Big Data 2020 Tutorial

Data Sources, Tools, and Techniques

Big Data-Driven Machine Learning in Heliophysics

A. Ahmadzadeh | D. J. Kempton | B. Aydin | R. A. Angryk

13 Dec 2020

www.dmlab.cs.gsu.edu
Data Mining Lab @ Georgia State University

Plan of Talk

(10 mins)

Introduction

Some details should be added here. This is just a placeholder. Some details should be added here. This is just a placeholder. Some details should be added here. This is just a placeholder.

(20 mins)

Image Parameter Dataset

■ Importance of Solar Observatories, ■ Solar Dynamics Observatory, ■ Image Data UI: Helioviewer, ■ Spatiotemporal Data: Sunpy, ■ Our Image-parameter Dataset, ■ The API & Applications

(30 mins)

Integrated Solar Event Database

(30 mins)

Data Wrangling for SWx Forecasting

Space Weather Forecasting * Rare-event Prediction * Role of Benchmark Datasets * Best Practices

(20 mins)

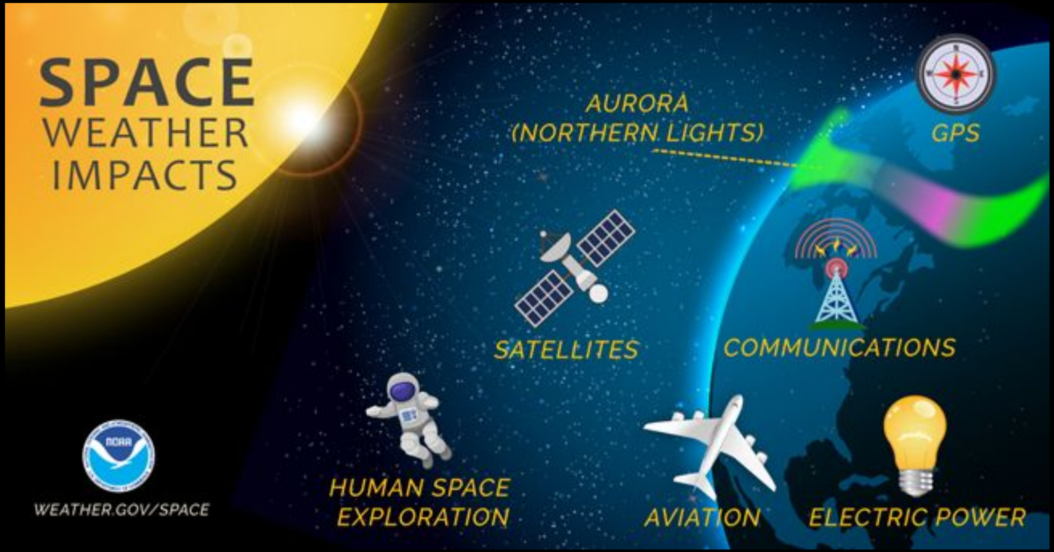
MVTS Data Toolkit

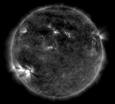
■ Basic Analysis of Raw Data, ■ Feature Extraction, ■ Sampling of MVTS data, ■ Normalization of MVTS data.

Data Wrangling for Space Weather Forecasting

Space Weather Forecasting

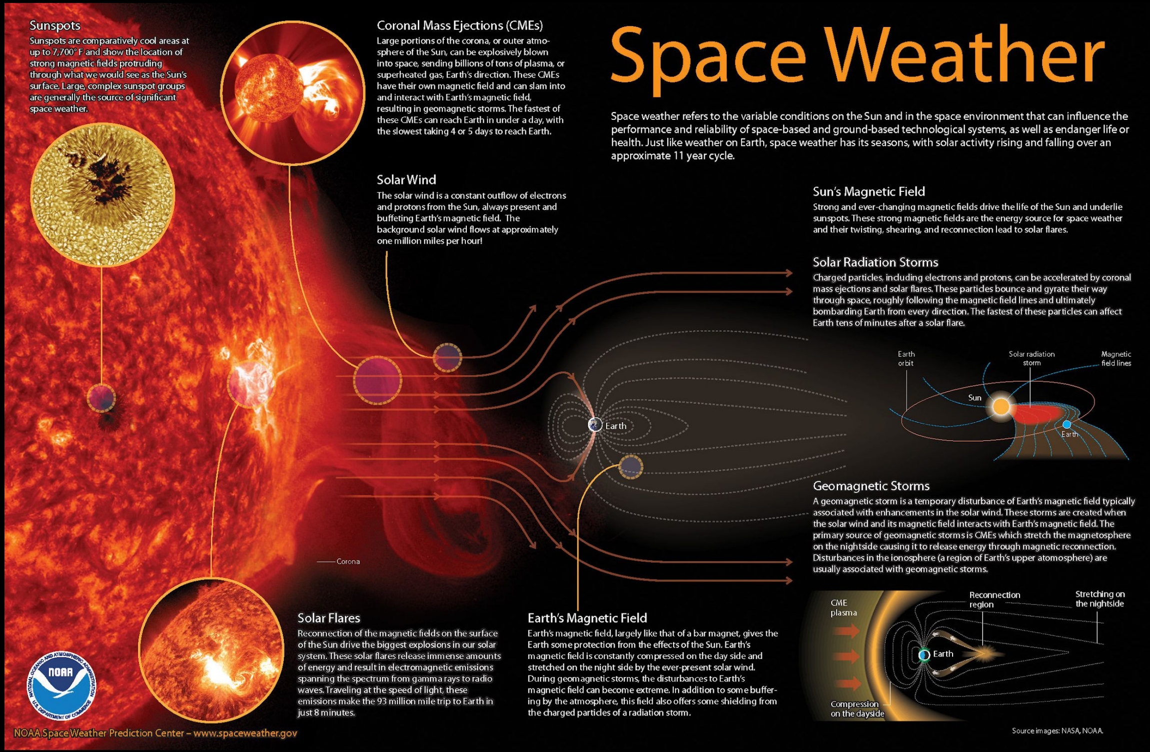
- Weather in space
- Solar storms (geomagnetic storms, solar radiation storms, radio blackouts)





Space Weather Forecasting

- Scale
- Variety
- Model-based vs ML-based
- Imbalance

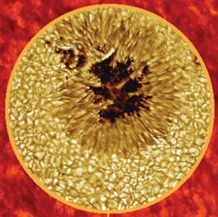


Space Weather

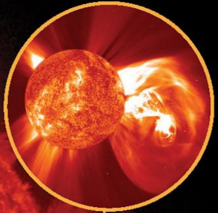
Space weather refers to the variable conditions on the Sun and in the space environment that can influence the performance and reliability of space-based and ground-based technological systems, as well as endanger life or health. Just like weather on Earth, space weather has its seasons, with solar activity rising and falling over an approximate 11 year cycle.

