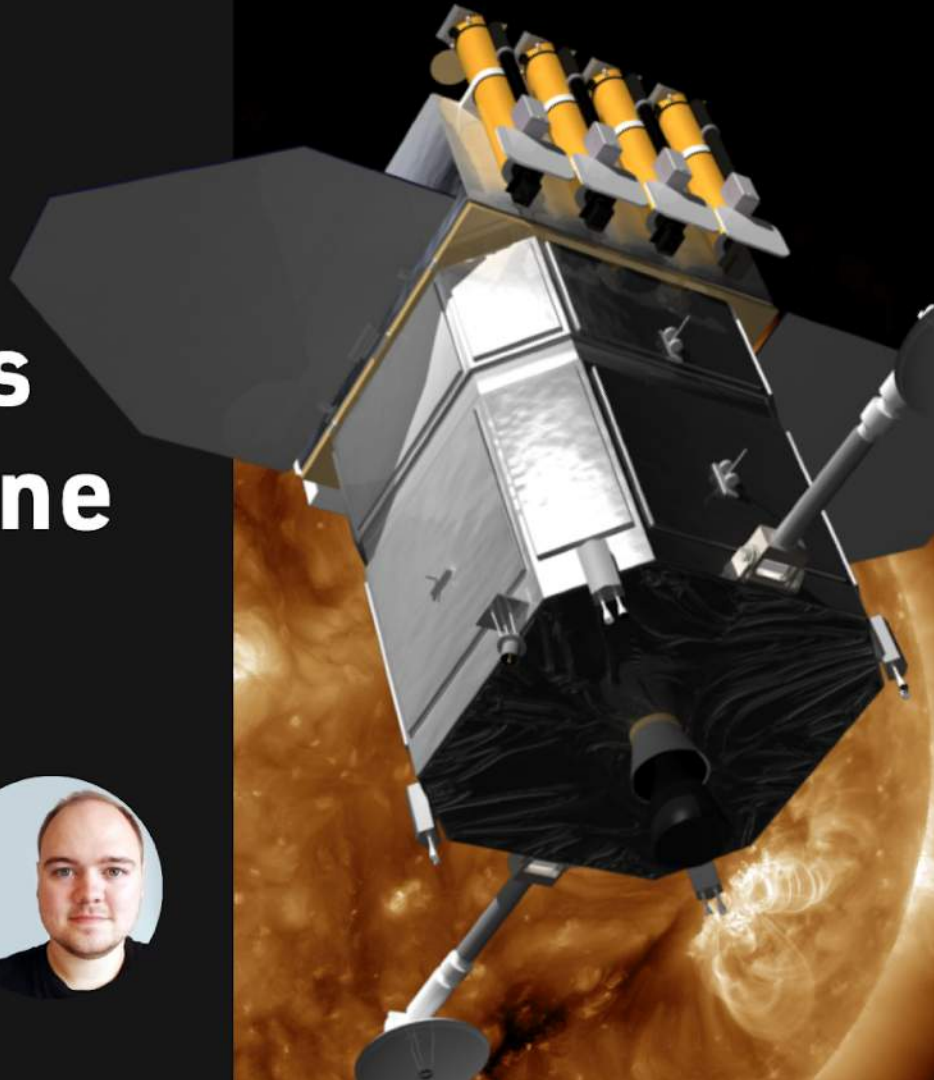


SPACEML 
Speaker Series

The Solar Dynamics Observatory Machine Learning Dataset

Paul J. Wright, Ph.D.
SpaceML Curator
& Researcher



Outline

1. Summary
2. The Solar Dynamics Observatory
3. SDOML Dataset

4. SpaceML
5. SDOML Notebook Demo
6. SDOML Projects

1. Summary

Solar Dynamics Observatory

- SDO has **three instruments**
Each observe the Sun over varying wavelengths and at varying cadence
- SDO has obtained > half-a-billion images
Petabyte-scale scientific dataset
- SDO data are easily accessible
However, pre-processing these data for a scientific analysis often requires specialised Heliophysics knowledge.

SDOML on SpaceML

- SDOML is a **cleaned, curated dataset** covering the entirety of the SDO mission 2010 - present
- SpaceML brings together:
 - data storage (Google Cloud),
 - compute (Google Colab/Cloud Platform),
 - well-commented, version-controlled notebooks for data access and reproducibility

2. The Solar Dynamics Observatory

Solar Dynamics Observatory


- SDO has **three instruments**
Each observe the Sun over varying wavelengths and at varying cadence
- Helioseismic and Magnetic Imager (SDO/HMI)
 - Photosphere
(equivalent to as-seen-by-eye)
 - 4096 x 4096 pixels
 - every 12 minutes

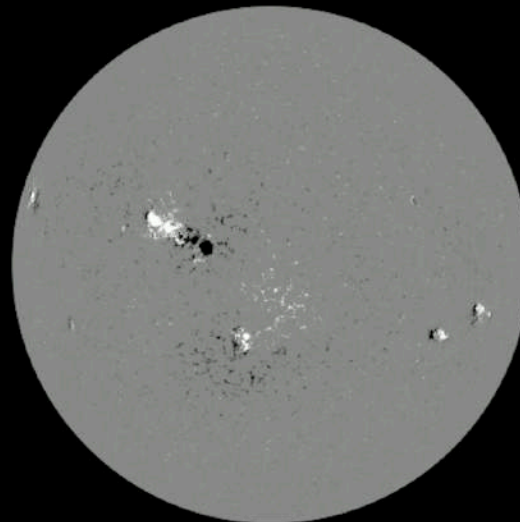


SDO/HMI
Continuum

Solar Dynamics Observatory

- SDO has **three instruments**
Each observe the Sun over varying wavelengths and at varying cadence
- Helioseismic and Magnetic Imager (SDO/HMI)
 - Vector magnetic field
 - Converted into line-of-sight (e.g. right)
 - 4096 x 4096 pixels
 - every 12 minutes

