UNIVERSITÉ DE GENÈVE

University of Applied Sciences and Arts Northwestern Switzerland

Lay Summary

Machine Learning based Analytics for Big Data in Astronomy

Project team
Prof. Svyatoslav Voloshynovskiy
Prof. Samuel Krucker
Prof. Martin Melchior
Prof. Lucia Kleint (external member)
Dr Cédric Huwyler
Dr Brandon Panos
Denis Ullmann

Contact addresses

Prof. Svyatoslav Voloshynovskiy
University of Geneva
Department of Computer Science
7, route de Drize
1227 Carouge/Geneva
+41 22 379 01 58
svolos@unige.ch

Prof. Samuel Krucker
Fachhochschule Nordwestschweiz FHNW
Hochschule für Technik
Bahnhofstrasse 6
5210 Windisch
+41 56 202 77 04
Samuel.krucker@fhnnw.ch

Prof. Martin Melchior
Fachhochschule Nordwestschweiz FHNW
Hochschule für Technik
Bahnhofstrasse 6
5210 Windisch
+41 56 202 77 07
martin.melchior@fhnw.ch

01.09.2021

FNSNF Swiss National Science Foundation

75 NRP Big Data National Research Programme

1. Background

The regular eruptions that take place on the sun can cause disturbances on Earth – for instance in radio and GPS positioning systems – as well as power outages. To date, scientists have been unable to explain the physical cause of solar eruptions, nor can they reliably predict them. The fact that solar eruptions occur in diverse complex spatial and temporal patterns makes it considerably more difficult to systematically analyse these phenomena.

2. Goals of the project

This project focuses on the analysis and possible prediction of solar flares. Solar flares are intense bursts of radiation originating from the release of magnetic energy associated with so-called active regions on the Sun. Strong flares can have widespread geomagnetic effects and can lead to the failure of power grids (as reported for example 1989 in Canada and 2003 in Sweden) which in our current technologically saturated society can have devastating repercussions for both the economy, as well as communication and navigation systems. The released high energy particles are able to affect or destroy satellite electronics through direct impact or discharge of immense electric potential differences that were built up. Transpolar flights have to be rerouted at the occurrence of solar flares to avoid unhealthy radiation doses for crew and passengers. Moreover, it has been reported that a strong solar flare could negatively affect the stratospheric ozone layer for months.

The data for our analysis are provided by NASA's Interface Region Imaging Spectrograph (IRIS)[1] to which a project member (Dr Lucia Kleint) contributed at the stage of mission development. The IRIS satellite is designed to observe the solar transition region, a complex dynamic interface region between the solar photosphere and corona. For this project, we have 30 TB of spectral and image data at our disposal that was never before examined with machine learning methods and also not at this large scale. We applied our machine learning based spectral and image processing tools in order to better understand the physics behind solar flares. **Our primary goals were the identification and classification of all flares in the data collected by IRIS, understanding their dynamics and if possible setting up a system that can predict them over a certain time horizon.**

Therefore, **the main objective of this project was to elucidate the physics underlying solar flares and to develop capabilities to predict them.** For this purpose, we systematically analysed spatio-temporal patterns observed in vast amounts of recently available solar data by applying big data analytic techniques. The patterns observed in the data are extremely complex and varied. Not surprisingly, conventional manual analysis conducted by solar scientists in the past has fallen short and no reliable method for forecasting flares has been developed so far. Instead, we automated this analysis by using state-of-the-art and custom-tailored machine learning algorithms for representing and classifying patterns. Ultimately, this facilitated models for predicting solar flares and generated new physical insights.