Sunspots

Sunspots are comparatively cool areas at up to 7,700° F and show the location of strong magnetic fields protruding through what we would see as the Sun's surface. Large, complex sunspot groups are generally the source of significant space weather.



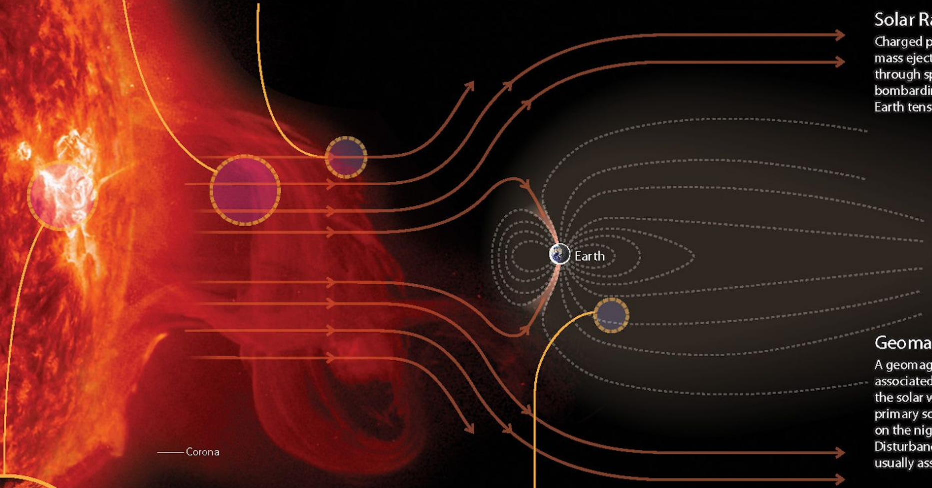
Coronal Mass Ejections (CMEs)



Large portions of the corona, or outer atmosphere of the Sun, can be explosively blown into space, sending billions of tons of plasma, or superheated gas, Earth's direction. These CMEs have their own magnetic field and can slam into and interact with Earth's magnetic field, resulting in geomagnetic storms. The fastest of these CMEs can reach Earth in under a day, with the slowest taking 4 or 5 days to reach Earth.

Solar Wind

The solar wind is a constant outflow of electrons and protons from the Sun, always present and buffeting Earth's magnetic field. The background solar wind flows at approximately one million miles per hour!

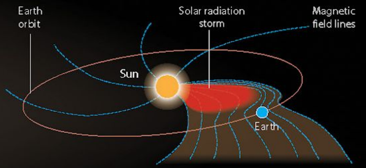


Sun's Magnetic Field

Strong and ever-changing magnetic fields drive the life of the Sun and underlie sunspots. These strong magnetic fields are the energy source for space weather and their twisting, shearing, and reconnection lead to solar flares.

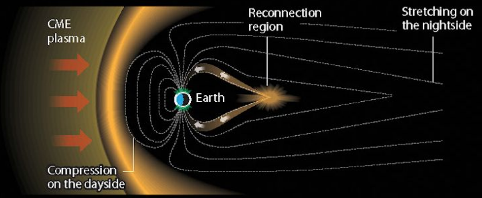
Solar Radiation Storms

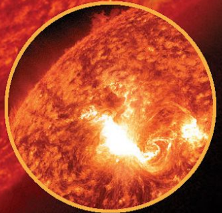
Charged particles, including electrons and protons, can be accelerated by coronal mass ejections and solar flares. These particles bounce and gyrate their way through space, roughly following the magnetic field lines and ultimately bombarding Earth from every direction. The fastest of these particles can affect Earth tens of minutes after a solar flare.



Geomagnetic Storms

A geomagnetic storm is a temporary disturbance of Earth's magnetic field typically associated with enhancements in the solar wind. These storms are created when the solar wind and its magnetic field interacts with Earth's magnetic field. The primary source of geomagnetic storms is CMEs which stretch the magnetosphere on the nightside causing it to release energy through magnetic reconnection. Disturbances in the ionosphere (a region of Earth's upper atomosphere) are usually associated with geomagnetic storms.





Solar Flares

Reconnection of the magnetic fields on the surface of the Sun drive the biggest explosions in our solar system. These solar flares release immense amounts of energy and result in electromagnetic emissions spanning the spectrum from gamma rays to radio waves. Traveling at the speed of light, these emissions make the 93 million mile trip to Earth in just 8 minutes.