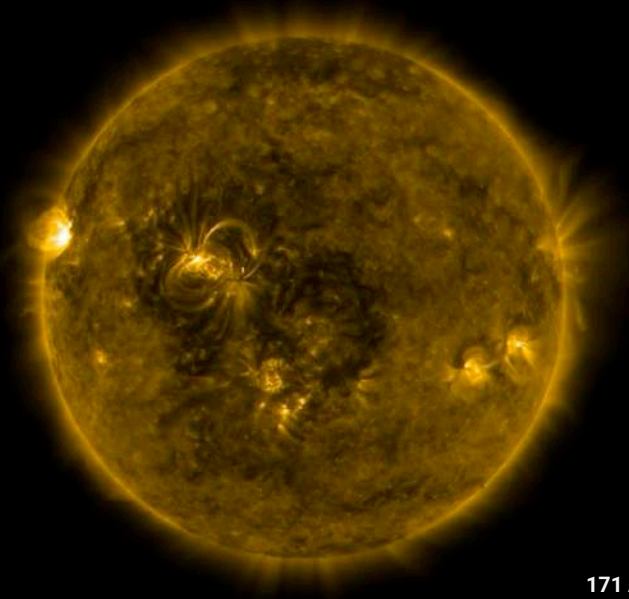
www.helioviewer.org 



SDO/HMI
Line-of-sight Magnetogram

Solar Dynamics Observatory

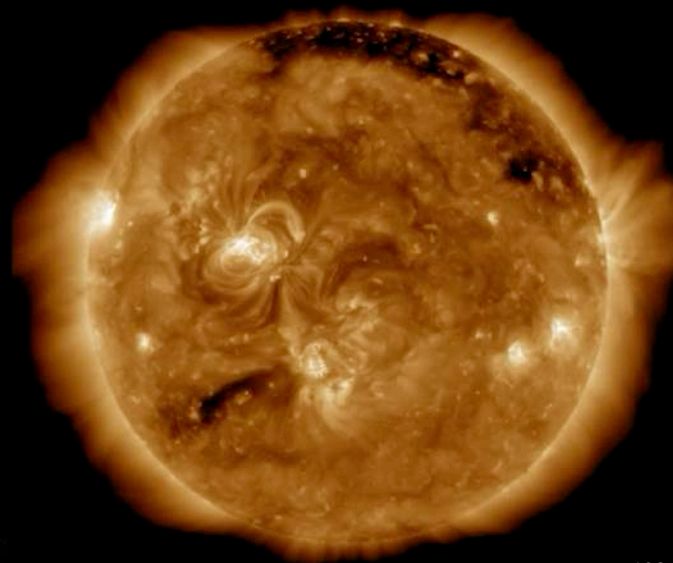
- SDO has **three instruments**
Each observe the Sun over varying wavelengths and at varying cadence
- Helioseismic and Magnetic Imager (SDO/HMI)
- Atmospheric Imaging Assembly (SDO/AIA)
 - 10 channels with varying temperature responses
 - Extreme Ultraviolet data (e.g. right) is obtained at 4096 x 4096 pixels every 12 seconds



SDO/AIA
171 Ångström
 $T_{\text{peak}} \sim 0.6 \text{ MK}$

Solar Dynamics Observatory

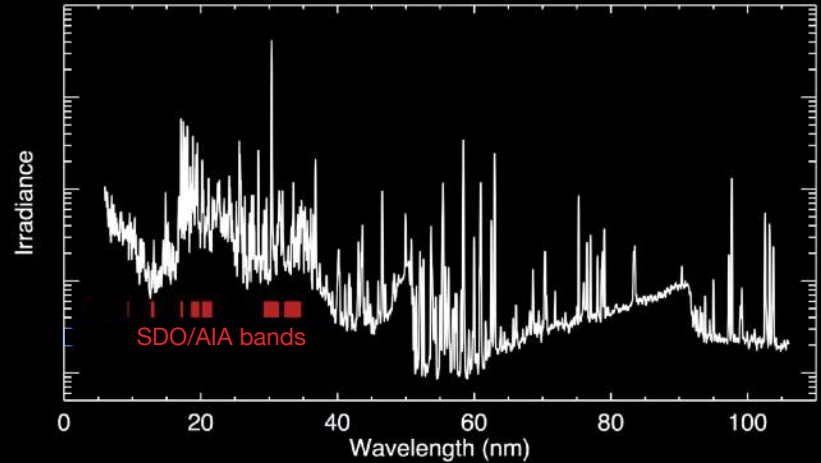
- SDO has **three instruments**
Each observe the Sun over varying wavelengths and at varying cadence
- Helioseismic and Magnetic Imager (SDO/HMI)
- Atmospheric Imaging Assembly (SDO/AIA)
 - Exposure time is variable
 - Telescope jitter is present
 - Eclipses occur
 - Instruments are sometimes offline



SDO/AIA
193 Ångström
 $T_{\text{peak}} \sim 1.5 \text{ MK}$

Solar Dynamics Observatory

- SDO has **three instruments**
Each observe the Sun over varying wavelengths and at varying cadence
- Helioseismic and Magnetic Imager (SDO/HMI)
- Atmospheric Imaging Assembly (SDO/AIA)
- Extreme Variability Experiment (SDO/EVE)
Provides sun-as-a-star spectra every 10 seconds, with data that overlaps SDO/AIA bands.