Our project was divided into a set of five work-packages with individual milestones that are designed to follow the timeline designed in the proposal. This is a short summary of the individual work packages: In WP1 (System Requirements, System Architecture, System Setup) we evaluated the system

[1] http://iris.gsfc.nasa.gov/

requirements for a big data analytics platform, the Solar Data Analysis Stack, that enables fast, interactive analysis loops usable by all three research groups. This included giving fast access to the original data, allowing for efficient and flexible pre-processing and providing a suitable computing infrastructure for all required, possibly large-scale machine learning applications. WP2 (Data Preparation) encompassed the creation of a data pre-processing pipeline, the creation of labelled IRIS data and the identification of training and test sets. WP3 (Machine Learning Techniques for Characterizing IRIS Data) involved the evaluation of existing machine learning algorithms for feature extraction and classification of IRIS data and also the creation of new algorithms customized for hyper-dimensional temporal datasets. In WP4 (Processing, Testing, Validating, Analysing Results) developed user-friendly tools for efficient evaluation of our algorithms and to synthesize results. Finally, we made our results available to different communities in WP5 (Dissemination of Results) with scientific publications in the fields of astrophysics and computer science and also with public outreach. In addition, the extended part of the project covered additional three work packages: Ext-WP1: Attention Mechanism, Ext-WP2: Incorporating IRIS2 (extended to multi spectral lines) and Ext-WP3: Dissemination. All work packages are completed and summarized in the second part of this report.

Solar flares are powered by the Sun's magnetic field, which can progressively become more complex as sub-photospheric currents slowly twist and contort the coronal magnetic fields into higher energy configurations. At some point the fields can reconnect, releasing large amounts of magnetically stored energy to accelerate charged particles to sub-relativistic speeds away from and into the lower solar atmosphere. This results in the heating of plasma and subsequent enhancement of the entire electromagnetic spectrum. Each part of this spectrum offers key insights into the underlying physical mechanisms of solar flares, and some parts have been explored more thoroughly than others.

There are several key deficiencies and holes within the field of solar flare physics, which our research has helped advance and close. Firstly, the development of the standard flare model, which is a theoretical framework explaining the evolution of flare energy, has slowly been accreted through the study of single flare events. Although large solar datasets have existed for some time now, the community has found it manually taxing to explore multiple events simultaneously. Therefore, statistical studies, which would be extremely instructive, are often avoided. In this light, ML has granted us easy access into the space of massive solar datasets and allowed us to perform large statistical studies containing many solar flares with ease. The results and methodologies birthed from this type of research can be used to test the standard flare model to an unprecedented scale.

Secondly, although the high energy/short-wavelength X-ray regime immediately preceding the magnetic reconnection event is well understood, the ultraviolet counterparts to this emission have received less attention. Since most of the flare energy is dissipated outside of the X-ray regime, this type of emission harbours critical information about flare physics. The coupling between IRIS and ML has allowed us to start analysing this critical period of a flare's lifecycle, called the impulsive phase, at longer near- and far-ultraviolet wavelengths. The regions of highest energy flux called the flare ribbons display extremely unique behaviour in the UV, which we have managed to start documenting.

Lastly, almost all flare forecasting publications to date have focused on a single dataset called the SHARP dataset. This dataset consists of a large number of features derived from continual full HMI vector magnetograms, and reportedly contains important information regarding the magnetic topology of the solar atmosphere. These magnetic maps are derived from the solar photosphere, where magnetic field strengths are significant enough to affect the atomic polarization states of certain atomic ions under the action of the Zeeman effect. Although this magnetic dataset undoubtedly accurately reproduces the state of the field in the lower atmosphere, it is uncertain how much information these maps contain with regards to the structure of the coronal fields, which is the key protagonist of solar flares, and is separated from the photosphere by a couple hundred thousand kilometres. Although the prediction methods and algorithms have become increasingly sophisticated over the last decade, the prediction results have stagnated. Our solution was to attack the field of flare prediction from an orthogonal angle, using spectral data for the first time instead of the standard photospheric magnetograms. First results indicate that IRIS spectral data could provide a useful enrichment of the standard magnetic dataset, as we managed to predict a solar flare half an hour before it occurred. The community is extremely excited about these results and further research is underway.

3. Methods

To cover all aspects of this domain-specific big data project, we proposed an interdisciplinary approach with a project team consisting of astronomers, who actively contributed to the development of the IRIS mission, experts in the development of machine learning and statistical information processing tools, and specialists in designing, implementing and maintaining data management systems for Big Data astronomy projects.