Earth's Magnetic Field

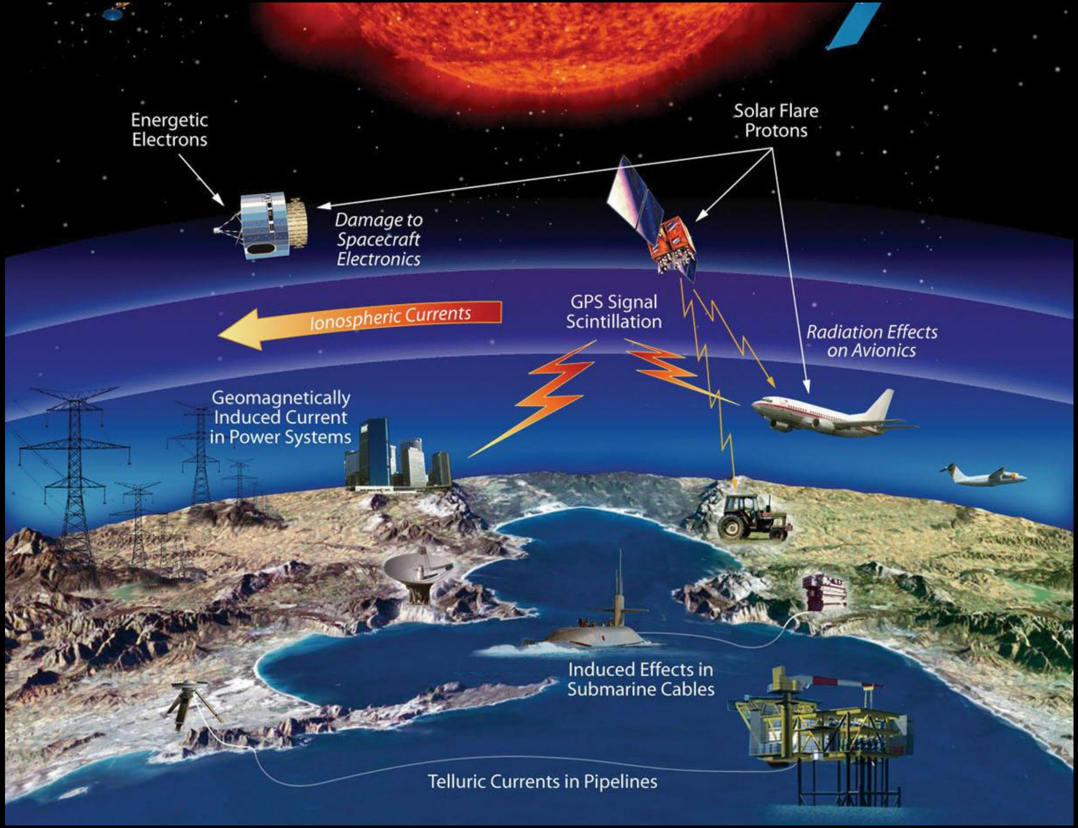
Earth's magnetic field, largely like that of a bar magnet, gives the Earth some protection from the effects of the Sun. Earth's magnetic field is constantly compressed on the day side and stretched on the night side by the ever-present solar wind. During geomagnetic storms, the disturbances to Earth's magnetic field can become extreme. In addition to some buffering by the atmosphere, this field also offers some shielding from the charged particles of a radiation storm.

NOAA Space Weather Prediction Center – www.spaceweather.gov

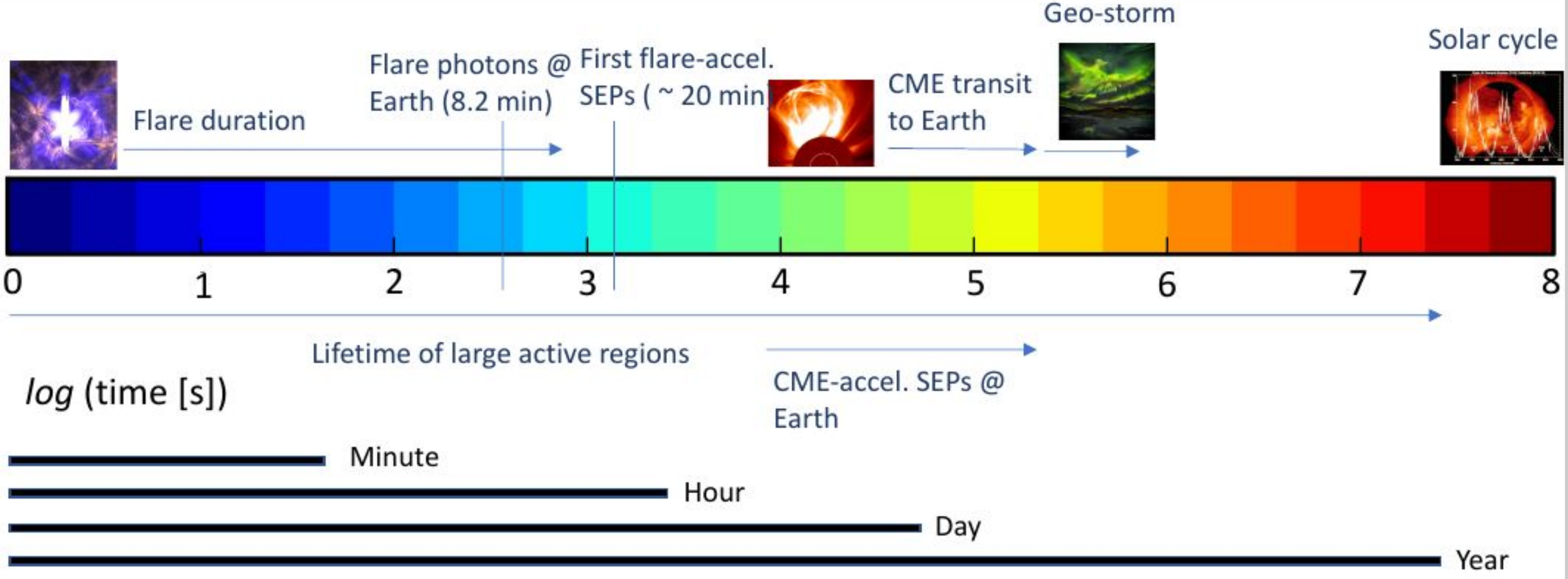
Source images: NASA, NOAA.

Stakeholders

- Satellite operators, electric utilities, airlines, oil drilling companies, precision agriculture, and federal agencies

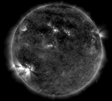


Time- and length-scales involved in SWx research



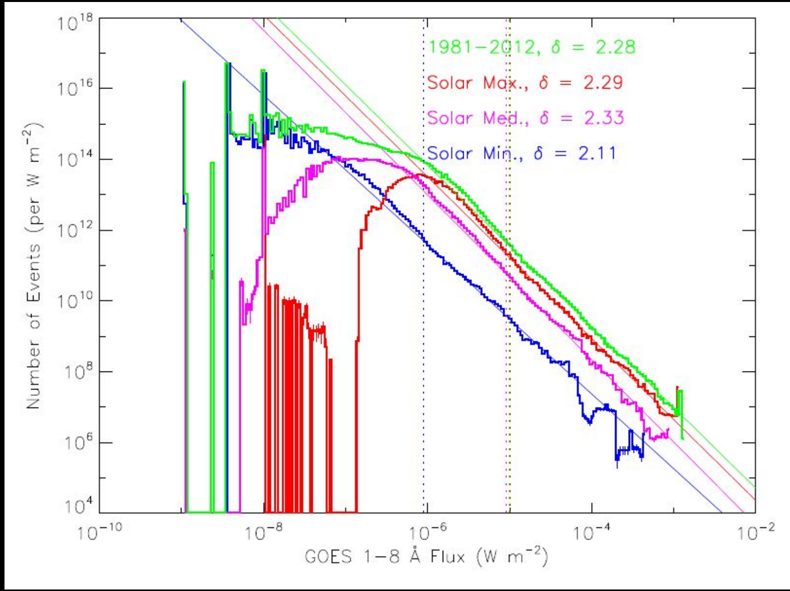
- ❑ A total span of > 8 orders of magnitude, from flare triggering to solar-cycle scales
- ❑ Different phenomena have trigger intertwined, interactive time scales