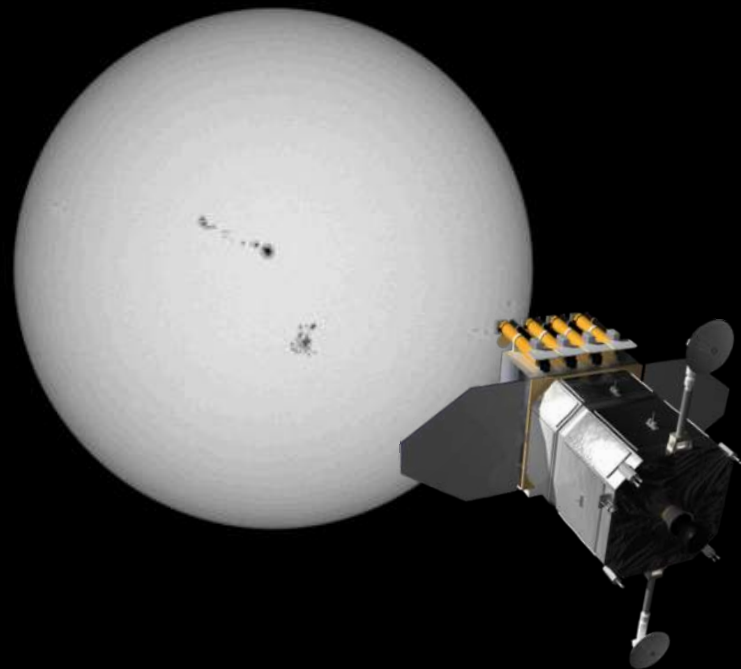


The solar extreme ultraviolet (EUV) spectrum measured by SDO/EVE, showing an overlap with SDO/AIA. EVE measures this spectrum every 10 seconds. Credit: NASA/LASP/CU

Solar Dynamics Observatory

- Since 2010, SDO has obtained Petabytes of high-quality/cadence scientific data
- With many unanswered questions in Heliophysics, there is an incredible potential for statistical/machine learning approaches.

SDOML combined with SpaceML can enable such discoveries, uncovering the mysteries of our closest star



3. SDOML Dataset



A project card for the SDOML dataset. The top section features a solar image with a 'PROJECT' label and a timestamp 'SDO/S4 103 2021-12-07 10:42:17 UT'. Below the image, the text reads: 'SDOML', 'A Machine Learning Dataset Prepared From the NASA Solar Dyna...', 'CHALLENGE AREA SPACE WEATHER', 'PROGRAM FDL US'.

PROJECT

SDO/S4 103 2021-12-07 10:42:17 UT

SDOML
A Machine Learning Dataset
Prepared From the NASA Solar
Dyna...

CHALLENGE AREA
SPACE WEATHER

PROGRAM
FDL US

SDOML Dataset

Galvez et al. 2019 ApJS

Data processing includes:

- Rotation to Solar North
- Co-alignment of images taken at the same time
- Correction for the Earth's elliptical orbit around the Sun
- Correction for AIA degradation
EUV filters degrade in orbit.

Resulting Cadence:

- AIA (12s -> 6 min.)
- HMI (12 min.)
- EVE (10s -> 1 min.)

THE ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES, 242:7 (11pp), 2019 May

<https://doi.org/10.3847/1538-4365/ab1005>

© 2019, The American Astronomical Society.

OPEN ACCESS



A Machine-learning Data Set Prepared from the NASA *Solar Dynamics Observatory* Mission

Richard Galvez¹, David F. Fouhey², Meng Jin^{3,4}, Alexandre Szentner⁵, Andrés Muñoz-Jaramillo⁶,
Mark C. M. Cheung^{1,7}, Paul J. Wright⁸, Monica G. Bobra⁹, Yang Liu⁹, James Mason⁹, and Rajat Thomas¹⁰

¹Center for Data Science, New York University, New York, NY 10003, USA, richard.galvez@nyu.edu

²University of Michigan, Ann Arbor, MI 48109, USA

³Lockheed Martin Solar & Astrophysics Laboratory, Palo Alto, CA, USA

⁴SETI Institute, Mountain View, CA 94043, USA

⁵University of Oxford, Oxford OX1 2JD, UK

⁶Southwest Research Institute, San Antonio, TX 78238, USA

⁷Hansen Experimental Physics Laboratory, Stanford University, Stanford, CA 94305, USA

⁸SUPA School of Physics & Astronomy, University of Glasgow, Glasgow G12 8QQ, UK

⁹NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA

¹⁰University of Amsterdam, 1012 WX Amsterdam, Netherlands

Received 2019 January 19; revised 2019 March 11; accepted 2019 March 13; published 2019 May 8