To answer our research questions and reach our research goals we used a broad range of methods from both the traditional supervised and unsupervised machine learning repertoire, as well as from very recent research with varying complexity. We found that many methods were not perfectly suited to our problem set and required modifications that could be generalized to the scope of many other big data problems, particularly in astrophysics. More specifically, we have used methods of unsupervised learning such as k-means clustering, variational and adversarial autoencoders, as well as extended the theoretical understanding of these algorithms based on the concept of the information bottleneck. We have also used traditional supervised methods covering SVN, trees and CNNs, more recent self-supervised methods based on contrastive learning and methods from weakly supervised learning, in particular multiple instance learning. Finally, we have used recent techniques of mutual information estimation such as MINE in applications to astrophysical data analysis.
The project is focused on the data provided by NASA's Interface Region Imaging Spectrograph (IRIS), at the beginning of the project, the newest and one of the most popular missions in solar physics. IRIS is part of the Heliophysics System Observatory fleet that consists of more than a dozen operating and planned satellites. The entire fleet's mission is to investigate different aspects of the Sun and its influence on Earth. IRIS is designed to observe the solar transition region, a complex dynamic interface region between photosphere and corona - whose understanding remains a challenge but is key for the

understanding of the physics of solar flares. Every few seconds, IRIS records images and spectra of varying targets on the Sun in the UV regime, with high resolution in space (1/3 arcsec ≈ 240 km on the Sun) and in time (2s).

After 8 years of operations, IRIS has accumulated an archive with more than 100 million spectra - far too many to be analysed manually - the use of automated and efficient big data analytics tools is therefore mandatory. Quite obvious to the modern data scientist, this example with the huge number of spectra and the huge variety of patterns found in them is a very promising use case for adopting machine learning techniques.

We used 30 TB of IRIS satellite data, which we were granted access to download from the Lockheed Martin Solar and Astrophysics Laboratory (LMSAL). All data have been transferred to a local NAS at FHNW to our Solar Data Analysis Stack, a system with 96 TB of storage space and sufficient computing power and also shared with the University of Geneva team. This setup ensures sufficiently fast data access for big data processing applications on the entire dataset and provides enough performance for state-of-the-art machine learning and deep learning methods. The Solar Data Analysis Stack is designed to manage all the computing tasks in our project: from data synchronisation with LMSAL and efficient data pre-processing to high-performance machine learning tasks. Data management and infrastructure maintenance was an essential and time-intensive part of the project that should not be underestimated.

4. Results

According to the main objective of the project to elucidate the physics underlying solar flares and to develop capabilities to predict them, we have developed a number of methods to address this objective. The developed methods were applied to the real data produced by the IRIS NASA mission. The main obtained results are reported below. These results are reflected in the corresponding publications as well as presented during international conferences and workshops. The main focus of our study was on the analysis of spatio-temporal patterns observed in the IRIS data. It should be pointed out that the targeted problem is very challenging due to the numerous factors such as dimensionality and volume of data, data multimodality, lack of labels as such, unbalanced representation of different events in the data such as the regions of quite sun, pre-flare and flare. The quiet sun event represents a dominating majority in the data. To address these challenges, we needed to adapt the existing ML approaches that led us to new insights in solar physics that are reported below.

The project objectives are further divided on the number of research questions that are covered in the referred papers:

