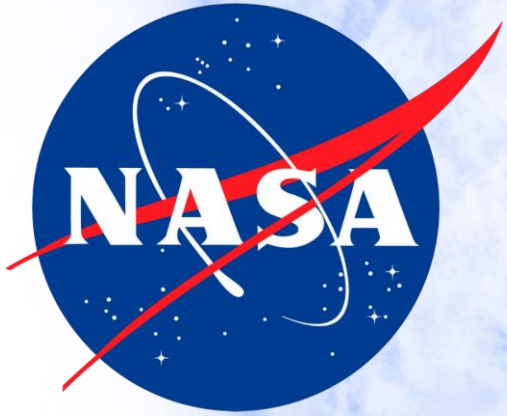
On Clustering of Machine Learning Attempts in Heliophysics: Examples and General Picture



Viacheslav Sadykov

Bay Area Environmental Research Institute

NASA Ames Research Center

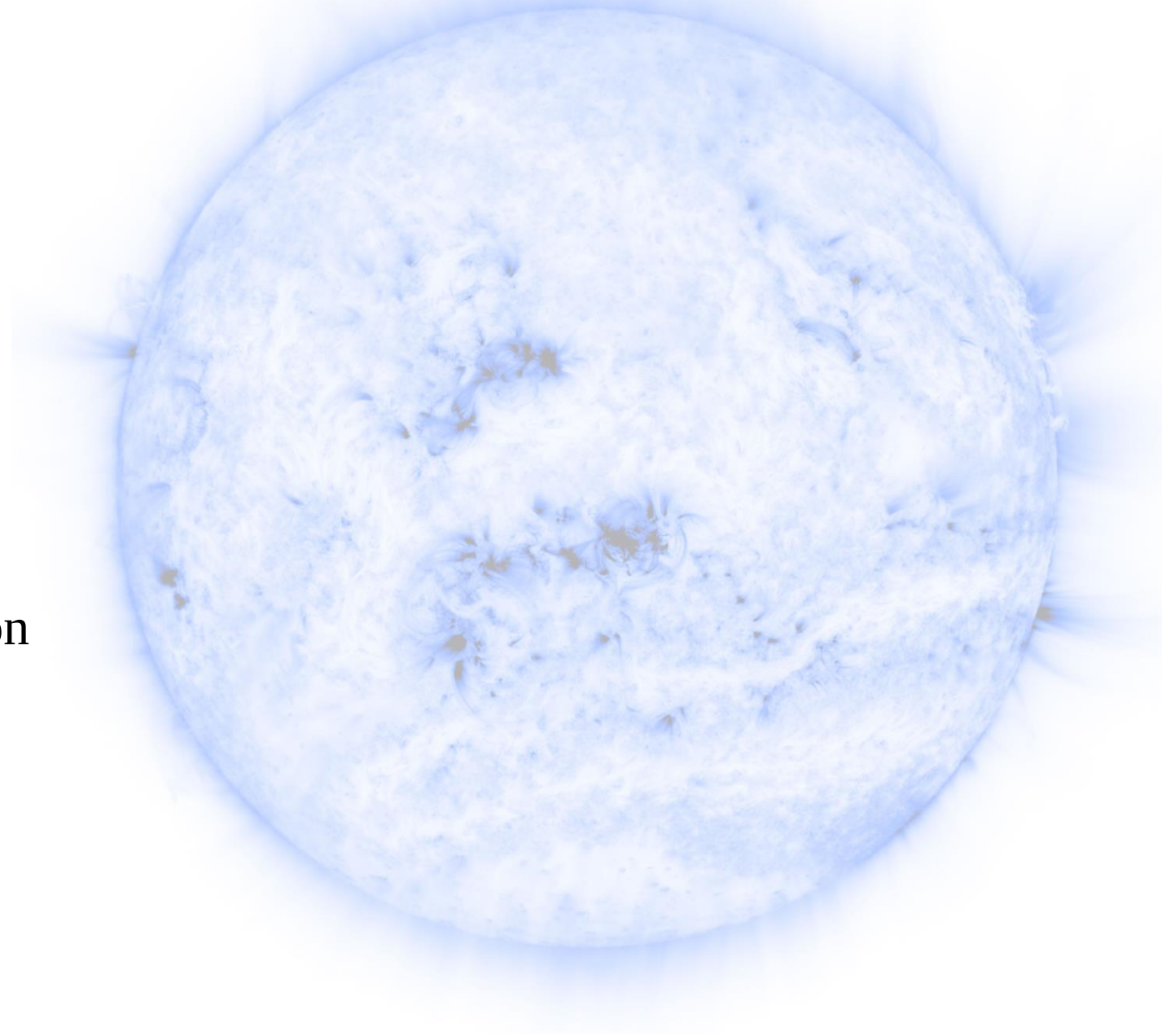
Bay Area Environmental Research Institute

EarthCube RCN Workshop: Machine Learning in Heliophysics and Space Weather Forecasting: Advances, Perspectives and Synergies

NJIT, January 16-17, 2020

Outline

- Introduction: the role of machine learning in Heliophysics
- Clustering of machine learning research publications
- Overview of each cluster: motivation and research examples
- Conclusions



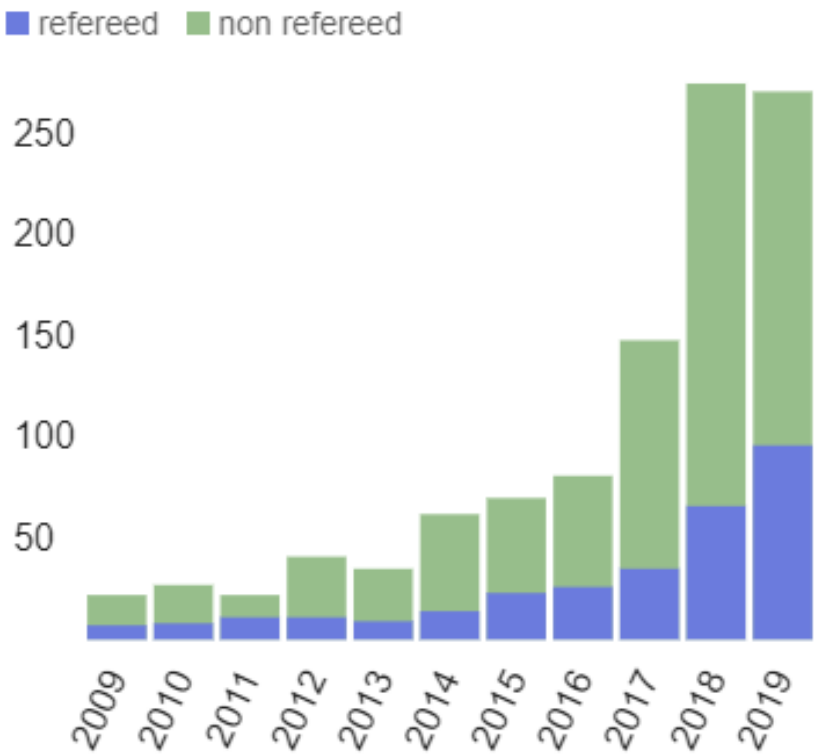
What is Machine Learning?

- Machine Learning: a study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instruction (Wiki)
- Employing machine learning is not magic: you usually must formulate very specific, very narrow tasks:
  - This does not work: “Let us predict solar flares.”
  - This may work: “Let us predict the probability of occurrence of solar flares of strength ≥ M1.0 GOES class in active regions within 24 hours after a certain considered time moment based on the data set X (full description of the data set including train-test separation).”
- To get an impression of how widely machine learning is currently used in Heliophysics, let us look at the publications.

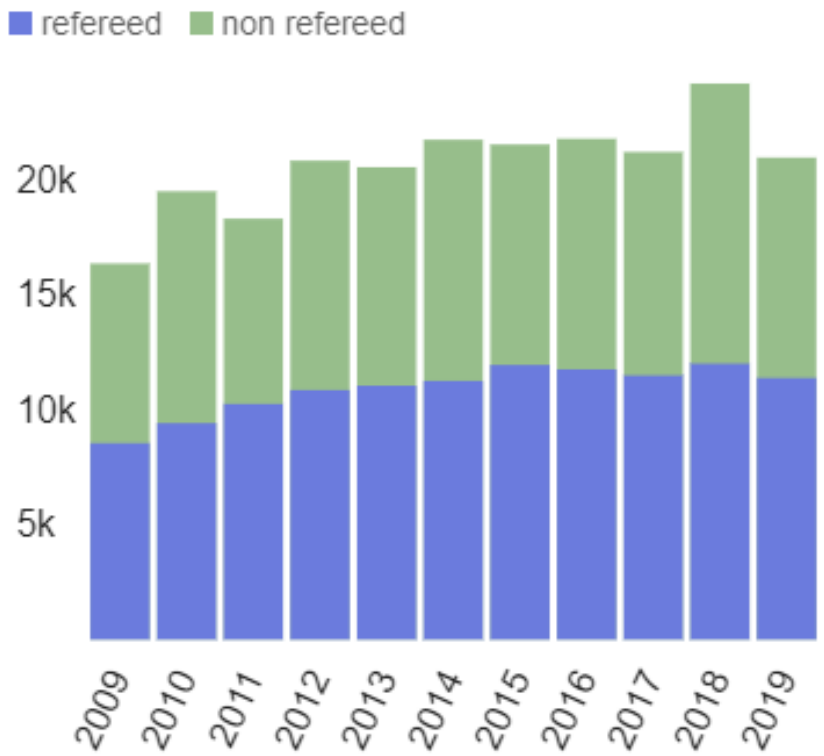
The role of the machine learning in Heliophysics: what do publication records tell us?