Abstract

In this paper, we present a curated data set from the NASA *Solar Dynamics Observatory* (*SDO*) mission in a format suitable for machine-learning research. Beginning from level 1 scientific products we have processed various instrumental corrections, down-sampled to manageable spatial and temporal resolutions, and synchronized observations spatially and temporally. We illustrate the use of this data set with two example applications: forecasting future extreme ultraviolet (EUV) Variability Experiment (EVE) irradiance from present EVE irradiance and translating Helioseismic and Magnetic Imager observations into Atmospheric Imaging Assembly observations. For each application, we provide metrics and baselines for future model comparison. We anticipate this curated data set will facilitate machine-learning research in heliophysics and the physical sciences generally, increasing the scientific return of the *SDO* mission. This work is a direct result of the 2018 NASA Frontier Development Laboratory Program. Please see the Appendix for access to the data set, totaling 6.5TBs.

Key words: astronomical databases: miscellaneous – catalogs – editorials, notices – miscellaneous – surveys

Version 1.0

Galvez *et al.* 2019 ApJS

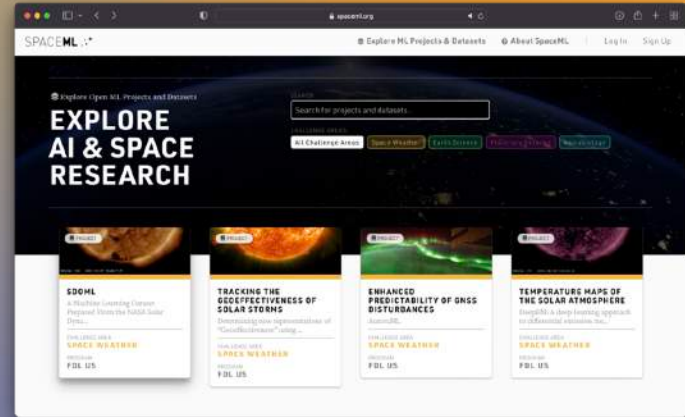
- 2010 - 2018
- Size: 7 TB
- Stored as .npz (image arrays) on the Stanford Digital Repository
- Problems with v1.0
 - Data is stored away from appropriate computing resources (Download is required)
 - The size of the data can be prohibitive.
 - No demonstration notebooks available.
 - No metadata associated with the images

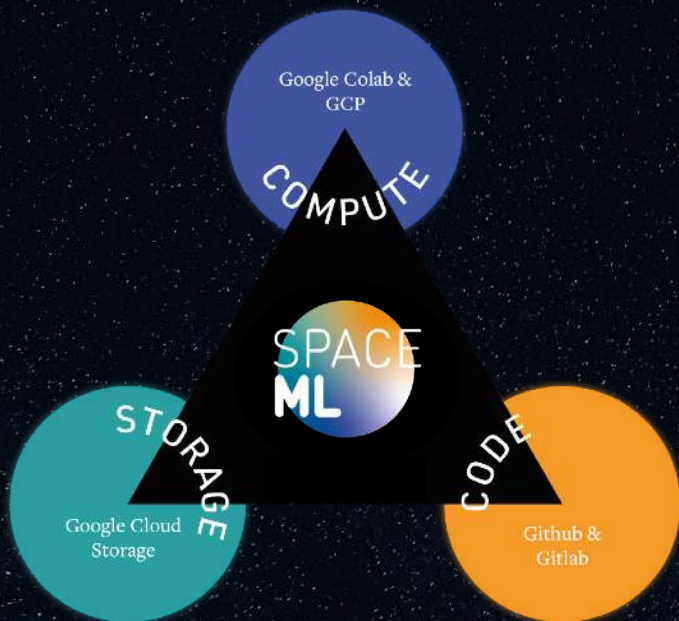
Version 2.0

SDO ML on SpaceML

- 2010 - **present**
- Stored as .zarr (image & metadata) on **Google Cloud/SpaceML**
 - Small dataset freely accessible with demonstration notebooks
 - Complete data access available on request
- **Updated degradation curves**
 - Based on the latest correction tables
- Provisions in place to **continually update (and version) the data**

4. SpaceML.org





SpaceML

is a machine learning toolbox and developer community building open science AI applications for space science and exploration.

- SpaceML hosts a number of freely available datasets for scientific analysis, including SDOML
- Project pages demonstrate published results with reproducible notebooks (that access data hosted by SpaceML)

Learn more about SpaceML @ spaceml.org

The screenshot shows the SpaceML website interface. At the top, there's a navigation bar with the SpaceML logo, a search icon, and links for 'Explore ML Projects & Datasets', 'About SpaceML', 'Log In', and 'Sign Up'. Below the navigation bar, the main content area features a large heading 'EXPLORE AI & SPACE RESEARCH' and a search bar with the placeholder text 'Search for projects and datasets...'. Underneath the search bar, there are several 'CHALLENGE AREAS' buttons: 'All Challenge Areas', 'Space Weather', 'Earth Science', 'Planetary Defense', and 'Astrobiology'. The main content area displays four project cards, each with a 'PROJECT' icon and a title. The first card is titled 'SDO ML' and describes a machine learning dataset prepared from NASA Solar Dynamics Observatory data. The second card is titled 'TRACKING THE GEOEFFECTIVENESS OF SOLAR STORMS' and describes determining new representations of 'Geoeffectiveness' using... The third card is titled 'ENHANCED PREDICTABILITY OF GNSS DISTURBANCES' and describes AuroraML. The fourth card is titled 'TEMPERATURE MAPS OF THE SOLAR ATMOSPHERE' and describes DeepEM: A deep-learning approach to differential emission me... Each card also includes a 'CHALLENGE AREA' (SPACE WEATHER) and a 'PROGRAM' (FDL US).