RQ1: "Identifying typical Mg II Flare spectra using machine learning" [1]: We performed a vector quantization on IRIS Mg II flare spectra using a modified version of the classical k-means algorithm. We found universal flaring spectra that appear in every flare. This spectra differs from the spectra in the quiet Sun significantly, not having the archetypal central reversal. We postulate and later substantiate, that these spectra are formed from the

conjoined action of a multilayered velocity field and the recoupling of the Plank and source function via increased temperature and densities within the upper chromosphere. Additionally, co-observations with X-ray instruments such as GOES and RHESSI indicate that these spectra occur spatially and temporarily with hard X-ray signatures. Performing a vector quantization via a modified version of k-means allowed us to construct a fundamental group of flaring profiles from millions of spectra across multiple flares. This step would prove to be more important than we initially anticipated, as it afforded us the opportunity to calculate abstract statistical quantities such as the mutual information (MI). Because this clustering algorithm is relatively straight forward, its results are easy to interpret, and physical insight can be extracted when coupling the quantized representations to other data sources such as IRIS slit-jaw images and GOES X-ray curves. The paper has accumulated many citations since its initial publication in Astrophysical Journal and was subsequently used by the group at Lockheed Martin as a way to perform fast inversions. The method now forms a core component of the IRIS2 database, which can be found on the IRIS missions main webpage https://iris.lmsal.com/. The methods of this paper are rapidly becoming popular in the community as a way to tractably analyse large solar datasets.

RQ2: "Exploring mutual information between IRIS spectral lines I" [2]: In this paper, we managed to harmoniously introduce methods to calculate information theoretic quantities on solar datasets. We showed that MI can be used as a measure of how connected the solar atmosphere is both under quiet Sun and flaring conditions. We used two methods to calculate the MI between pairs of spectral lines sourced over a variety of formation heights, from the photosphere (FeII) to the transition region (SiIV & FeXXI). The first method involved a categorical transformation and the use of the maximum information criterion (MIC) to select the most suitable granularity for our problem. The second method used a recently developed neural network called a mutual information neural estimator (MINE-network) which avoids the use of a categorical transformation and eliminates the free granularity variable, allowing us to calculate the MI on the raw spectral data. Both methods converged to the same results, indicating weak atmospheric coupling within the quiet Sun, and strong correlations between different heights during solar flares. The pairs of MI will now serve as a benchmark for simulations to try and reconstruct and has received much interest from one of the community's leading simulation experts Dr Graham Kerr.

RQ3: "Exploring mutual information between IRIS spectral lines II" [3]: This was a follow up paper to the first instalment of our MI study. From the first publication, it became clear that some lines couple very strongly with high MI scores. This implies that if we know one of the line's output, the probability distribution for the other line's output should be calculable. In this paper, we therefore calculated the conditional probabilities over all IRIS spectral windows in response to a single known Mg II spectrum. This has led to a number of additional constraints for simulations and a "deforestation" of the degenerate thermodynamic solution space. We also calculated the point-MI and were able to isolate the regions that were most responsible for the high correlations, both temporarily and spatially. We found that the MI was maximum in step with the GOES derivative, in a location directly over the flare ribbons. This result shows definitively that the solar atmosphere is most connected over regions of largest energy deposition.

RQ4: "Real-time flare prediction based on distinctions between flaring and non-flaring active regions spectra" [4]. This paper has captured the imagination of the community and introduced a fresh approach for predicting solar flares. Instead of using the same HMI magnetic dataset as almost all previous publications, we attempted to predict solar flares based on spectra alone. This was done by teaching a neural network the difference between spectra taken from active regions that don't experience a flare, from active regions (called pre-flare regions) that do. It is natural to assume that the network will make fewer labelling errors the closer we come to a flare in time. Under this assumption, the performance of the network can be used as an early warning signal for solar flares in

real time. This has prompted the community to think about incorporating additional datasets alongside the standard magnetograms from HMI and has also won an invited talk at the next IRIS meeting in Washington DC.

All of our published articles have received much interest from the community, and the Astrophysical Journal publications are all featured as IRIS highlights on the IRIS instruments main page. The link to the highlight page is given here
https://iris.lmsal.com/science_highlights/?cmd=view-archive.