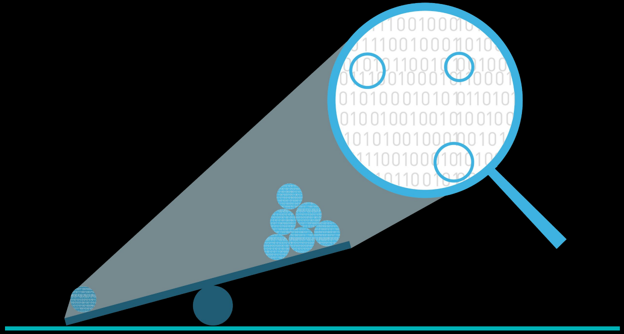
Image Credit: Georgoulis (2018)

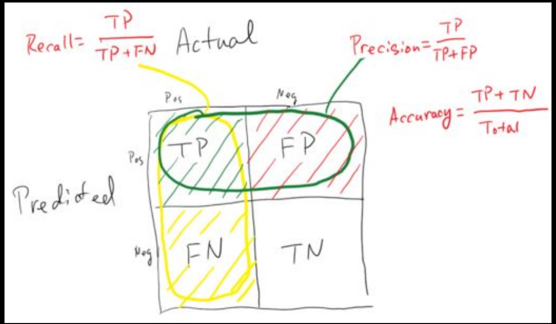


Rare-event Prediction

- Power-law distribution
- More interesting the event less instances of it is available
- Imbalanced Dataset
- Alternative evaluation metrics







Predictive Tasks for Space Weather

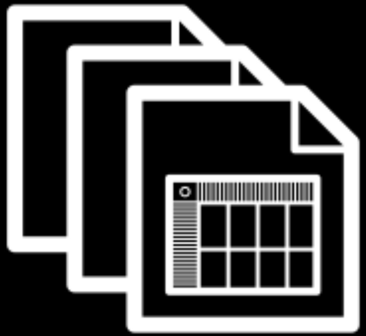
- Flare
  - Occurrence (binary/probability), magnitude (individual event or augmented over time)
  - Full-disk aggregated, active region-based
- Coronal Mass Ejections
  - Occurrence, characteristics (width, direction, velocity)
- SEP Events
  - Occurrence (binary/probability), magnitude, profile
- Solar wind
  - Magnitude, speed
- Geomagnetic storms, radio bursts, aurora, …

Benchmark Dataset Creation for Space Weather Analytics

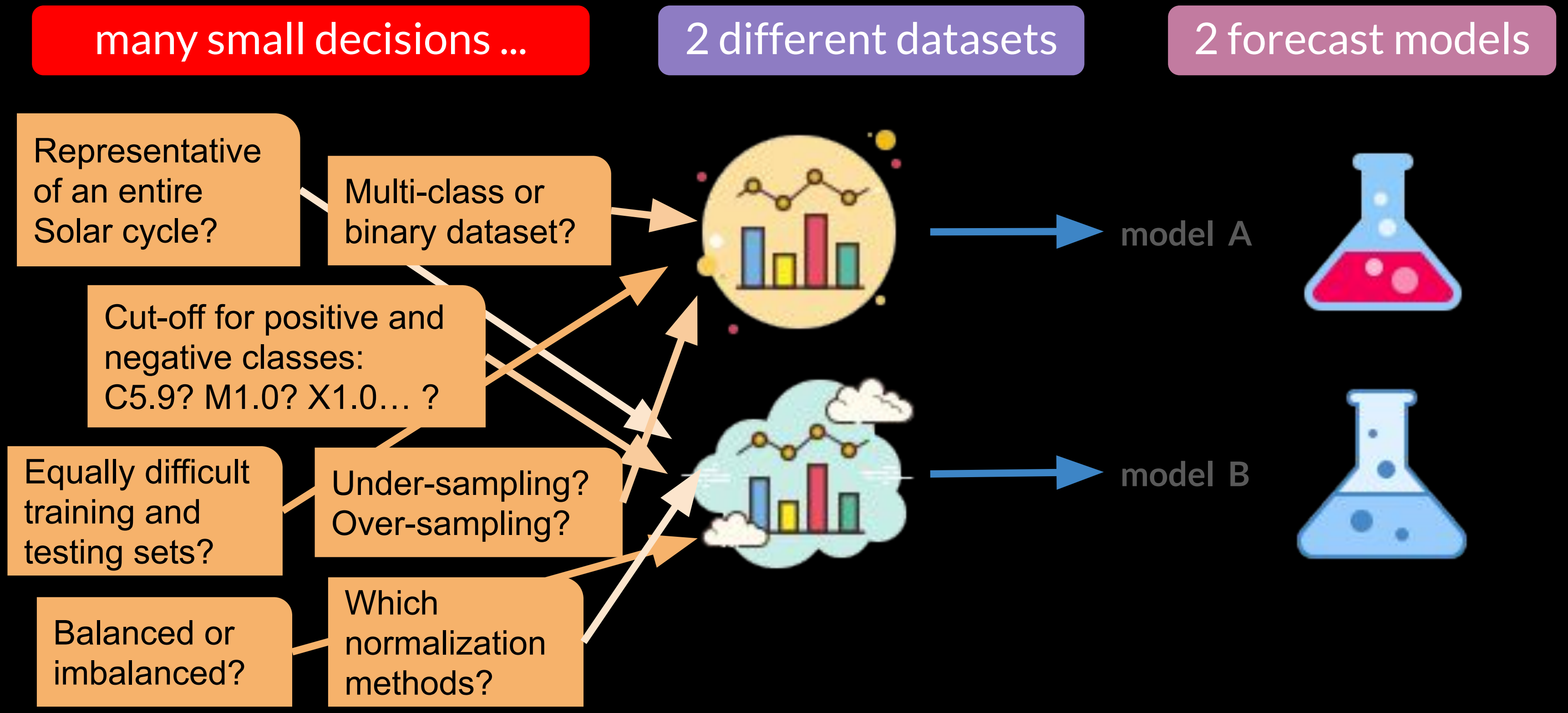
- Processing
- Transformations
- Cleaning
- Integration