SPACEML

Explore ML Projects & Datasets | About SpaceML | Log In | Sign Up

Explore Open ML Projects and Datasets

SEARCH

Search for projects and datasets...

CHALLENGE AREAS

All Challenge Areas | Space Weather | Earth Science | Planetary Defense | Astrobiology

PROJECT

SDO ML
A Machine Learning Dataset Prepared From the NASA Solar Dyna...

CHALLENGE AREA
SPACE WEATHER

PROGRAM
FDL US

PROJECT

TRACKING THE GEOEFFECTIVENESS OF SOLAR STORMS
Determining new representations of "Geoeffectiveness" using ...

CHALLENGE AREA
SPACE WEATHER

PROGRAM
FDL US

PROJECT

ENHANCED PREDICTABILITY OF GNSS DISTURBANCES
AuroraML

CHALLENGE AREA
SPACE WEATHER


PROGRAM
FDL US

PROJECT

TEMPERATURE MAPS OF THE SOLAR ATMOSPHERE
DeepEM: A deep-learning approach to differential emission me...

CHALLENGE AREA
SPACE WEATHER

PROGRAM
FDL US



SPACEML 

Explore ML Projects & Datasets | About SpaceML | Log In | Sign Up

REPO → PROJECT

SDOML

A Machine Learning Dataset Prepared From the NASA Solar Dynamics Observatory Mission

SHARE PROJECT  

TEAM

Researchers:
Richard Galvez, Alexandre Szenicer, Paul J. Wright, Rajat Thomas

Faculty:
Mark Cheung, Andrés Muñoz-Jaramillo, David Fouhey, Meng Jin

Posted by:
[Leosilverberg](#)



DETAILS

Program:
FDL US

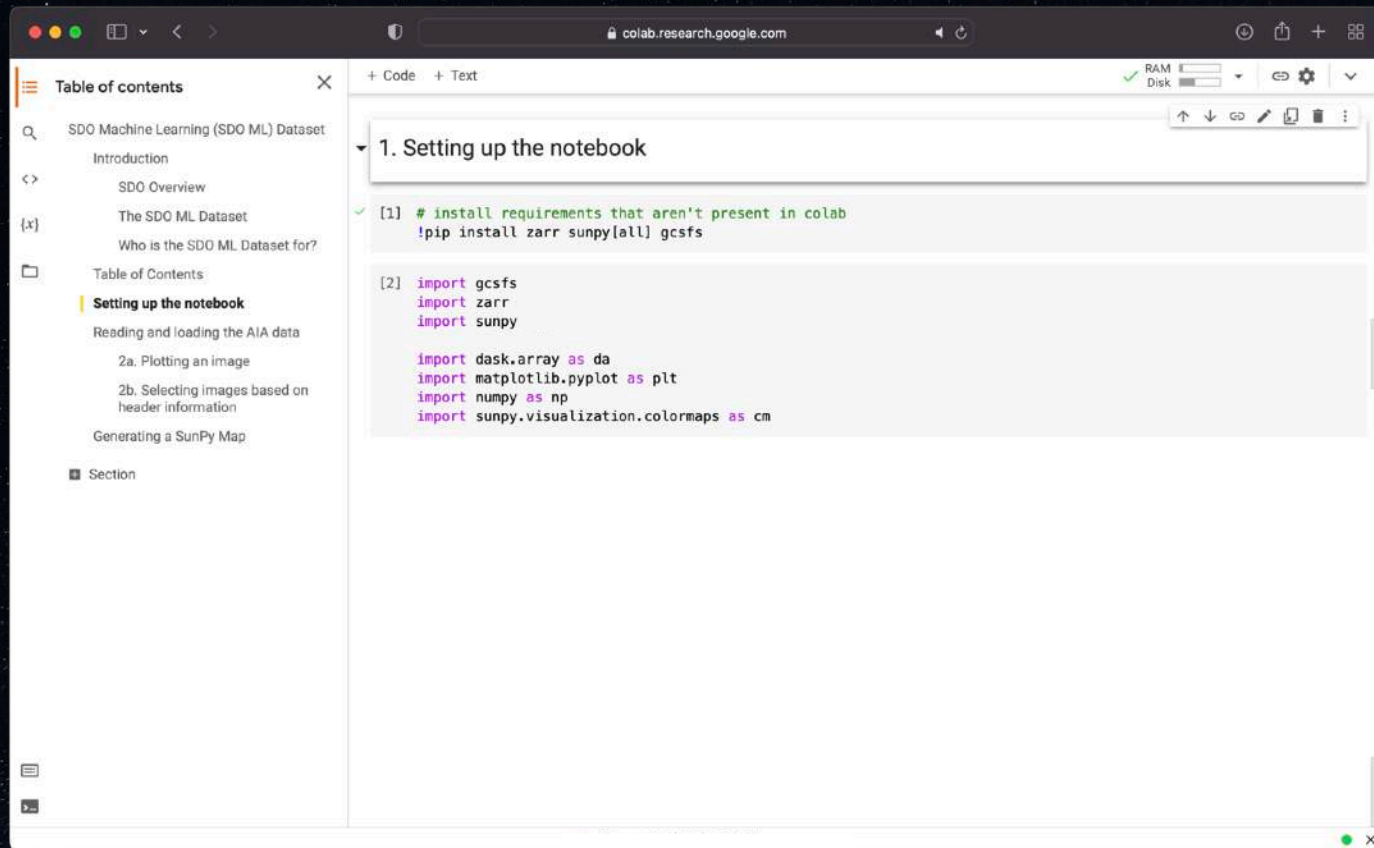
Institution:
New York University, University Of Oxford, University Of Glasgow, University Of Amsterdam, Lockheed Martin Solar & Astrophysics Lab, UC Berkeley, SETI Institute, Frontier Development Lab