At the same time, we have also addressed several important technical issues related to the processing, sharing and reduced complexity training and classification of Big Data:

RQ5: "DCT-Tensor-Net for solar flares detection on IRIS data" [5]: In this work, we focus on 8 terabytes of raw data with only a small fraction representing approximately a thousand of flares. Our data are asymmetric and also not labelled, which oriented us toward unsupervised or semi-supervised methods. We proposed a tool for automatic flare detection and analysis of IRIS videos, based on a video tensor harmonic analysis, by composing time component discrete cosine transform (1D- DCT) with spatial discrete cosine transform of the images (2D-DCT). This is one of the first tools for detecting flares based on IRIS images/videos. Our method reduces the false detections of flares by taking into consideration their specific local spatial and temporal patterns. A proposed DCT Tensor Network that reduces the size of the input video tensor by harmonic decomposition of the signal by empirically selected range of frequency domains for flare detection. This method is a straightforward handcrafted Network for which all available data is used for testing and for the empirical selection of the frequency domain for flare detection. Finally, we produce an output of a classifier indicating the probability of flare observation. The proposed method for flare detection on IRIS data and may also be of interest for further labelling and analysis of the flares. We believe that this structure could be useful for other scientific video observations analysis. The classification or clustering results of our method should be subjected to future cross-validation and correlation with the analysis made from the spectral data of IRIS in order to better understand the solar flare phenomenon.

RQ6: "Solar activity classification based on Mg II spectra: towards classification on compressed data" [6]: In this paper, paper we have used the information-theoretic framework to demonstrate that the reliable classification of Mg II spectra is still possible based on the compressed data with a negligible loss with respect to the raw data. The training of classifiers of several families on the compressed data confirmed this finding. To achieve this result we have used an information bottleneck formulation for the supervised and unsupervised systems and then merged the encoder of the trained system producing the latent space representation with the classifier of the supervised one. Such a compressed representation allows to largely reduce the complexity of storage, communication and training in Big Data applications that can be extended well beyond the addressed problem.

RQ7: "Information Bottleneck Classification in Extremely Distributed Systems, Entropy journal" [7]: In this work, we further extended the concept of the information bottleneck to a new problem formulation when the data obtained in different observations, missions or centres cannot be shared or communicated to the centralized server for the classifier training for various reasons covering big volume, privacy, restricted communication bandwidth, etc.. For the first time, we have demonstrated a principal feasibility to train a novel system where no information is shared with the centralized server or between the nodes where partial information is stored. The proposed architecture consists of a number of information bottleneck auto-encoders representing lossy compression modules. Once trained for a given class, each module provides an optimal encoding and decoding for the bounded latent space with a fixed compression rate. The results obtained on several public data sets confirm the fact that the proposed system provides the state-of-the-art performance comparable with those of fully supervised classifiers where all samples from all classes are present at the centralized node for training. We

believe that such a finding can be of great interest for future observation missions and remote sensing where the communication bandwidth is limited to provide data in raw format to the centralized nodes. Furthermore, we also believe that the proposed system might be of interest for hospitals and research labs dealing with Big Data subject to privacy concerns.

RQ8: "Prediction of solar spectra based on inpainting self-supervised learning" [8]: This is our ongoing work originating from the current project. In this work, we present a deep neural network approach designed to predict multi-dimensional time-sequences representing solar activity recorded by IRIS. We proceed with a time-sequence encoded into a two-dimensional representation in the form of images. Accordingly, we use the recent advances in image-extension to predict an unseen part of an image from a given one. We show a conceptual link between the proposed approach and classical time-series modelling and predictions by Recurrent Neural Networks (RNN) and Long-Short-Term-Memory (LSTM). Actually both time-series and pixels of images share an order property. But we claim that the convolutions and overall structure of the image-extension network have some flexibility with the cadence of the time-sequence, which is not the case for RNNs and LSTMs. This approach is useful when the time-sequence data that we want to predict present many possible cadences. There are several benefits of such predictions: to predict the missing ending of an observation and then overcoming the memory constraints of the satellite, to assist in the planning of observations and to predict the physics of the sun including critical events.

