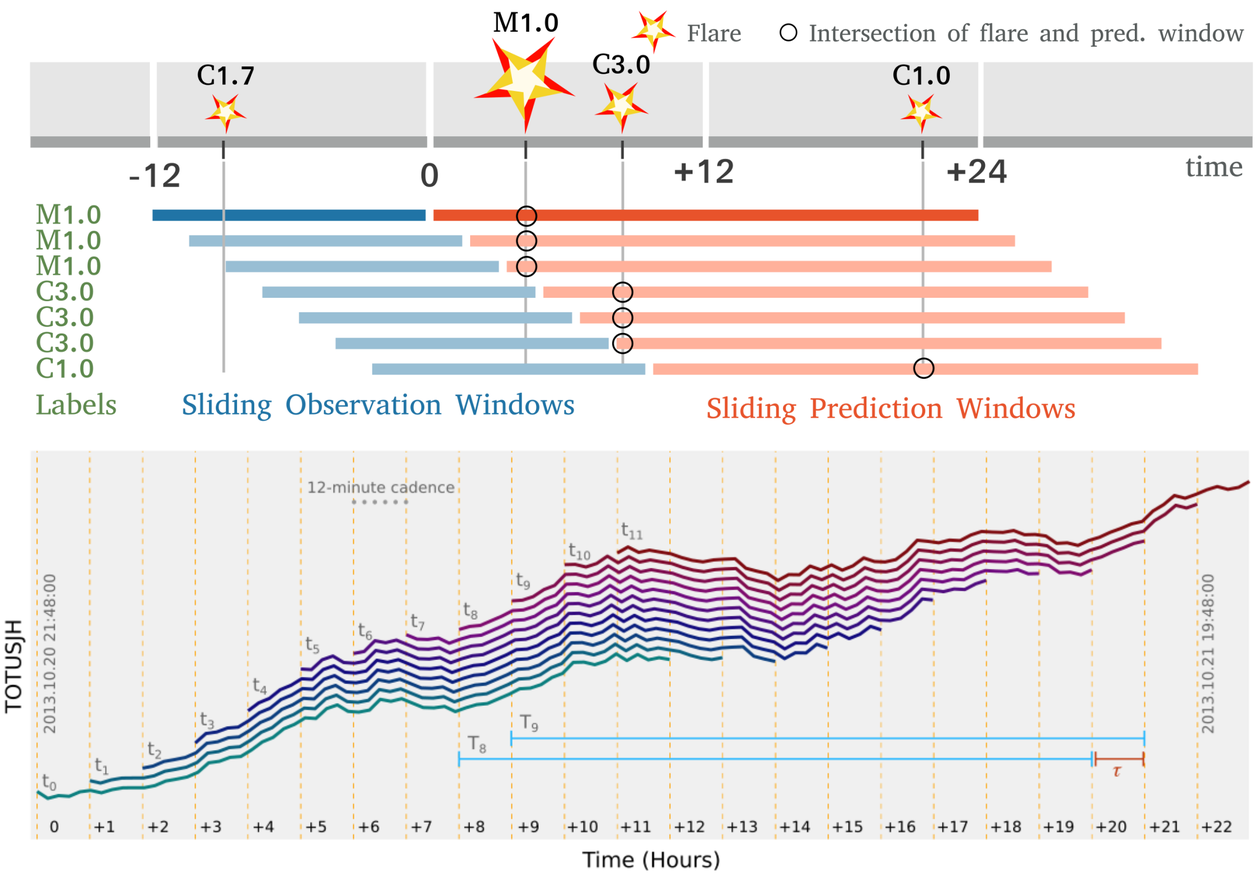



## Extreme Class Imbalance

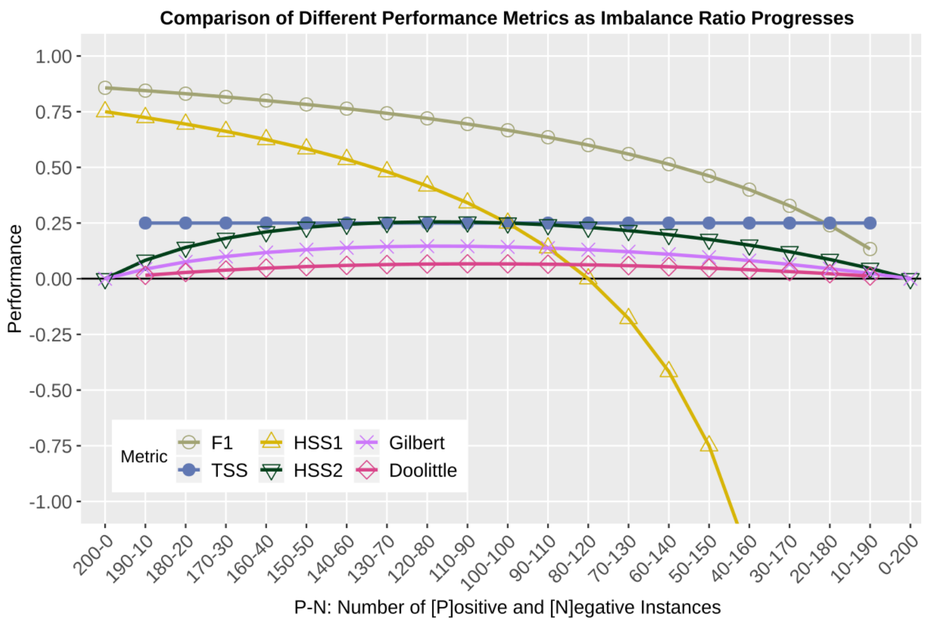



If we focus on binary classification, the imbalance gets event worse (e.g., Class X vs. Classes M, C, B and N in the 5th fold: 1:5,000)

## Temporal Coherence




## First ML Results



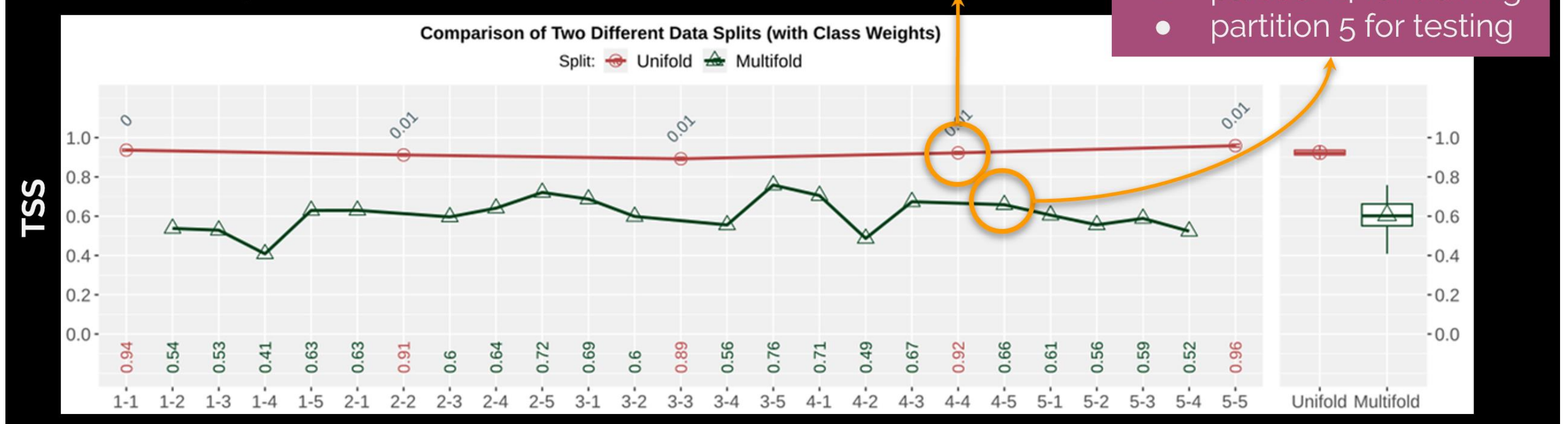

**On the left:** Susceptibility of various ML performance metrics to the class imbalance.

**At the bottom:** The impressive True Skill Statistic (TSS) of the red curve come from the training and testing on the same partition.

The impressive TSS is the direct result of ***temporal coherence***.

randomly sampled from partition 4:
- ⅔ for training
- ⅓ for testing

Used entire
- partition 4 for training
- partition 5 for testing




dmlab.cs.gsu.edu