Year:
2018

Challenge area:
Space Weather

OVERVIEW DATASETS  OPEN IN COLAB  OPEN IN GITHUB / GITLAB

5. SDOML Notebook Demo



The screenshot shows a Google Colab notebook window. The browser address bar displays `colab.research.google.com`. The interface includes a left sidebar with a "Table of contents" panel, a top toolbar with "Code" and "Text" tabs, and a main workspace area. The "Table of contents" panel lists the following items:

- SDO Machine Learning (SDO ML) Dataset
 - Introduction
 - SDO Overview
 - The SDO ML Dataset
 - Who is the SDO ML Dataset for?
- Table of Contents
 - Setting up the notebook**
 - Reading and loading the AIA data
 - 2a. Plotting an image
 - 2b. Selecting images based on header information
 - Generating a SunPy Map
 - Section

The main workspace area shows a code cell titled "1. Setting up the notebook" with the following Python code:

```
[1] # install requirements that aren't present in colab
!pip install zarr sunpy[all] gcsfs

[2] import gcsfs
import zarr
import sunpy

import dask.array as da
import matplotlib.pyplot as plt
import numpy as np
import sunpy.visualization.colormaps as cm
```

Table of contents

- SDO Machine Learning (SDO ML) Dataset
 - Introduction
 - SDO Overview
 - The SDO ML Dataset
 - Who is the SDO ML Dataset for?
 - Table of Contents
 - Setting up the notebook
 - Reading and loading the AIA data**
 - 2a. Plotting an image
 - 2b. Selecting images based on header information
 - Generating a SunPy Map
 - Section

2. Reading and loading the AIA data

```
[3] gcs = gcsfs.GCSFileSystem(access="read_only")
    loc = "fdl-sdoml-v2/sdomlv2_small.zarr/"
    store = gcsfs.GCSMap(loc, gcs=gcs, check=False)
```

The SDO ML dataset is stored in the Zarr format, a format for the storage of chunked, compressed, N-dimensional arrays with Numpy dtype. For an in-depth overview, see <https://zarr.readthedocs.io/en/stable/tutorial.html>.

```
[4] # first, we create a group with the store data located on GCP.
    root = zarr.group(store)
```

```
[5] # Using `root.tree()`, we are able to display the hierarchy (of `loc`).
    print(root.tree())
```

```

/
├── 2010
│   ├── 131A (6135, 512, 512) float32
│   ├── 1600A (6135, 512, 512) float32
│   ├── 1700A (6135, 512, 512) float32
│   ├── 171A (6135, 512, 512) float32
│   ├── 193A (6135, 512, 512) float32
│   ├── 211A (6135, 512, 512) float32
│   ├── 304A (6135, 512, 512) float32
│   ├── 335A (6135, 512, 512) float32
│   └── 94A (6135, 512, 512) float32

```

As shown in the tree, the hierarchy consists of groups, each shown with their respective shape, and data type. In this example, we will primarily look at the 171 Å channel from August 2010. This consists of 6135 512x512 images, stored as float32, and can be accessed as follows

```
data = root["2010"]["171A"]
```

colab.research.google.com

RAM
Disk

Table of contents

- SDO Machine Learning (SDO ML) Dataset
 - Introduction
 - SDO Overview
 - The SDO ML Dataset
 - Who is the SDO ML Dataset for?
 - Table of Contents
 - Setting up the notebook
 - Reading and loading the AIA data**
 - 2a. Plotting an image
 - 2b. Selecting images based on header information
 - Generating a SunPy Map
 - Section

+ Code + Text Unsaved changes since 19:52

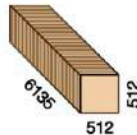
Loading with Dask

We can then load this data into an array using dask.

```
[6] data = root["2010"]["171A"]  
    all_images = da.from_array(data)
```

```
[7] all_images
```

	Array	Chunk
Bytes	6.43 GB	125.83 MB
Shape	(6135, 512, 512)	(120, 512, 512)
Count	53 Tasks	52 Chunks
Type	float32	numpy.ndarray



As shown above, the data has the shape (6135, 512, 512), and is split into 52 chunks of (120, 512, 512), each of 125.83 MB; this is further visualised on the right. The data is now in a form to be manipulated like a Numpy array.