- NASA ADS (https://ui.adsabs.harvard.edu/) was used to gather statistics on the number of publications, citations, and number of reads.
- Statistics for the last 11 years (2009-2019) and for 2019 alone were compiled.
- The following search keywords were used to extract machine learning papers in Heliophysics:
  - “Solar” or “Heliosphere” or “Heliophysics” or “Space Weather” in Abstract
  - “Machine learning” or “Artificial Intelligence” or “Data Science” in Abstract
  - Search with no machine learning keywords for comparison

Citation patterns: Heliophysics (Credit: NASA ADS)



Number of papers
(Machine learning)



Number of papers
(all papers)

- The number of machine learning papers in Heliophysics is growing exponentially with respect to the number of all papers in the field. This is an intensively growing field now.
- A smaller fraction of machine-learning papers are refereed on average, but the situation is improving.

Citation patterns: 2009-2019 summary table (Credit: NASA ADS)

HELIOPHYSICS	Machine Learning	All papers
Number of papers	1,056 (0.46% from all)	227,403 (100%)
Number of citations	6,163 (5.84 per paper)	2,017,218 (8.87 per paper)
Number of reads (last 90 days)	49,985 (47.33 per paper)	2,553,061 (11.23 per paper)

- The number of reads during last 90 days is 4.2 times higher for machine learning papers.
- Machine learning papers are cited much less per paper on average than heliophysics papers overall. This may be an effect of the exponential growth in recent years.

Citation patterns: 2019 summary table

(Credits: NASA ADS)

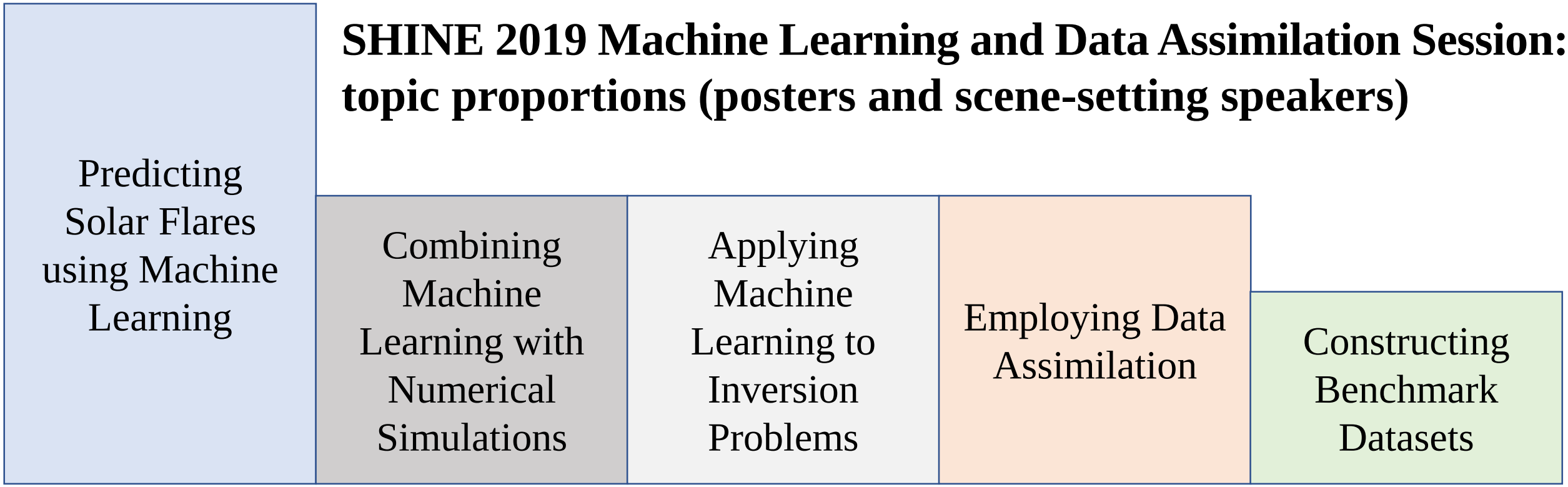
HELIOPHYSICS	Machine Learning	All papers
Number of papers	273 (1.29% from all)	21,204 (100%)
Number of citations	256 (0.94 per paper)	21,134 (1.00 per paper)
Number of reads (last 90 days)	31,834 (115.00 per paper)	881,780 (41.6 per paper)

- (Answering the question from the previous slide) Yes, this is an effect of the field's growth. Recent machine learning papers are cited at approximately the same rate as an average paper in 2019.
- The fraction of machine learning papers has grown with time (compare 1.29% for 2019 with 0.46% averaged over the last 11 years)
- Machine learning papers are just slightly less cited in average, but almost 3 (!) times more read compared to an average paper.

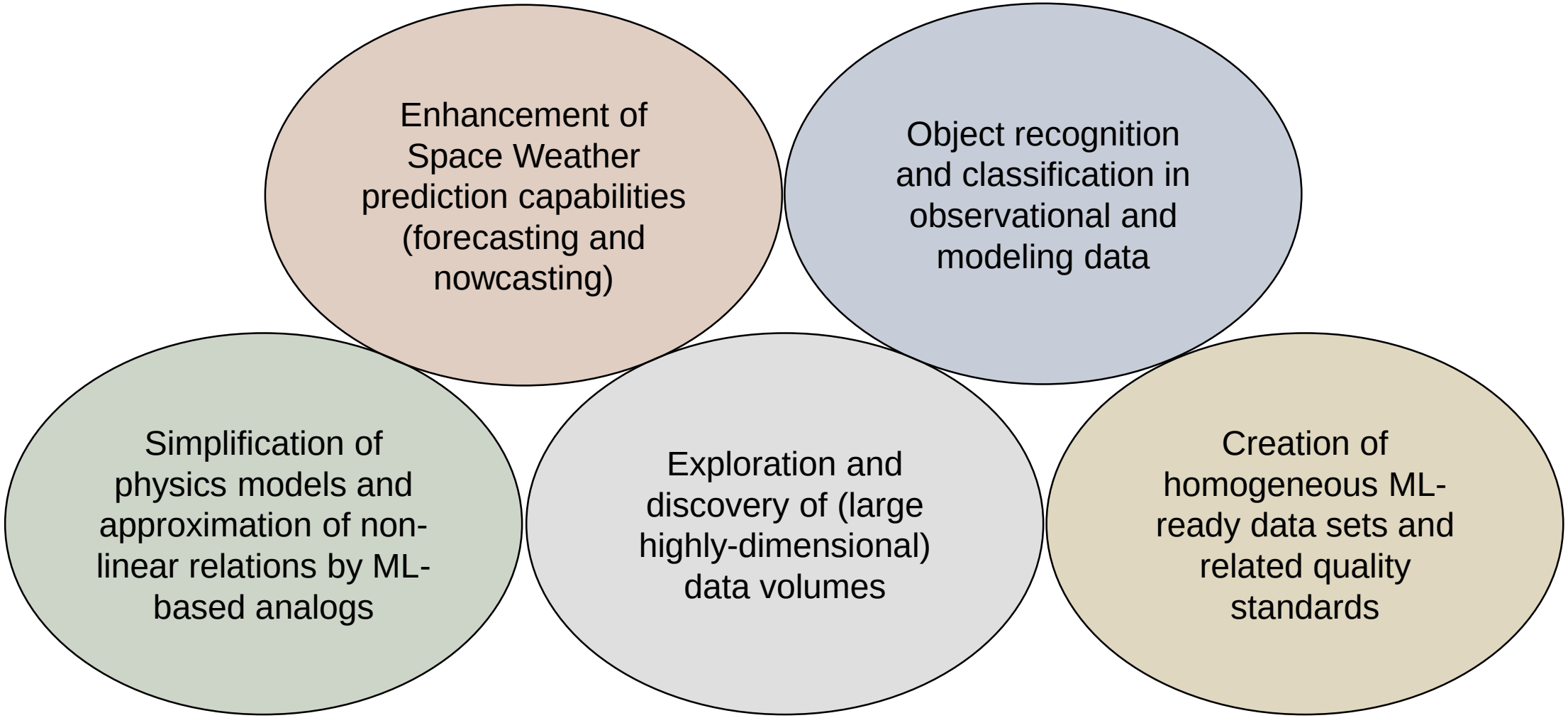
Can we classify machine learning research?