Benchmark Datasets

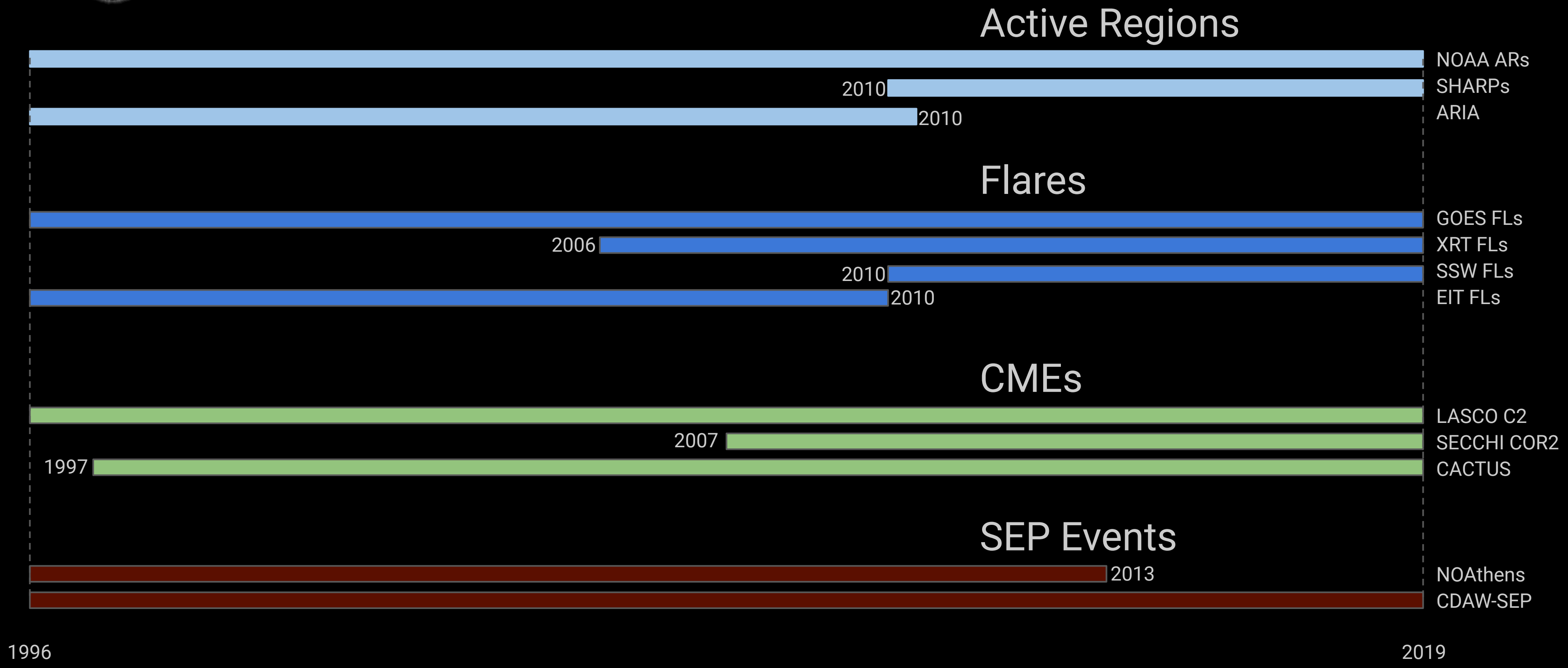
- Allows for fair comparisons among different algorithms and models
- Many small decisions impacting the outcome
- Veracity (accuracy and completeness)

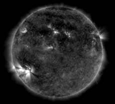


Motivation


many small decisions ...
2 different datasets
2 forecast models
Representative of an entire Solar cycle?
Multi-class or binary dataset?
Cut-off for positive and negative classes: C5.9? M1.0? X1.0… ?
Equally difficult training and testing sets?
Under-sampling? Over-sampling?
Balanced or imbalanced?
Which normalization methods?
model A
model B