colab.research.google.com

RAM
Disk

Table of contents

- SDO Machine Learning (SDO ML) Dataset
 - Introduction
 - SDO Overview
 - The SDO ML Dataset
 - Who is the SDO ML Dataset for?
 - Table of Contents
 - Setting up the notebook
 - Reading and loading the AIA data
 - 2a. Plotting an image**
 - 2b. Selecting images based on header information
 - Generating a SunPy Map
 - Section

2a. Plotting an image

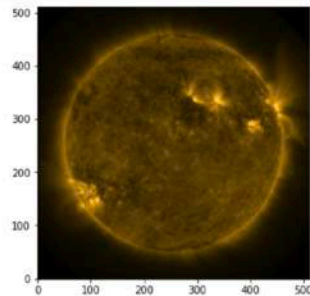
```
[8] # select the image with index 6000
image=all_images[6000,:,:]

# set the figure size
plt.figure(figsize=(5,5))

# set the colourmap from SunPy
colormap = plt.get_cmap('sdoaia171')

# As the dynamic range of these images are large,
# we can plot the square root of the data:
image = np.sqrt(image)

# Display the data.
# To display as would be seen on Helioviewer, use 'origin='lower''
plt.imshow(image, origin='lower', cmap=colormap);
```



colab.research.google.com

RAM
Disk

Table of contents

- SDO Machine Learning (SDO ML) Dataset
 - Introduction
 - SDO Overview
 - The SDO ML Dataset
 - Who is the SDO ML Dataset for?
 - Table of Contents
 - Setting up the notebook
 - Reading and loading the AIA data
 - 2a. Plotting an image
 - 2b. Selecting images based on header information**
 - Generating a SunPy Map
 - Section

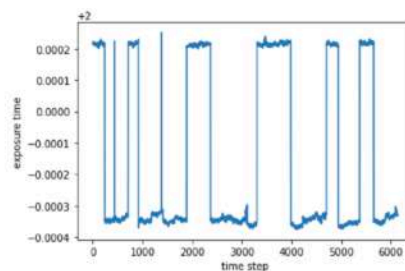
2b. Selecting images based on header information

SDO ML v2.0 includes all fits header information with the same keywords. To find out the AIA keyword definition, one can refer to the following online document: http://jsoc.stanford.edu/~jsoc/keywords/AIA/AIA02840_K_AIA-SDO_FITS_Keyword_Document.pdf

We can extract the exposure (and observation) time from the data attributes (the header information), and downsample our data based upon that information.

```
[9] # generate numpy arrays with "EXPTIME"
     exptime = np.array(data.attrs["EXPTIME"])

[10] plt.plot(exptime)
      plt.xlabel('time step')
      plt.ylabel('exposure time');
```



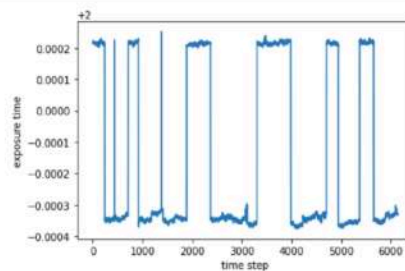
colab.research.google.com

RAM Disk

Table of contents

- SDO Machine Learning (SDO ML) Dataset
 - Introduction
 - SDO Overview
 - The SDO ML Dataset
 - Who is the SDO ML Dataset for?
 - Table of Contents
 - Setting up the notebook
 - Reading and loading the AIA data
 - 2a. Plotting an image
 - 2b. Selecting images based on header information**
 - Generating a SunPy Map
- Section

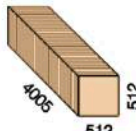
```
[10] plt.plot(exptime)
plt.xlabel('time step')
plt.ylabel('exposure time');
```



```
[11] # select indices where the exposure time is less than 2 seconds
index = np.where(exptime < 2.0)
selected_images = all_images[index[0], :, :]
```

```
[12] selected_images
```

Array	Chunk
Bytes 4.20 GB	125.83 MB
Shape (4005, 512, 512)	(120, 512, 512)
Count 92 Tasks	39 Chunks
Type float32	numpy.ndarray



SDO ML DEMO

colab.research.google.com

RAM Disk

3. Generating a SunPy Map

SunPy is an open-source Python library for Solar Physics data analysis and visualization.

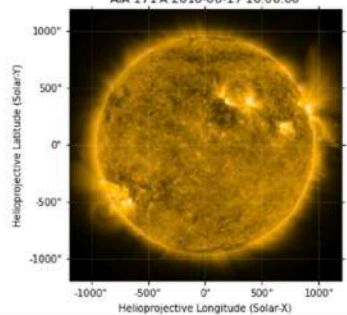
In this section we will demonstrate how SunPy's `Map` (`sunpy.map.Map`) with the Zarr-formatted data. We demonstrate this for a single index.

```
[13] # select the respective image, and header required for sunpy.map.Map()
selected_image = all_images[6000, :, :]
selected_hdr = {keys: values[6000] for keys, values in data.attrs.items()}

[14] my_map = sunpy.map.Map((np.array(selected_image), selected_hdr))

[15] plt.figure(figsize=(5,5))
ax = plt.subplot(projection=my_map)
my_map.plot();
```

AIA 171 Å 2010-08-17 10:06:00



SDOML Projects

The screenshot shows a web browser window at spaceml.org displaying a grid of eight project cards. Each card features a header image, a 'PROJECT' label, a title, a brief description, a challenge area, and a program affiliation.