- It seems that machine learning research in Heliophysics can be classified into several categories, not based on type/algorithms but based on which physical problems they address.
- Example: classification of the contributions at the SHINE 2019 ML&DA Session.
- Can we build a more general classification?

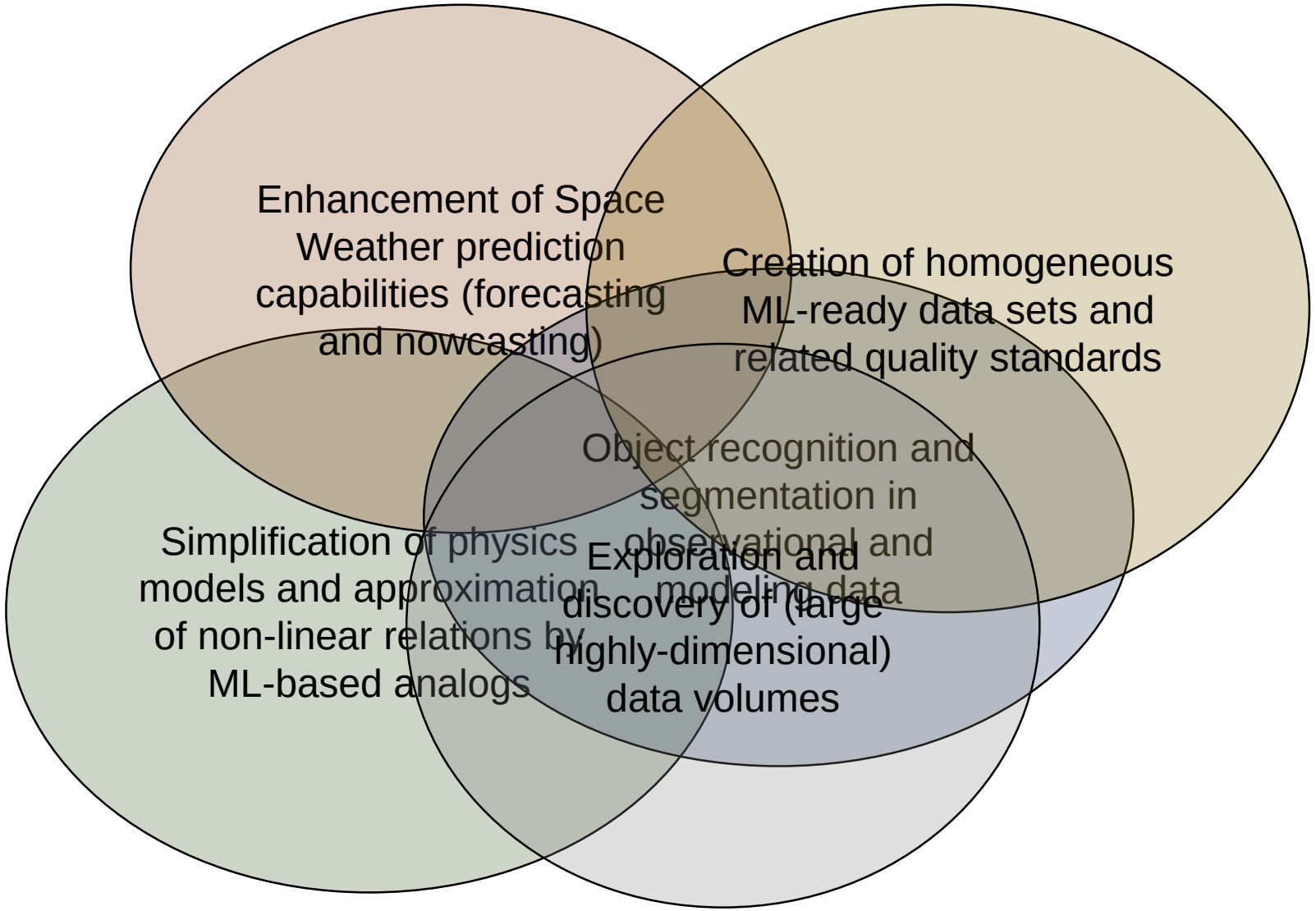
SHINE 2019 Machine Learning and Data Assimilation Session: topic proportions (posters and scene-setting speakers)



Machine learning research clusters in Heliophysics: personal impression



Machine learning research clusters in Heliophysics: personal impression



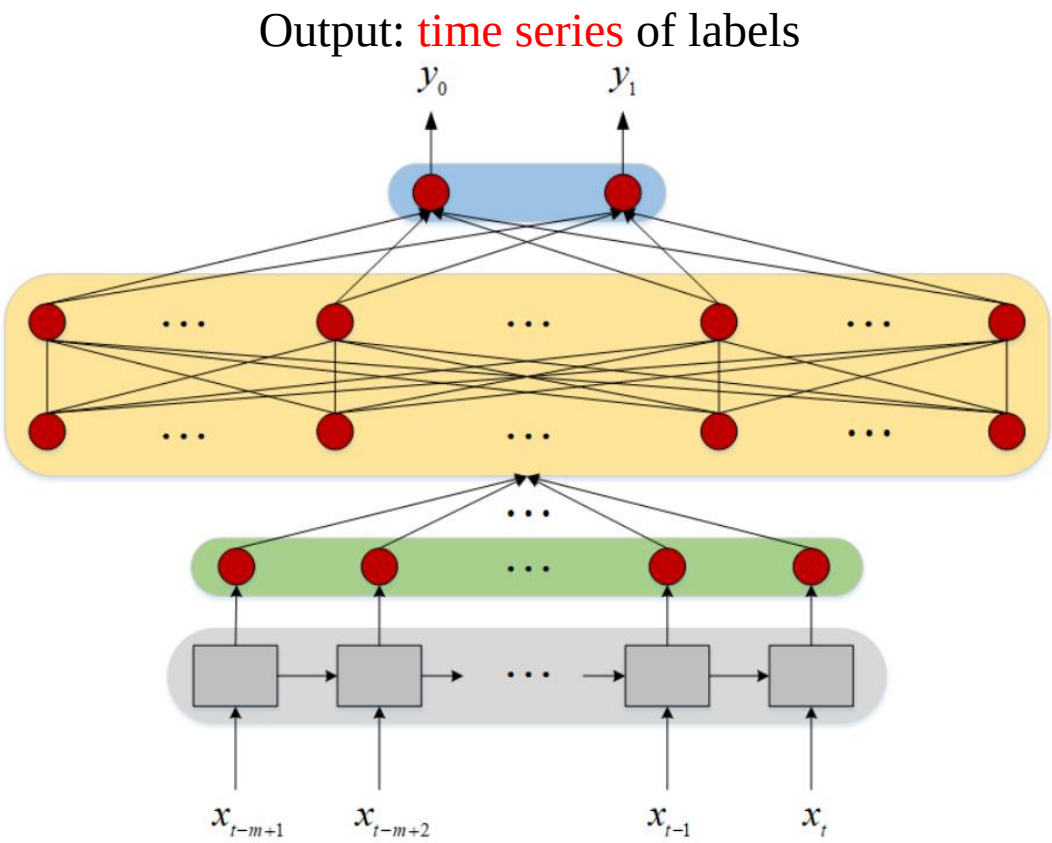
- Probably, this is a more realistic visualization of these clusters
- Clusters overlap with each other. A variety of research attempts can be associated with several clusters.
- Let us talk more about the selected research examples in each cluster.

Cluster 1: Enhancement of Space Weather prediction capabilities (forecasting and nowcasting)

- We are still trying to understand the triggers and drivers of solar transient and longer-term activity and the related terrestrial impacts.
- Availability of large observational data volumes and (often) a very clearly-defined task allows us to formulate precise classification or regression tasks for machine learning.
- Probably, the largest category in Heliophysics where machine learning is applied so far: prediction of solar flares, CMEs, SEPs, ionospheric scintillations, geomagnetic indexes, sunspot numbers, solar irradiance, etc.
- Together with a significant research component, this category is expected to have a strong impact on operational forecasting of Space Weather

Example 1.1: Deep learning as an emerging tool for flare prediction

- Convolutional Neural Network-based methods[1] using magnetograms as input images give results comparable to previous feature-based approaches.
- Recently several attempts utilizing Recurrent Neural Network[2] architectures have appeared. Although feature-based, the presented methods consider time series instead of static descriptors.