RQ9: "Using Multiple Instance Learning to Predict Solar Flares" [9]: In this work we tackle the problem of missing labels at the spectral level using the Multiple Instance Learning (MIL) paradigm. In contrast to standard supervised learning, in the MIL setting label information is not available at the level of single instances, but only for bags thereof. Nevertheless, it is possible to learn common concepts in bags with the same label that can then be used to classify new bags and to identify instances that contribute strongly to the decision process. MIL is often used for automatic image segmentation given only image-level labels. We prepare a manually selected set of active region observations that either produce a flare within the next few hours (pre-flare / PF) or do not (active region / AR). Each observation timestep can be treated as a bag of spectra and is labelled either as AR or PF. With MIL, this dataset can be leveraged to a) predict flares at the observation level, b) propagate the bag labels AR/PF to single spectra and c) to produce saliency maps that highlight regions of interest for flare prediction relying on an attention-based mechanism. We have not yet seen this approach for solar data and see it as a tool with enormous potential to deal with heterogeneous data that is labelled only at a higher hierarchical level. In further research we plan to apply this method to other astronomical images where events are indicated to have occurred in an image, but the precise location has not been labelled.

To have the necessary tools to address the above research questions we developed the Python library IRISreader[2]. At the start of the project, no Python library to access IRIS data was available and everybody relied on tools in the Interactive Data Language (IDL) that is a Fortran descendant and more suited to perform analysis than efficient programming. For this analysis, we needed a tool to build easy and efficient end-to-end data and machine learning pipelines in Python. IRISreader has been acknowledged by the IRIS community[3].

Refereed publications:

[2] https://github.com/i4Ds/IRISreader

[3] https://iris.lmsal.com/itn45/IRIS-LMSALpy_chapter1.html

1. B. Panos, L. Kleint, C. Huwyler, S. Krucker, M. Melchior, D. Ullman, and S. Voloshynovskiy, *Identifying Typical Mg II Flare Spectra Using Machine Learning,* Astrophysical Journal, Vol. 861, Number 1, p.62, 2018.
2. B. Panos, L. Kleint, S. Voloshynovskiy, *Exploring Mutual Information between IRIS Spectral Lines. I. Correlations between Spectral Lines during Solar Flares and within the Quiet Sun*, Astrophysical Journal, Vol. 912, Number 2, p.121, 2021.
3. B. Panos, L. Kleint. *Exploring mutual information between IRIS spectral lines. II. Calculating the most probable response in all spectral windows,* Astrophysical Journal, Vol. 915, Number 2, p.77, 2021.
4. B. Panos, L. Kleint, *Real-time Flare Prediction Based on Distinctions between Flaring and Non-flaring Active Region Spectra*, Astrophysical Journal, Vol.891, Number 1, p.17, 2020.
5. D. Ullmann, S. Voloshynovskiy, L. Kleint, S. Krucker, M. Melchior, C. Huwyler, and B. Panos, *DCT-Tensor-Net for solar flares detection on IRIS data*, 7-th European Workshop on Visual Information Processing (EUVIP), Tampere, Finland, 2018.
6. S. Ivanov, M. Tsizh, D. Ullmann, S. Voloshynovskiy, *Solar activity classification based on Mg II spectra: towards classification on compressed data*, Astronomy and computing, 2021.
7. D. Ullmann, S. Rezaeifar, O. Taran, T. Holotyak, B. Panos, and S. Voloshynovskiy, *Information Bottleneck Classification in Extremely Distributed Systems*, Entropy journal, No 11, Article number 1237, in special issue Information-Theoretic Methods for Deep Learning Based Data Acquisition, Analysis and Security, 2020.
8. D. Ullmann, S. Voloshynovskiy, L. Kleint, S. Krucker, M. Melchior, C. Huwyler, and B. Panos, *Prediction of solar spectra based on inpainting self-supervised learning,* To be submitted in October 2021.
9. C. Huwyler, M. Melchior. *Using Multiple Instance Learning to Predict Flares.* In preparation. To be submitted in August 2021.

