PI:

Project title: Machine Learning based Analytics for Big Data in Astronomy

75 NRP **Big Data**
National Research Programme

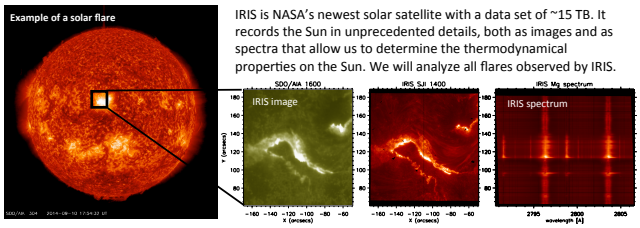
Big Data in Astronomy and Solar Physics

Astronomical observatories record very large data sets that cannot be fully analyzed manually. The solution is to apply big data analytics techniques to such data sets to derive scientifically valuable information.

Main Domain Science Objective

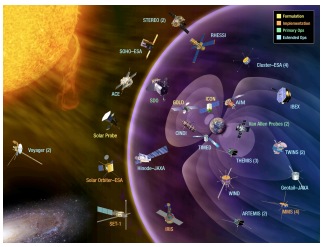
The main objective of this project is to elucidate the physics underlying solar flares and to develop capabilities to predict them. Solar flares are highly energetic explosions on the Sun, which may affect the Earth's communication, power grids, satellites, and which cause auroras.

Solar Observatories, IRIS and the Data



IRIS is NASA's newest solar satellite with a data set of ~15 TB. It records the Sun in unprecedented details, both as images and as spectra that allow us to determine the thermodynamical properties on the Sun. We will analyze all flares observed by IRIS.

IRIS is part of NASA's Heliophysics System Observatory, a fleet of satellites designed to study the Sun and its influence on Earth.

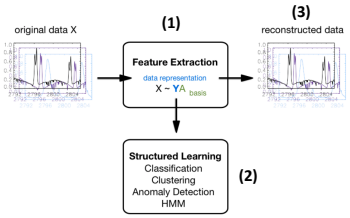


IRIS was launched in 2013 and has since observed more than 500 solar flares, resulting in millions of spectra to analyze.

Methodology

Efficient Representations and Machine Learning Techniques

In the roughly 200 million spectra a huge variety of patterns is found. Massive models with large parameter space will be needed to capture this variety. This will lead to sparse ('shallow overcomplete') repres-entations of the data and will provide the basis for **feature extraction (1)** .



The extracted features will be used for classifying the spectra and which is the basis for learning more about the underlying physics (**structured learning (2)**).

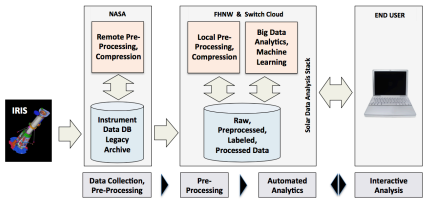
The sparse representations can also be used for designing efficient data compression schemes which are indispensible for an efficient transport of the data from the observatories / data archives to the data processing centers where the **data** can be **reconstructed (3)**.

Data Management

A solidly engineered data management infrastructure is key for large scale execution of machine learning algorithms and for evaluating the quality and efficiency of different algorithms applied to our data.

Existing platforms and platforms such as Spark, TensorFlow or Theano, R or WEKA will be reused and combined as far as possible into an efficient data analysis software stack. Particular emphasis will be given to providing a scalable and extensible system that can also by re-used for the analysis of other (even larger) datasets. This also includes turning the involved algorithms into scalable programs that can efficiently be executed in large computing infrastructures.

Finally, we will develop interactive tools that will allow end users to efficiently perform the analysis of the huge dataset.

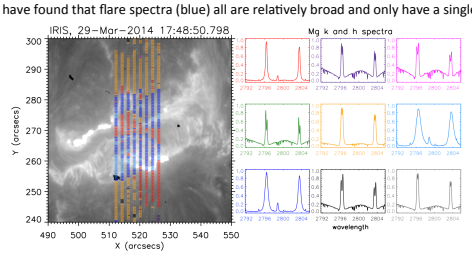


Aim of Project and Expected Results

Domain Science	We plan to answer the following domain science questions: • What types of spectra do exist? Are there common spectral features in solar flare spectra? Common features indicate common underlying physics, which in turn may help predicting flares. • Are there unusual spectra, e.g. spectra that differ from most others? Unusual spectra are equally interesting, because they hint at unusual conditions on the Sun. • Do certain types of spectra occur in spatially or temporally coherent ways? This gives us information on the solar evolution and answers whether all flares work in a similar way.
Machine Learning	By adopting a systematic approach to test the performance of different algorithms in representing and classifying the data we will identify the methods that work particularly well for the given problem at hand. In addition, we expect to come to a judgment to what extent the methods can be generalized to other data or even other science domains.
Data Mgmt	We expect to provide a robust and scalable data management infrastructure that will also provide visual tools to interact with the data.

Promising Preliminary Tests

In a preliminary test, we have applied a clustering technique (here k-means) to identify spectra of similar shapes. We have found that flare spectra (blue) all are relatively broad and only have a single peak.



Collaboration with other Projects