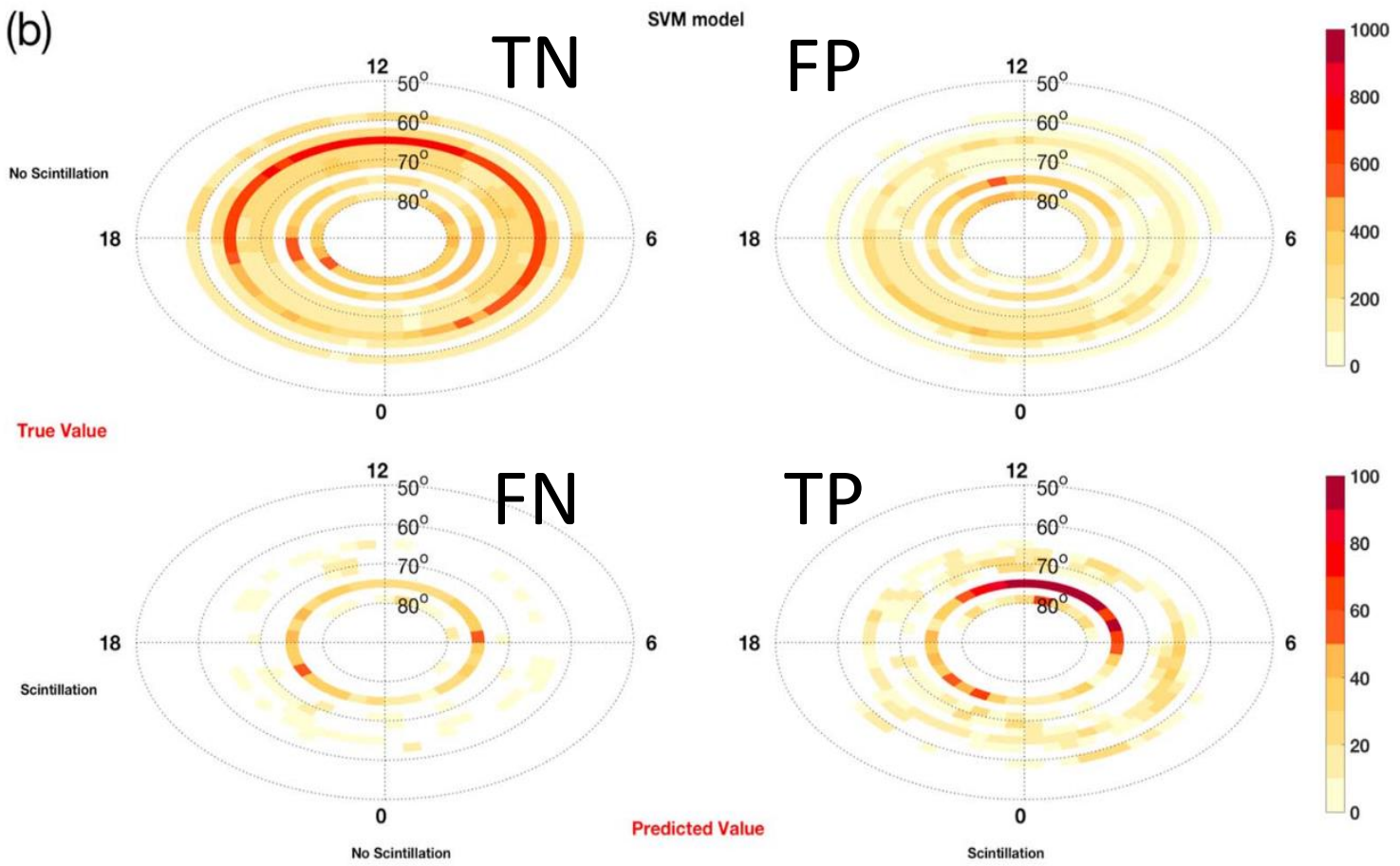


Example of the LSTM network architecture.
Credits: Liu et al. 2019

[1] Huang et al. 2018, Jonas et al. 2018
[2] Long-Short Term Memory Network (LSTM), Liu et al. 2019, Sun et al. 2019

Example 1.2: Predicting ionospheric scintillations

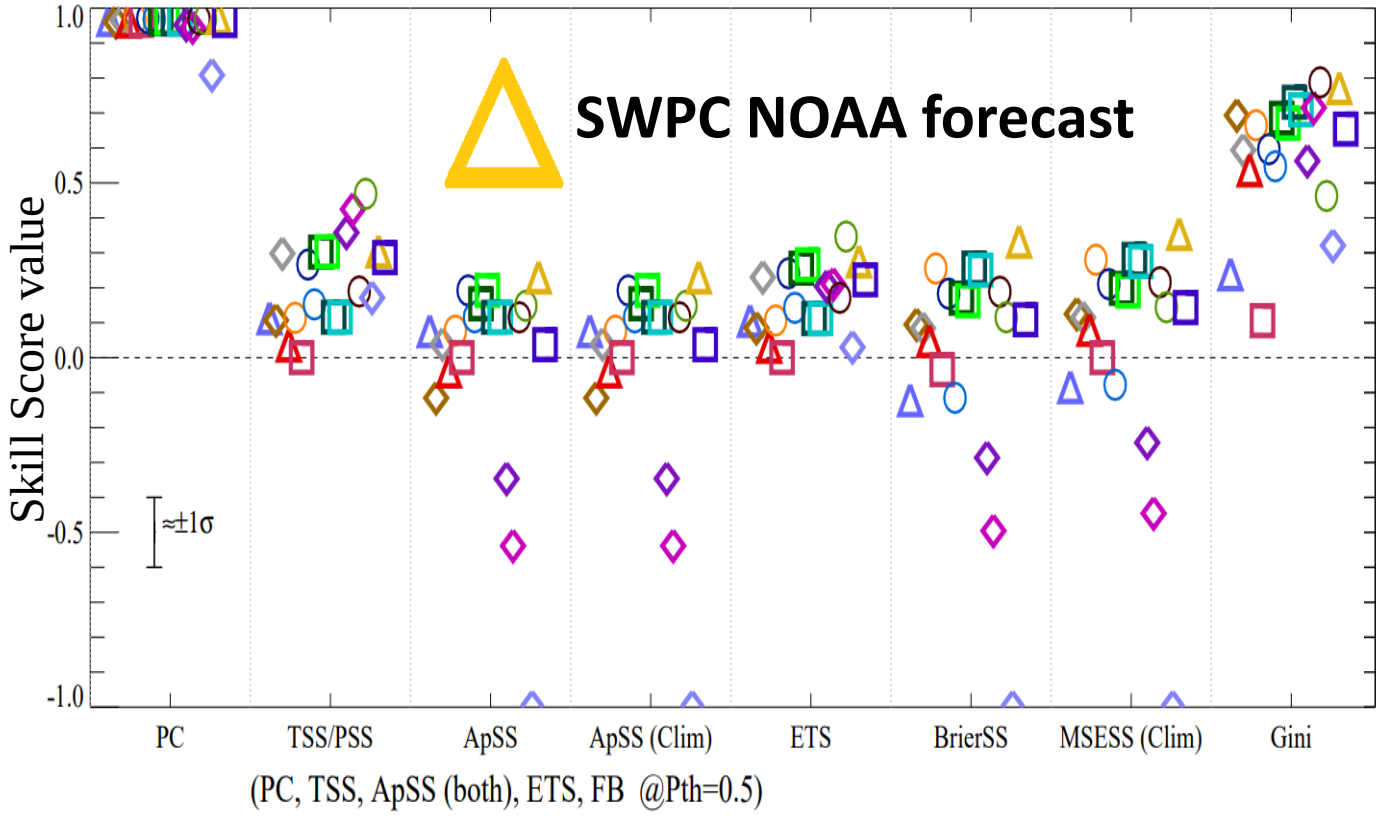
- McGranaghan et al. (2018) utilized solar wind data, geomagnetic activity, and particle precipitation data, as well as ionospheric data, to predict high-latitude ionospheric phase scintillations
- The authors employed a Support Vector Machine (SVM) classifier as a machine learning algorithm and performed careful analysis of the model performance with respect to lead time variations and definition of the scintillation.



Visualized confusion matrix. Credits: McGranaghan et al. (2018)

Challenges while moving from research to operations

- Prediction depends on the data set, on the definition of train/test subsets, on the metrics targeted to maximize, etc.:
  - It is almost impossible to cross-compare the performance of different methods.
  - It is very hard to move any developed method to operational status.
- Possible solution: apply prediction algorithms from different efforts under the same conditions (right: Leka et al. 2019). Another solution: match temporal and spatial scales of operational data sets (Sadykov et al. AGU 2018).
- Research to Operations pipeline requires validation of the research attempts on longer-term data sets.