Project Title	Challenge Area	Program
SDOML A Machine Learning Dataset Prepared From the NASA Solar Dyna...	SPACE WEATHER	FDL US
TRACKING THE GEOEFFECTIVENESS OF SOLAR STORMS Determining new representations of "Geoeffectiveness" using ...	SPACE WEATHER	FDL US
ENHANCED PREDICTABILITY OF GNSS DISTURBANCES AuroraML	SPACE WEATHER	FDL US
TEMPERATURE MAPS OF THE SOLAR ATMOSPHERE DeepEM: A deep-learning approach to differential emission me...	SPACE WEATHER	FDL US
SDO: AUTOCALIBRATION Multichannel auto-calibration for the Atmospheric Imaging As...	SPACE WEATHER	FDL US
NASA GIBS WORLDVIEW SIMILARITY SEARCH A No-Code, Self-Supervised Learning and Active Labeling Tool...	EARTH SCIENCE	FDL US
SUPER-RESOLUTION MAPS OF THE SOLAR MAGNETIC FIELD State of the art deep neural networks to calibrate and super...	SPACE WEATHER	FDL US
DIGITAL TWIN EARTH Can we lower the cost of accurate global precipitation forec...	EARTH SCIENCE	FDL EUROPE

Table of contents

- Expanding the Capabilities of SDO: SDO/AIA Autocalibration
 - Introduction
 - Table of Contents
 - Setting up the notebook
 - Reading and loading the SDO/AIA data
 - Autocalibration Inference
 - 3a. Multi-channel Model
 - Converting all times to astropy time.
 - Defining two functions to use as a median filter
 - Plotting the Degradation Curve
 - Downloading & Correcting AIA images
 - Generating the AIA Autocalibration correction table
 - Plotting**

+ Code + Text

RAM
Disk

```

[18] fig = plt.figure(figsize=(len(maps) * 3, 6))
plt.subplots_adjust(wspace=-0.2, hspace=0)

for i, (m, mc) in enumerate(zip(maps, maps_corrected)):
    ax = fig.add_subplot(2, len(maps), i + 1, projection=m)
    m.plot(axes=ax, norm=norm, annotate=False)
    ax.set_title(m.date.datetime.year)
    ax.coords[0].set_ticks_visible(False)
    ax.coords[0].set_ticklabel_visible(False)
    ax.coords[1].set_ticks_visible(False)
    ax.coords[1].set_ticklabel_visible(False)
    ax.set_aspect("equal")

    ax = fig.add_subplot(2, len(maps), i + 1 + len(maps), projection=mc)
    mc.plot(axes=ax, norm=norm, annotate=False)
    ax.coords[0].set_ticks_visible(False)
    ax.coords[0].set_ticklabel_visible(False)
    ax.coords[1].set_ticks_visible(False)
    ax.coords[1].set_ticklabel_visible(False)
    ax.set_aspect("equal")

```

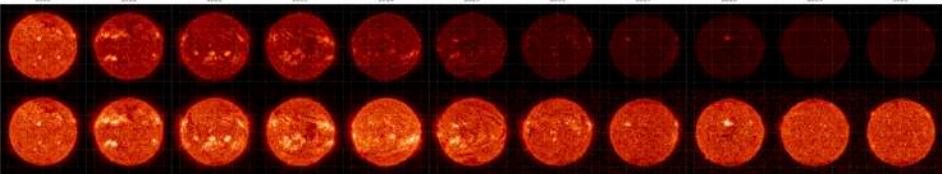


Figure 3: Top: Yearly observations of SDO/AIA 304 A data as observed by SDO/AIA, plotted on the same scale. Bottom: As top, but with corrected for time-dependent degradation using the AIA Autocalibration method.

22s completed at 19:45
✕

Summary

SDOML (v2.0) on SpaceML

- SDOML is a **cleaned and curated** dataset covering 2010 - present
- SDOML (v2.0) is available on SpaceML
- Notebook in place to demo data access
 - Small subset freely accessible
 - Full dataset available on request
- SpaceML provides examples of ML projects that use SDOML, along with notebooks to reproduce the results (e.g. AIA Auto-calibration)

SpaceML

Projects & Links

SDOML:

Galvez et al. 2019: [arxiv:1903.04538](https://arxiv.org/abs/1903.04538)

Nowcasting SDO/EVE with SDO/AIA

Szenicer et al. 2019 [science.org/10.1126/sciadv.aaw6548](https://science.org/doi/10.1126/sciadv.aaw6548)

SDO/AIA Autocalibration:

Dos Santos et al. 2021: [arxiv:2012.14023](https://arxiv.org/abs/2012.14023)

SpaceML: spaceml.org

Github: [spaceml-org/helionb-sdoml](https://github.com/spaceml-org/helionb-sdoml)



SPACEML The logo graphic for SpaceML, consisting of a small blue dot, a small orange dot, and a red plus sign, with a small white plus sign below the orange dot.

**SpaceML is a machine learning
toolbox and developer community
building **open science** AI applications
for space science and exploration.**

Find out more at spaceml.org

This enhanced data product was made possible by:



PARTNER



FRONTIER
DEVELOPMENT
LAB



FDL is a public private partnership between NASA, the SETI Institute, Trillium Technologies and leaders in commercial AI, space exploration and Earth Science. NASA provides funding for the Frontier Development Lab (FDL) through a co-operative agreement with the SETI Institute.