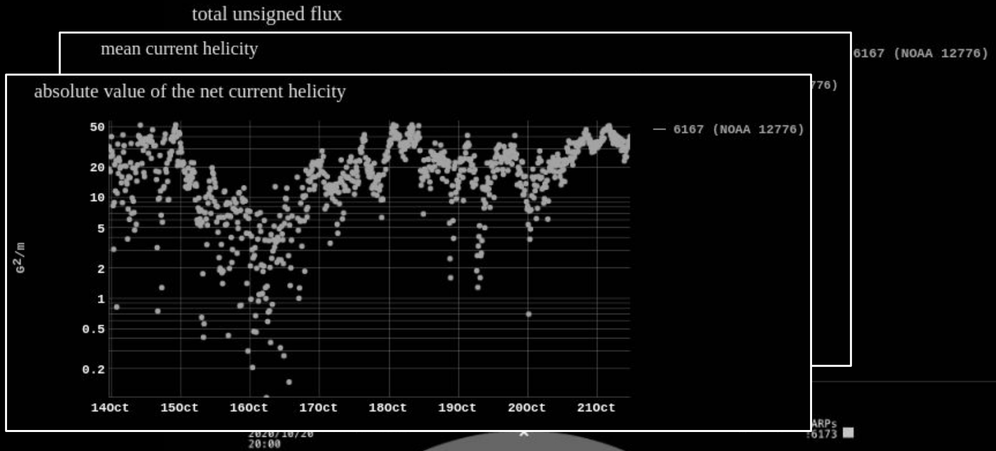
Temporal Coverage

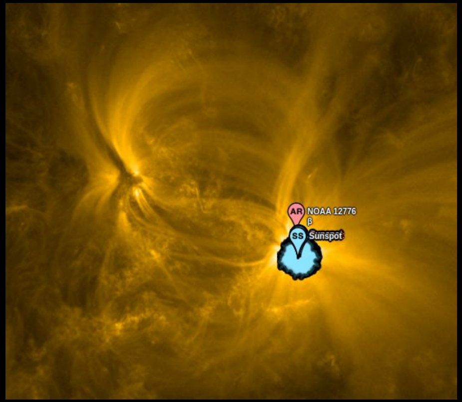


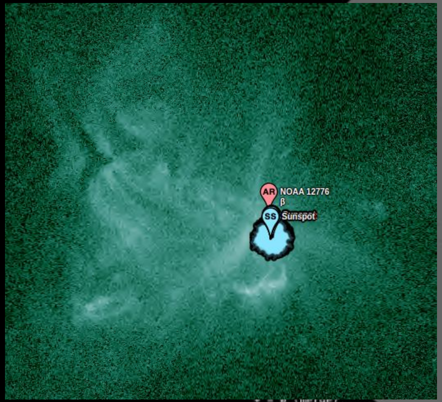


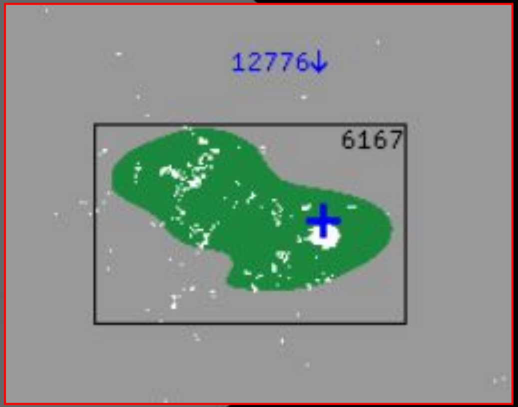
Solar Data

- High-dimensional
- Heterogeneous
- Multi-faceted

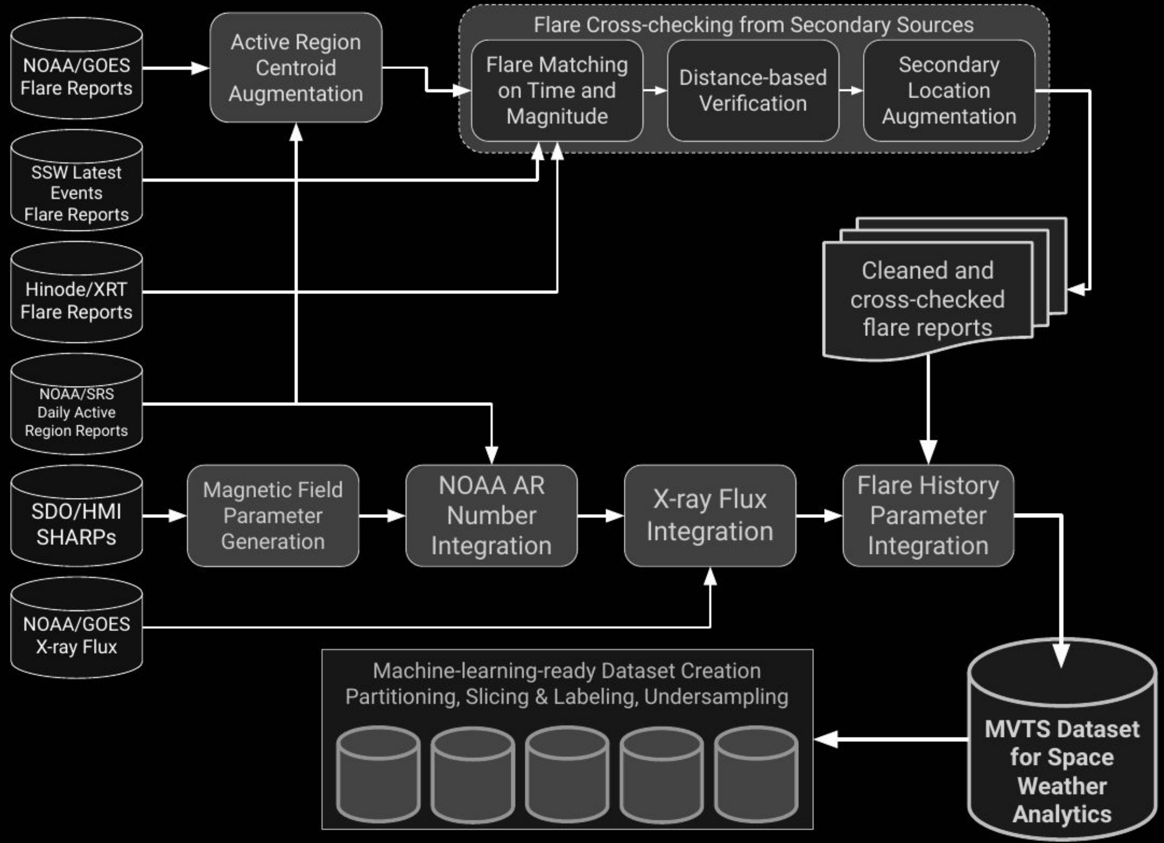




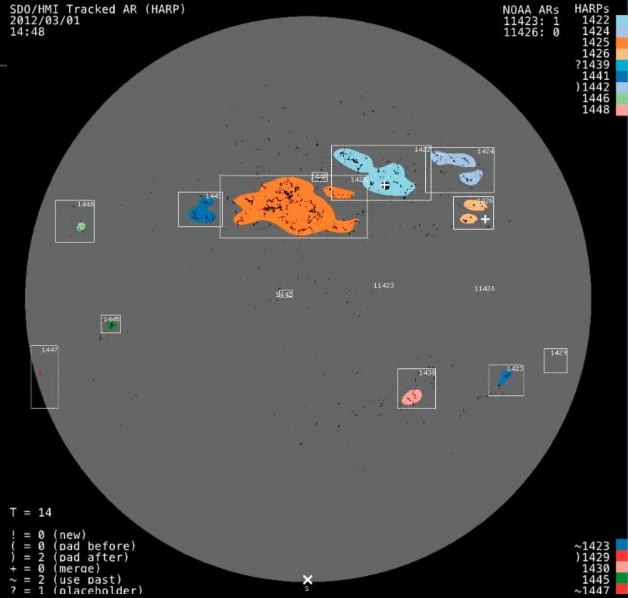


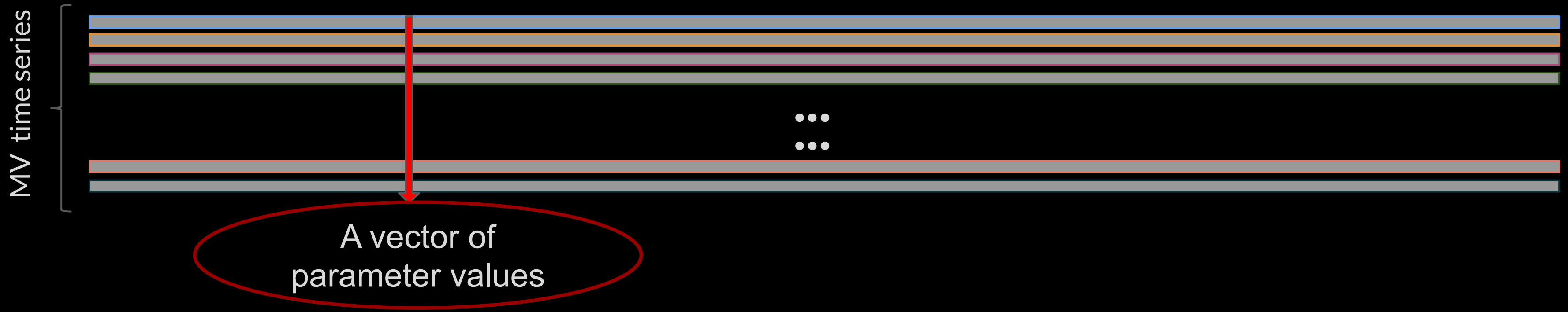