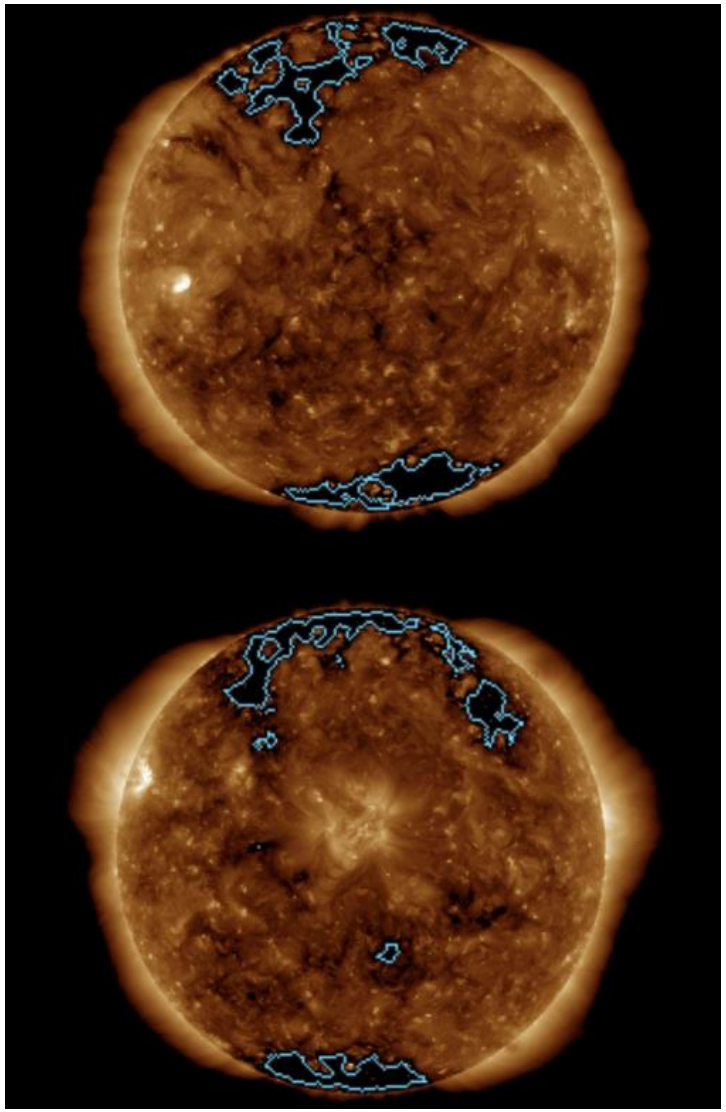


Comparison of different forecasting methods. Each symbol represents an operational method, either from space weather warning centers or research facilities. Credits: Leka et al. 2019.

Cluster 2: Object recognition and classification in observational and modeling data

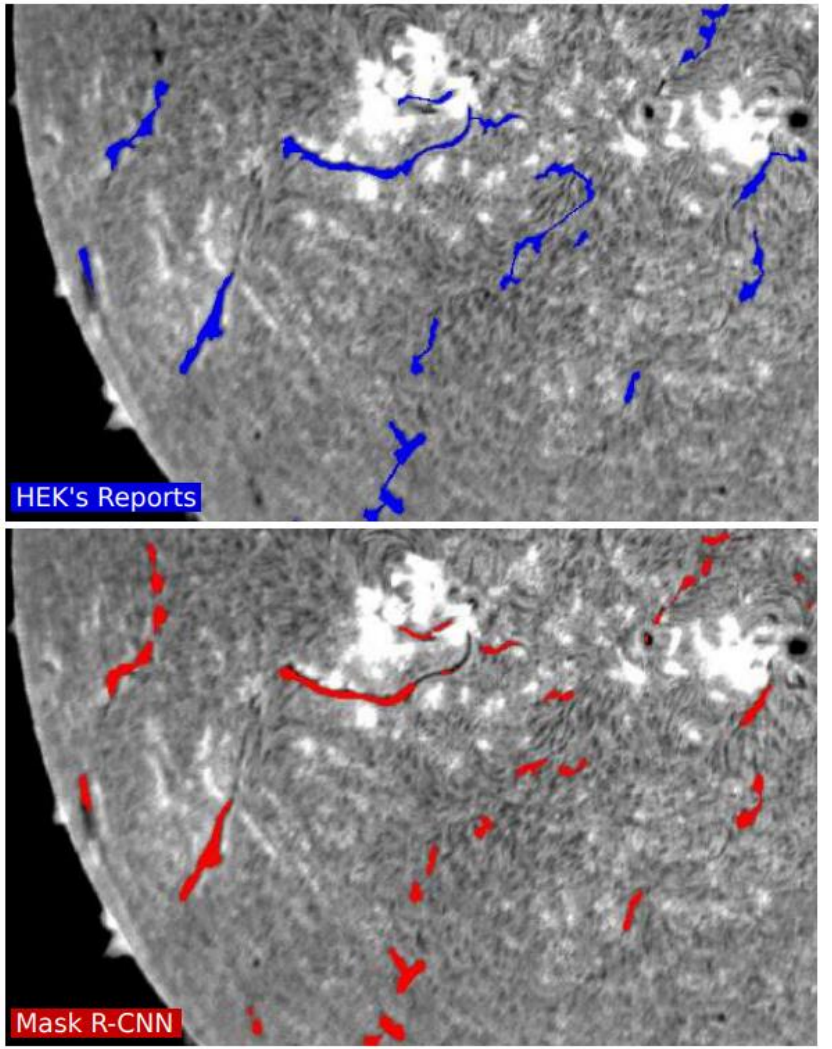
- Although this task is often performed in support of other research directions (Space Weather forecasting, derivation of features of recognized objects, statistical studies, etc.), it can be considered as a self-contained task.
- Object recognition is often used to replace (and enhance) human-based detection.
- Often represents a pure classification task in machine learning

Examples 2.1: Employing deep learning for automatic detection of coronal holes and filaments



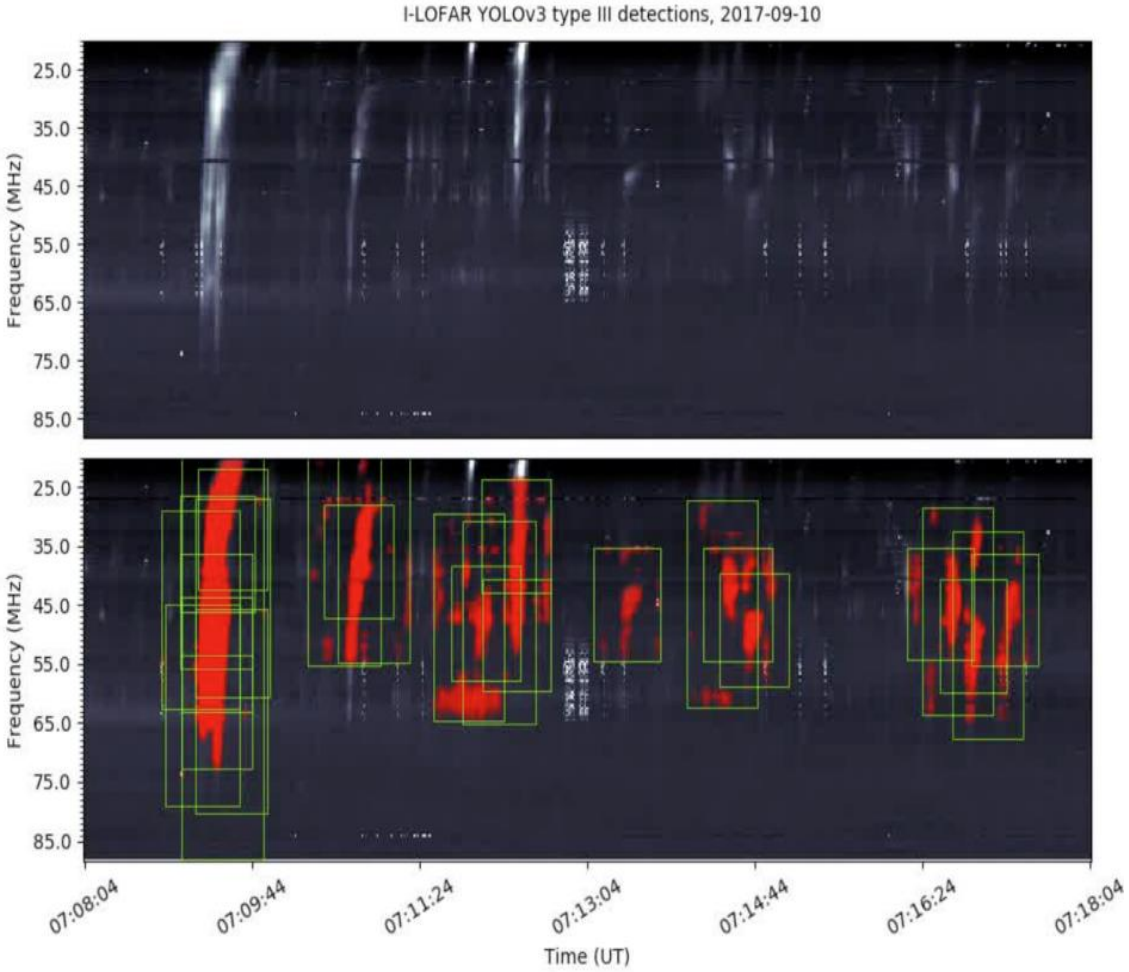
Left: Examples of coronal hole detection by U-Net. Credits: Illarionov and Tlatov, ML-Helio 2019.