Talks and Posters at Conferences:

1. *Multi-wavelength observations of major solar flares* (invited)
   Lucia Kleint, April 3, 2018.
   EWASS 2018, Arena & Convention Centre, Liverpool, UK
2. *Do all flares share the same chromospheric physics?*
   Brandon Panos, June 19, 2018.
   17th RHESSI Workshop, Trinity College, Dublin, Ireland
3. *Do all flares share the same chromospheric physics?*
   Brandon Panos, June 29, 2018.
   IRIS-9, Max Planck Institute for Solar System Research, Göttingen, Germany.
4. *Flare detection from discrete Fourier transforms of SJI* (Poster)
   Denis Ullmann, June 25-29, 2018.
   IRIS-9, Max Planck Institute for Solar System Research, Göttingen, Germany
5. *IRISreader - A Python Library for IRIS Data Processing* (Poster)
   Cédric Huwyler, June 25-29, 2018.
   IRIS-9, Max Planck Institute for Solar System Research, Göttingen, Germany
6. *Invited Seminar, including some parts of this project*
   Lucia Kleint, October 19, 2018
   Solar Physics Seminar, University of Peking, China
7. *DCT-Tensor-Net for Solar Flare Detection on IRIS Data*
   Denis Ullmann, November 26, 2018.
   EUVIP 2018, Tampere University of Technology, Tampere, Finland.
8. *Tracking Coronal Rain with Machine Learning Methods*
   Brandon Panos, January 21, 2019
   ISSI Meeting on Observed Large-scale Variability of Coronal Loops as a Probe of Coronal Heating, Bern, Switzerland
9. *Machine Learning and Solar Flares.*
   Brandon Panos, March 6, 2019
   3rd SCOSTEP workshop, PMOD/WRC, Davos, Switzerland

10. *Match-Making in Big Data with Academia and Industry*, NRP-75 talk by Cédric Huwyler, November 10, 2020
11. *The connected Sun: Correlations between spectral shapes from different ions*, Brandon Panos, Seminar at the University of Glasgow, February 11, 2021
12. *Learning Interpretable Features*, Martin Melchior, June 8, 2021
    First International Symposium of the Science of Data Science
13. *SPD Harvey prize talk: "A journey from quiet Sun magnetic fields to Flares"*, (invited), Lucia Kleint, June 2021,
    SPD AAS meeting / online

Completed PhDs:

This project led to one completed PhD (Brandon Panos, *"The Analysis of Solar Flares Using Machine Learning"*, defended 10 May 2021) and one PhD that is at final stage (Denis Ullman).

5. Significance of the results for science and practice

In this project we targeted to better understand the physics under solar flares and to develop techniques to predict them. The main focus of our project was on the machine learning analysis of spatio-temporal patterns produced by the IRIS mission. The main scientific implications for practice and science are:

- We have shown that high resolution spectral data from the ultraviolet regime contains a high potential for a deeper understanding and a possible prediction of solar flares and solar activity in general. Solar activity has a continuous impact on Earth's geomagnetic field that encompasses the Earth's surface and the near-Earth space environment. Thus it can affect our daily lives and other fields of study: long distance radio communication can be disturbed or disrupted, satellite electronics can be damaged or destroyed, transpolar flight trajectories have to be rerouted due to increased radiation doses and communication problems (which leads to increased CO2 emission), the ozone layer can be weakened for several months and power grids can be disturbed or disrupted completely for several hours.
- The presence of labelled data is highly unlikely in many Big Data applications. One important scientific implication of the project is the development of unsupervised machine learning tools for automatic data clustering and analysis of statistical relationships. We think that these techniques are of high significance for both science and practice.
- One more implication is a possibility to perform the reliable classification of complex physical phenomena on specially designed compressed data that leads to the considerable simplification of training complexity and requirements to computational infrastructures. Furthermore, such a compression might be moved directly on data-sensors and data acquisition devices that will in addition drastically reduce the communication burden in Big Data applications. Finally, the developed techniques might be of great interest for privacy preserving applications where the utility attributes can be encoded into the compressed representations while the privacy sensitive attributes will be compressed and removed.

We are convinced that the scientific findings and technical outcomes of this project might be of great interest for many interdisciplinary projects facing similar challenges related to Big Data.