SWAN-SF Overview

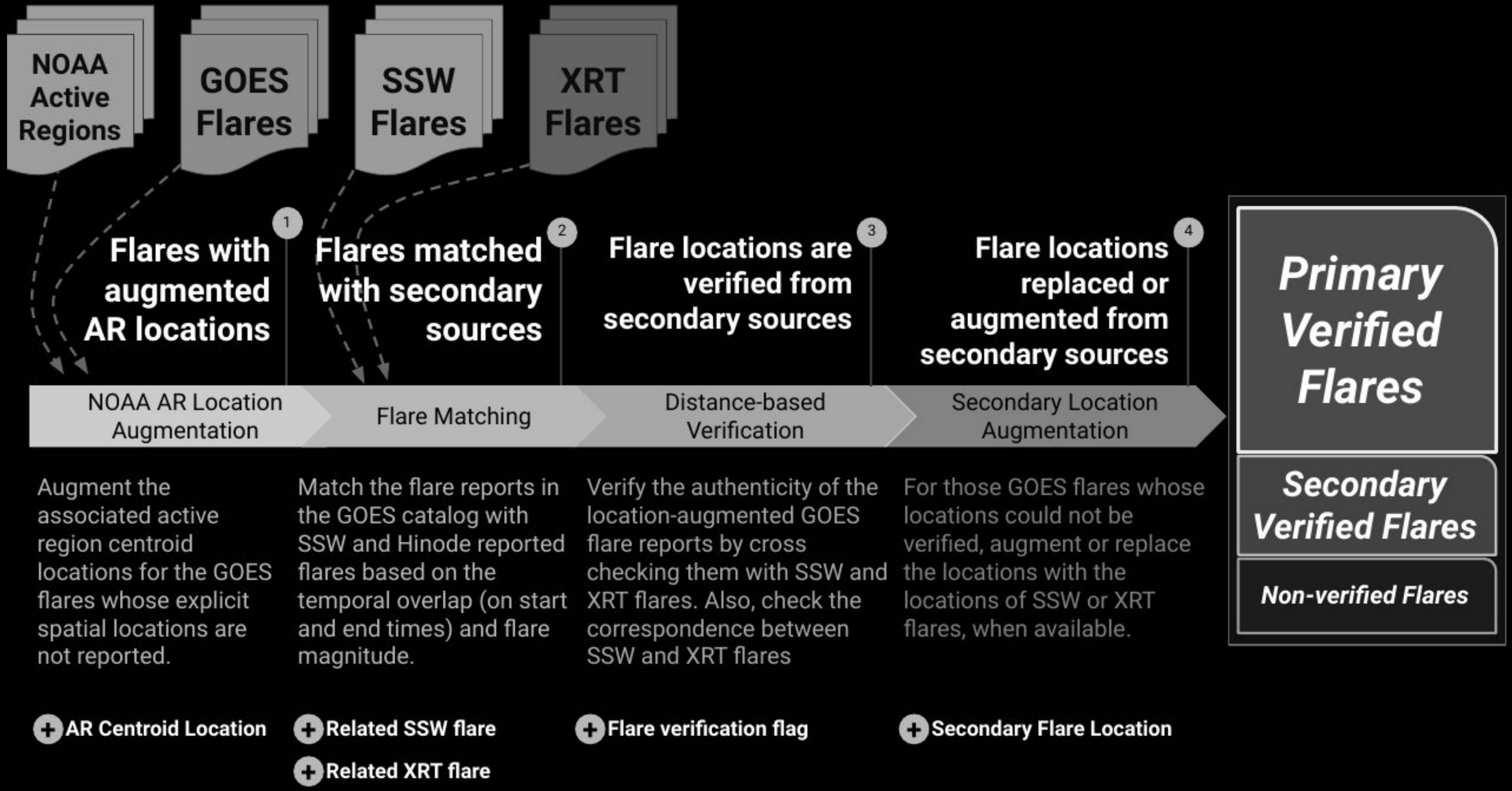


SWAN-SF – Active Region Data

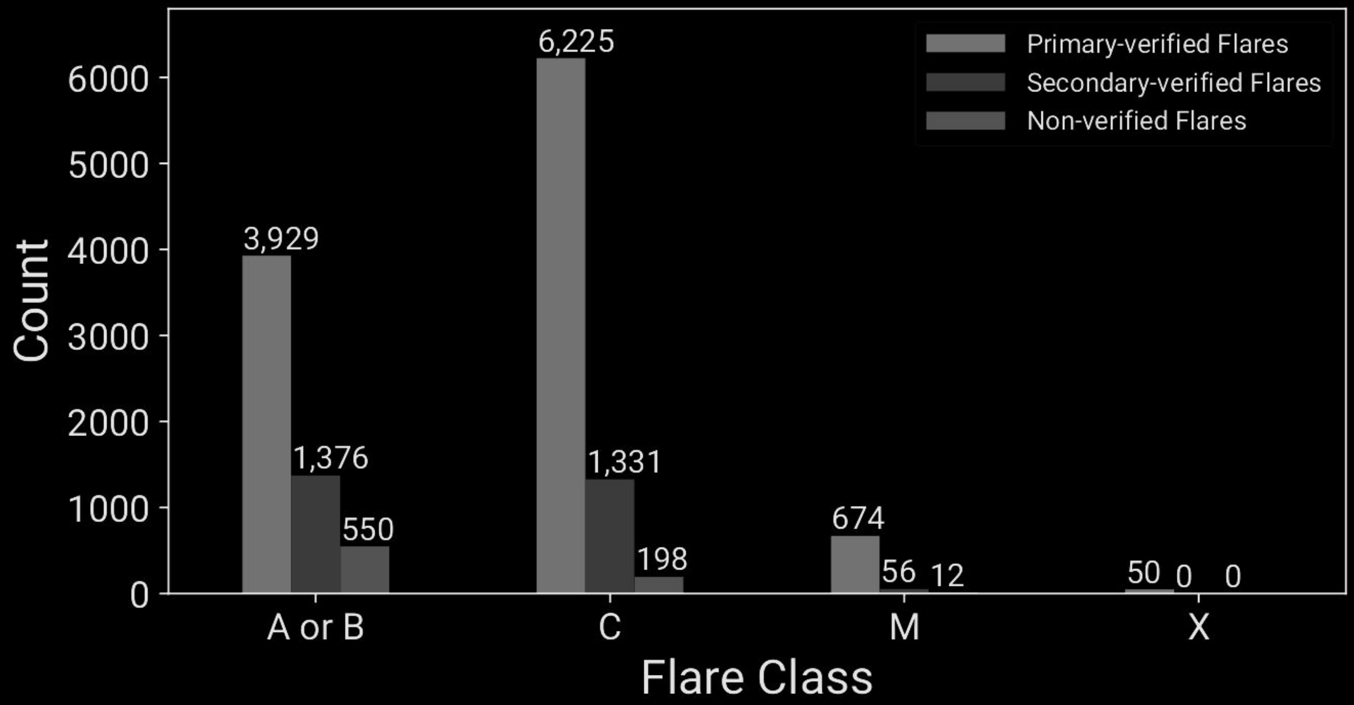




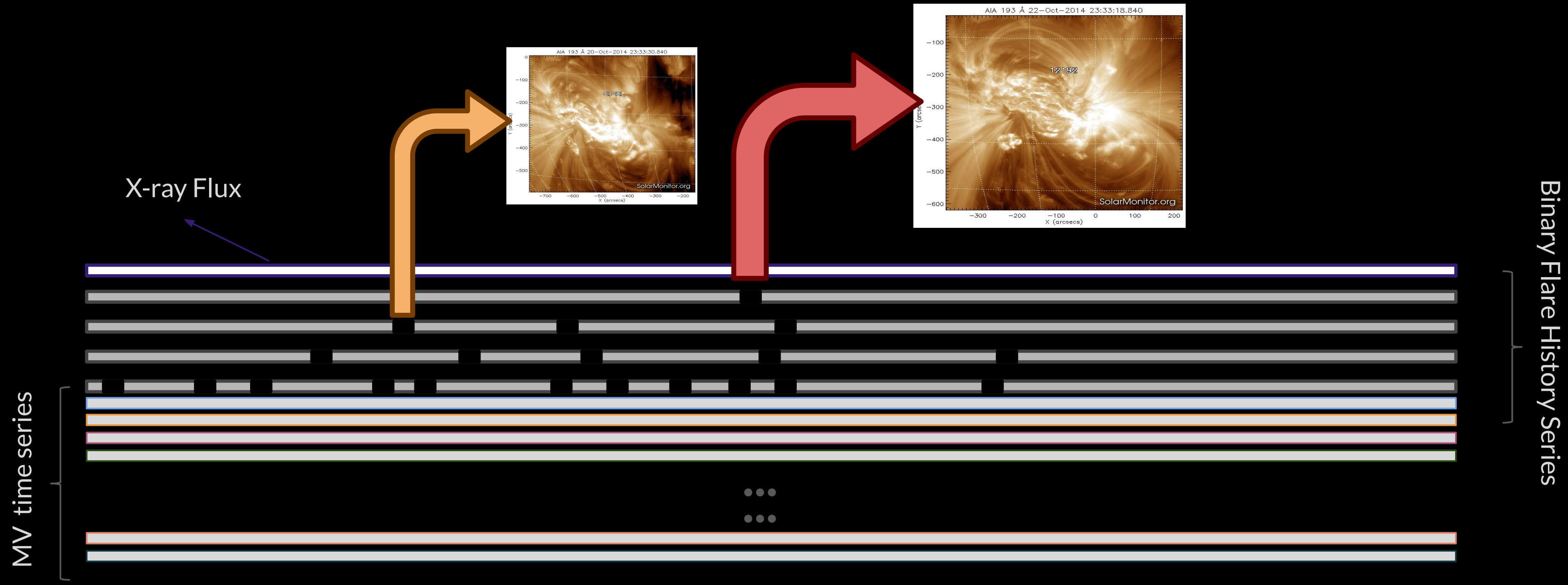
SWAN-SF – Flare Data



SWAN-SF – Flare Data

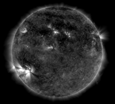


Data Integration



Harvard Dataverse - Usage

Practical Space Weather Analytics using SWAN-SF Dataset




Data Quality
Benchmark Datasets
Learning Models

Data Issues for Space Weather Analytics

Data Accuracy

Validation of data

Cross verification of data sources

Imbalance (skill scores) and Sampling

Model verification/CV/Data Partitioning

ML-Ready Dataset Creation

Hands-on Data Validation

Hands-on Imbalance and Skill Scores

Hands-on Temporal Coherence