- Statement: deep learning can be successfully applied to object detection in solar images in the presence of accurate labeling and sufficient size of the training set.



Right: Example of filament segmentation as reported to HEK (top) and detected by Mask R-CNN (bottom). Credits: Ahmadzadeh et al. 2019.

Other examples of event detection

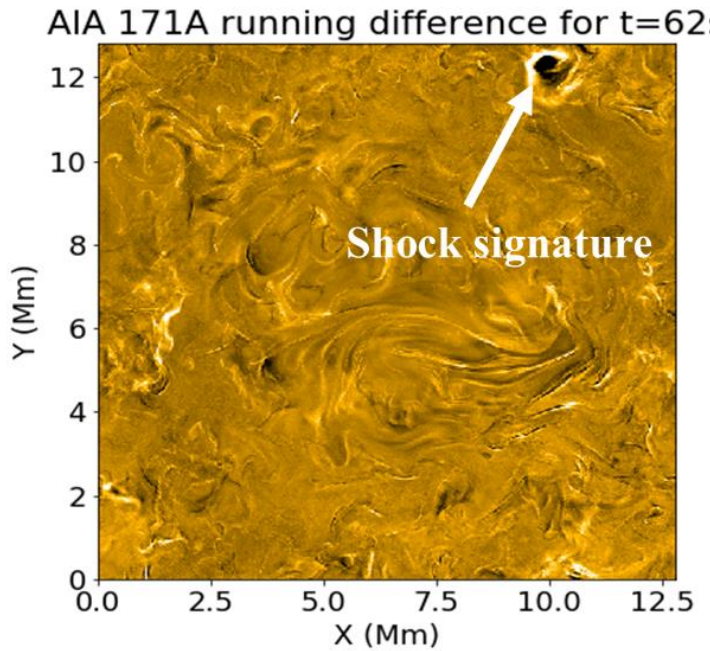


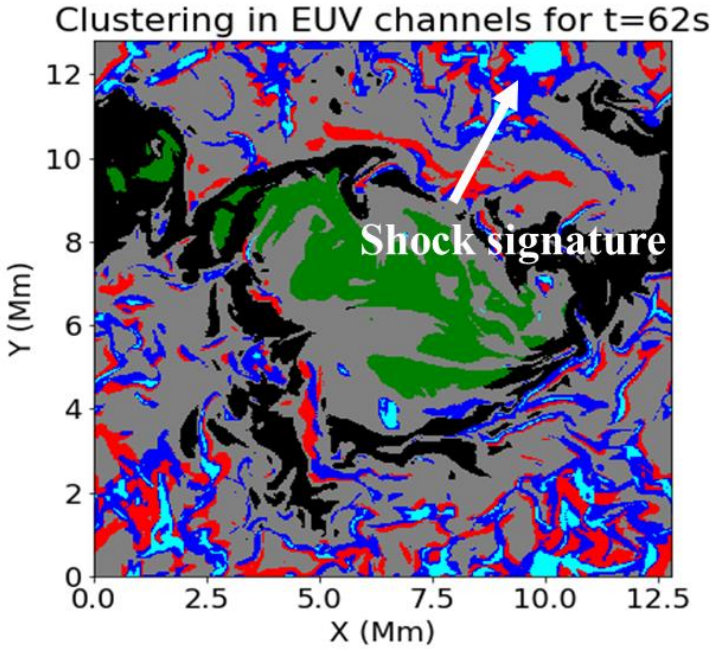
Top: an example of automatic detection of radio bursts using the YOLO deep learning algorithm. Credits: Carley et al. (ML-Helio 2019).

Right: illustration of recognition of shockwaves in quiet Sun simulations performed with the StellarBox RMHD code (Wray et al. 2015, 2018)

Shockwaves are prominent in running-difference images of synthesized SDO/AIA emission (top panel). After clustering is applied for synthesized SDO/AIA emission, the shocks and tend to be in one cluster (bottom panel)

Credits: Sadykov et al. (2019).



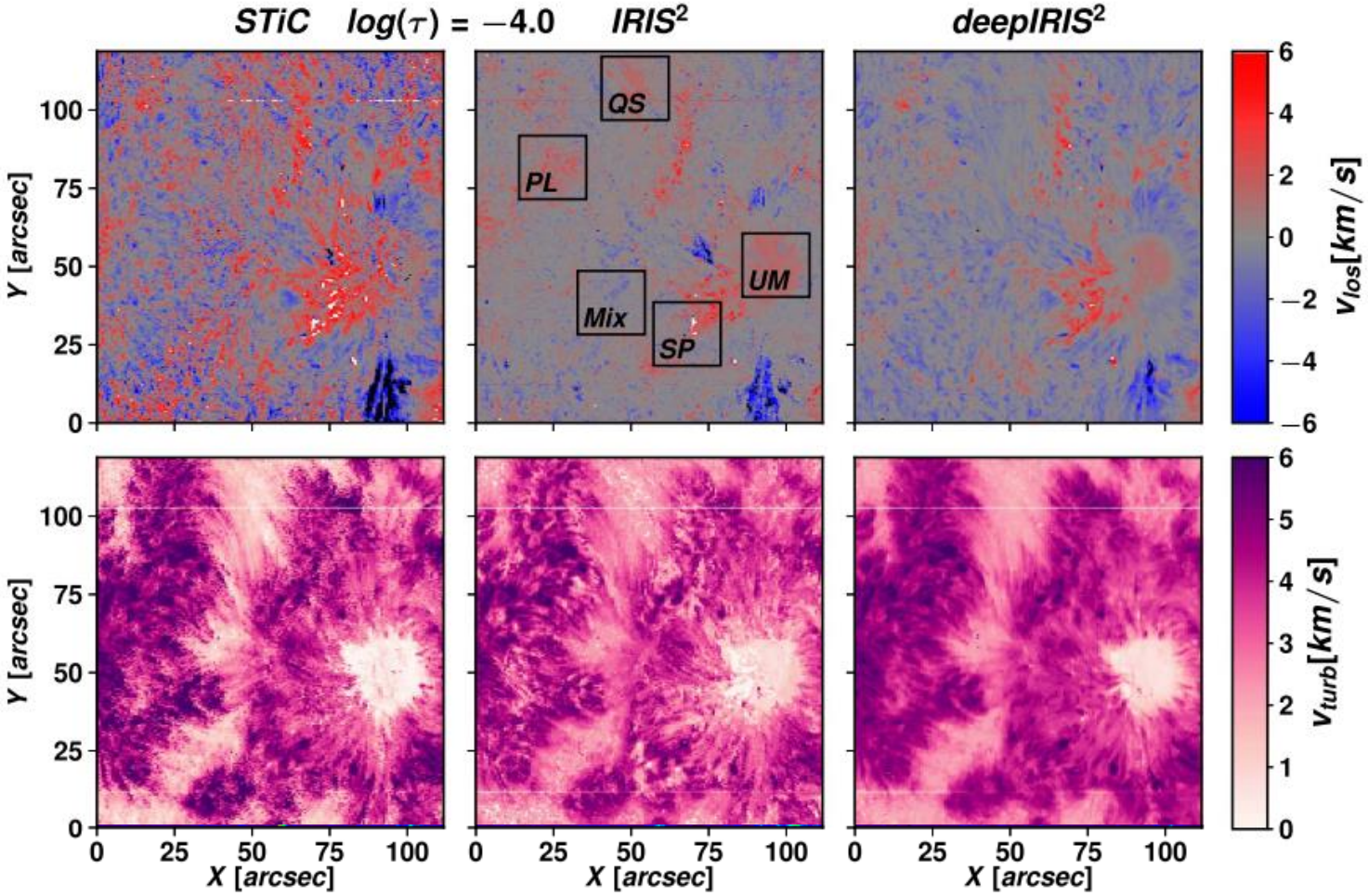


Cluster 3: Simplification of physics models and approximation of non-linear relations by ML analogs

- Many problems of physics are non-linear, non-local, ill-posed in nature, and are very expensive to solve computationally .
- Sometimes even finding the appropriate physical description of the relations between data sets (for example, photospheric magnetic fields of active regions and related EUV emission of coronal loops) is complicated.
- In the presence of enough training data, machine learning can help us either to replace some portions of a model by its faster ML-driven analog, or to fully replace the original model.
- Further analysis of ML-driven models can potentially enhance our understanding of the physical processes.

