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

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.

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

Integrated Solar Event Database

Data Wrangling for SWx Forecasting
Space Weather Forecasting * Rare-event Prediction * Role of Benchmark Datasets * Best Practices

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.

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, Earthward. These CMEs have their own magnetic field and can slam into and interact with Earth's magnetic field, resulting in geomagnetic storms. The extent 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.

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 right side by the present solar wind. During geomagnetic storms, the disturbances in Earth's magnetic field can become extreme. In addition to some buffeting by the atmosphere, this field also offers some shielding from the charged particles of a radiation storm.

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 solar storms, where 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 become energized by these events and can eventually bombard 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 atmosphere) are usually associated with geomagnetic storms.

Source images: NASA, NOAA.
Stakeholders

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

- Solar Flares
- Coronal Mass Ejections (CMEs)
- Particle Events (SEP, SPE, ESP)
- Solar Wind, Geomagnetic Storms
- Radio Burst/Blackout
- Solar Cycle

Time- and length-scales involved in SWx research

- Flare photons @ Earth (8.2 min)
- First flare-accel. SEPs (~20 min)
- CME transit to Earth
- Geo-storm
- Solar cycle

Log (time [s])

- Lifetime of large active regions
- CME-accel. SEPs @ Earth

- 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 ...

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?

2 different datasets

2 forecast models

model A

model B
Temporal Coverage

Active Regions
- NOAA ARs
- SHARPs
- ARIA

Flares
- GOES FLs
- XRT FLs
- SSW FLs
- EIT FLs

CMEs
- LASCO C2
- SECCHI COR2
- CACTUS

SEP Events
- NOAthens
- CDAW-SEP
Solar Data

- High-dimensional
- Heterogeneous
- Multi-faceted
SWAN-SF Overview

- NOAA/GOES Flare Reports
- SSW Latest Events Flare Reports
- Hinode/XRT Flare Reports
- NOAA/SSRS Daily Active Region Reports
- SDO/HMI SHARPs
- NOAA/GOES X-ray Flux

Active Region Centroid Augmentation

Flare Cross-checking from Secondary Sources
- Flare Matching on Time and Magnitude
- Distance-based Verification
- Secondary Location Augmentation

Cleaned and cross-checked flare reports

- Magnetic Field Parameter Generation
- NOAA AR Number Integration
- X-ray Flux Integration
- Flare History Parameter Integration

Machine-learning-ready Dataset Creation
- Partitioning, Slicing & Labeling, Undersampling

MVTS Dataset for Space Weather Analytics
SWAN-SF – Active Region Data

A vector of parameter values
**SWAN-SF – Flare Data**

1. **Flares with augmented AR locations**
   - Augment the associated active region centroid locations for the GOES flares whose explicit spatial locations are not reported.
   - **AR Centroid Location**
   - **Related SSW flare**
   - **Related XRT flare**

2. **Flares matched with secondary sources**
   - Match the flare reports in the GOES catalog with SSW and Hinode reported flares based on the temporal overlap (on start and end times) and flare magnitude.
   - **Flare Matching**
   - **Related SSW flare**
   - **Related XRT flare**

3. **Flare locations are verified from secondary sources**
   - Verify the authenticity of the location-augmented GOES flare reports by cross checking them with SSW and XRT flares. Also, check the correspondence between SSW and XRT flares.
   - **Flare verification flag**

4. **Flare locations replaced or augmented from secondary sources**
   - For those GOES flares whose locations could not be verified, augment or replace the locations with the locations of SSW or XRT flares, when available.
   - **Secondary Flare Location**

**Primary Verified Flares**

**Secondary Verified Flares**

**Non-verified Flares**
SWAN-SF – Flare Data

Count

Flare Class

A or B
C
M
X

Primary-verified Flares
Secondary-verified Flares
Non-verified Flares

3,929
1,376
1,331
674
56
12
50
0

550
198
0
0
Data Integration

- X-ray Flux
- Binary Flare History Series

MV time series
Harvard Dataverse - Usage
Practical Space Weather Analytics using SWAN-SF Dataset
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