Example 3.1: Fast inversion of Mg II line profiles using Deep Learning

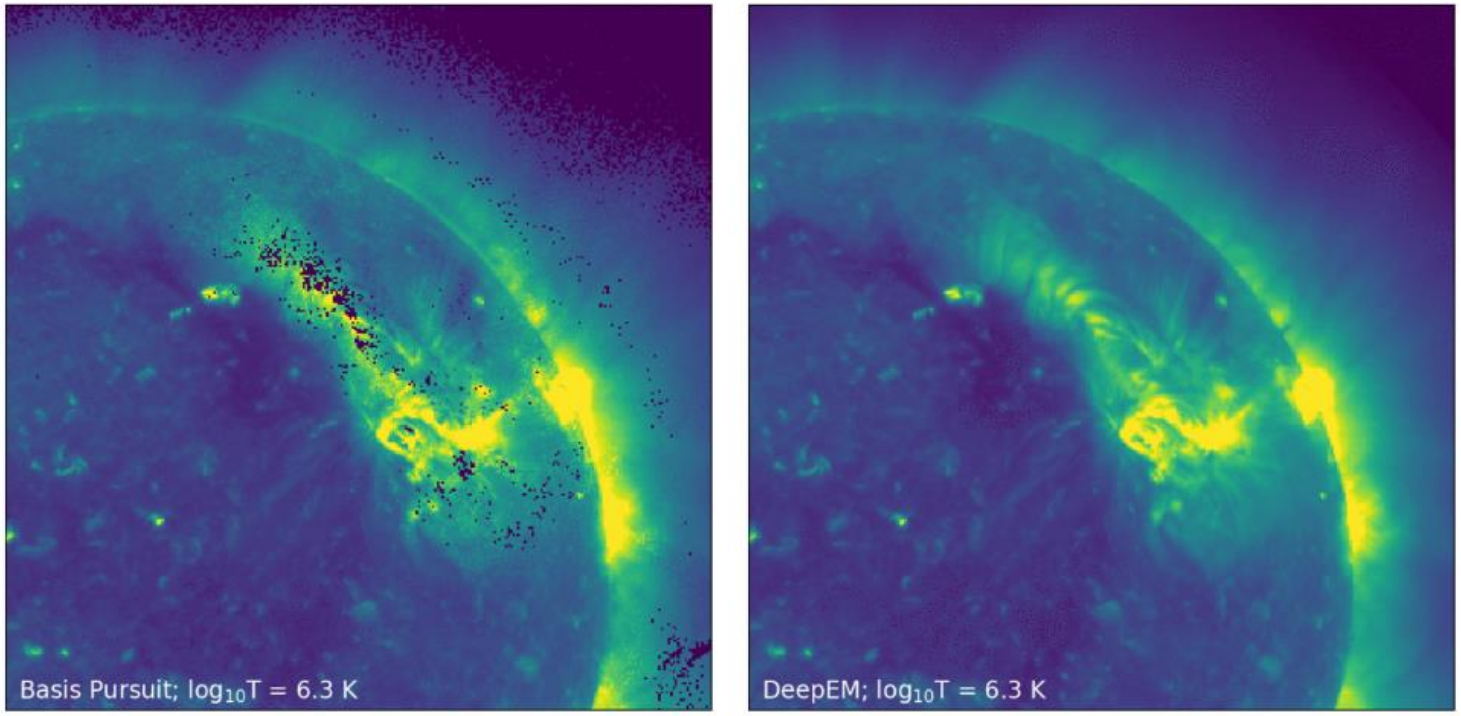
- Sainz Dalda et al. (2019) developed a faster model for non-LTE inversion of the Mg II lines for various conditions of the solar atmosphere (including flaring atmospheres):
  - The k-Means clustering technique was applied for identification of typical line profiles.
  - Cluster centers were inverted with STiC code (a physics model).
  - The neural network was trained on the inverted cluster centers to "interpolate" the solutions of the inverse problem.



Reconstruction of atmospheric parameters from Mg II lines.
Credits: Sainz Dalda et al. (2019)

Example 3.2: Fast inversion of EUV emission using Deep Learning

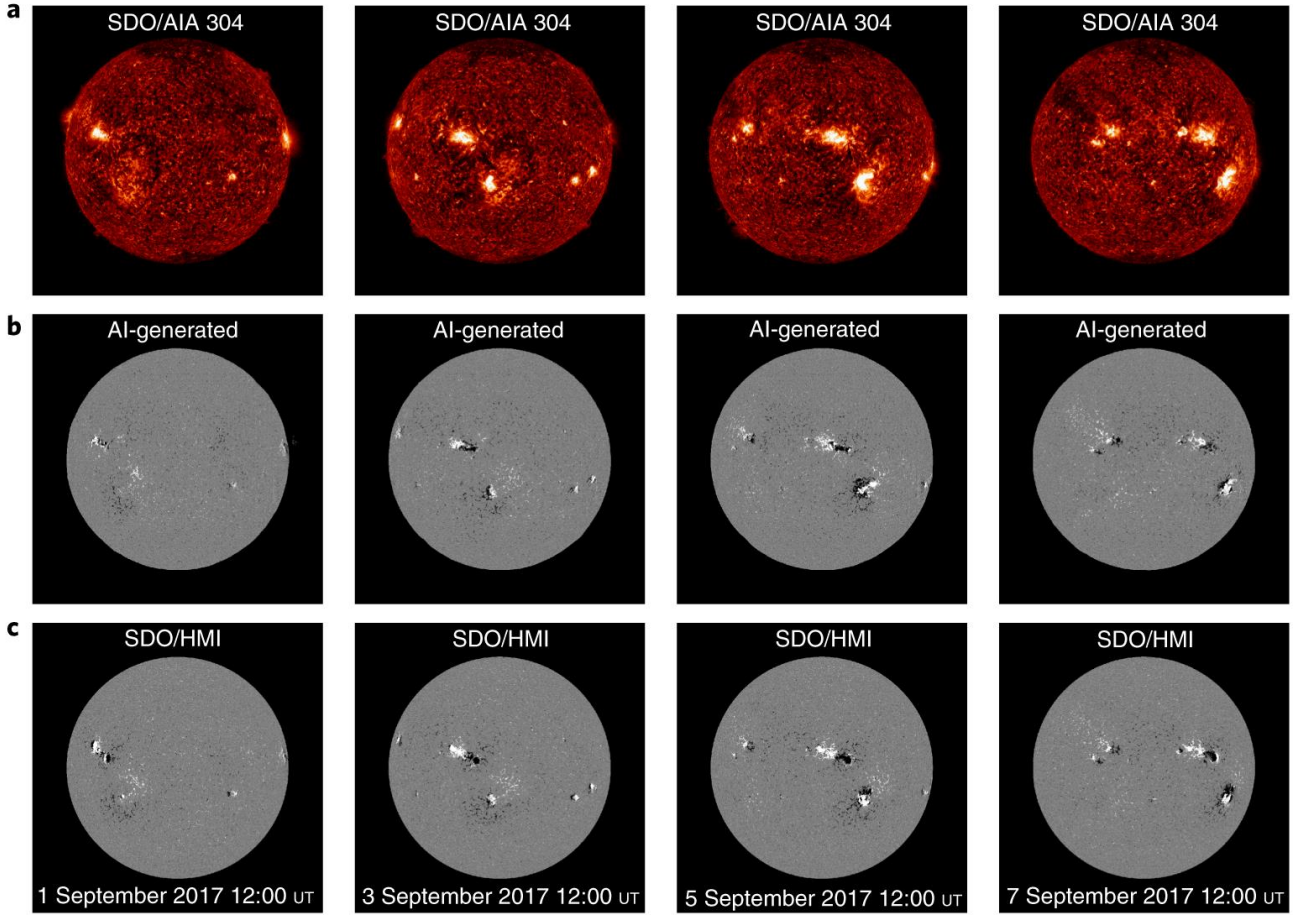
- Wright et al. (2019) employed deep learning to replicate the differential emission measure (DEM) derivation from SDO/AIA images.
  - The training set represented a set of calibrated SDO/AIA observations and physics-based DEM solutions
  - The neural network was trained on the inverted data set to replicate the physical model.
  - This attempt was followed by generation of synthetic MEGS-A data from SDO/AIA observations after the instrument failed (FDL 2018).



Comparison of physics-based and deep learning-based solutions. Credits: Wright et al. 2019, in HelioML ebook.

Example 3.3: Far-side magnetic field maps from STEREO/EUVI and far-side helioseismology

- Kim et al. (2019) developed a model to reconstruct magnetic field maps from EUV images (SDO/AIA and STEREO/EUVI 304 A observations).
- STEREO/EUVI had periods of time when it observed the far side of the Sun. This opened the possibility of reconstructing the far-side magnetic field maps.
- Chen et al. (AGU 2019) used far-side helioseismology to deduce magnetic flux maps based on previous EUV-HMI pairing.



Comparison of SDO/HMI magnetic field maps with generated from SDO/AIA 304A by AI. Credits: Kim et al. (2019)

Cluster 4: Exploration and discovery of (large highly-dimensional) data volumes

- Precursor: growing observational and modeling capabilities result in significantly larger data complexity, rates, and volumes with respect to what was handled previously. It is impossible to explore these volumes without projecting them to a simpler, more compact space
- Data clustering has already been applied to a variety of problems (selection of training samples for deep learning, classification of spectroscopic line profiles and EUV emission). Attempts to employ representation learning were shown at FDL 2019.
- Personal opinion: this direction is under-explored and has a strong potential for discovery.

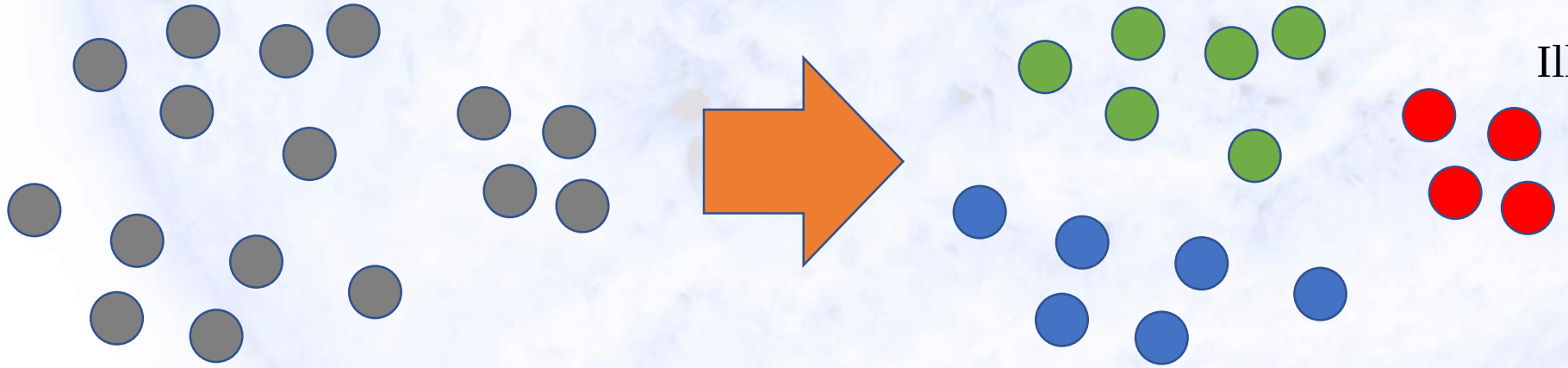
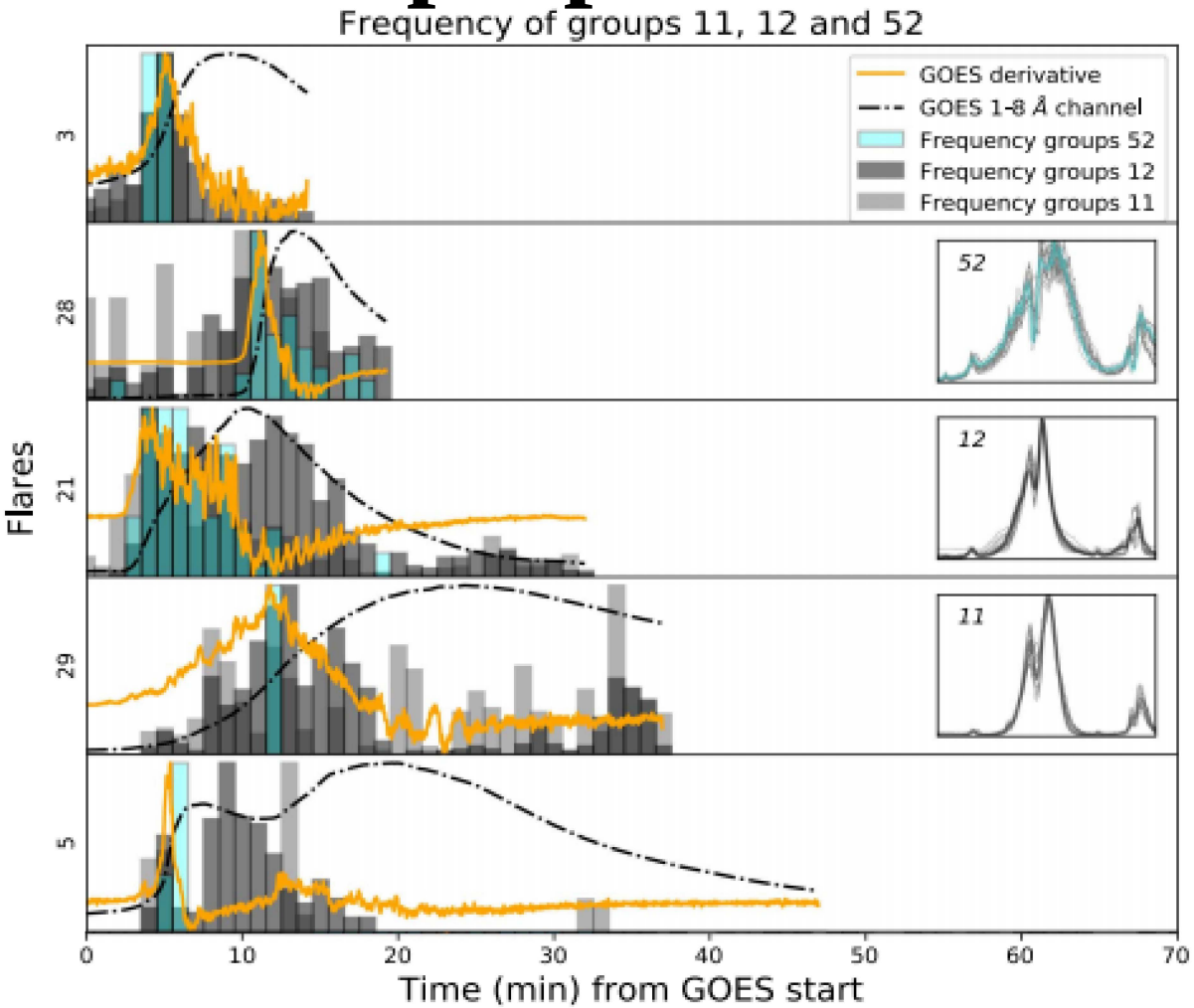


Illustration of clustering of data points

Example 4.1: Classification of spectroscopic line profiles and correlation with flare properties

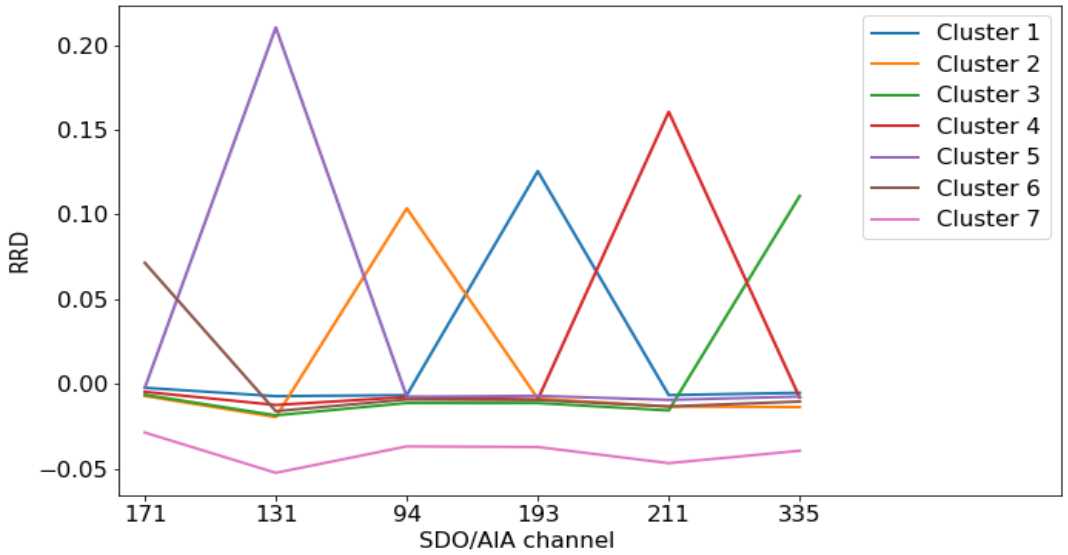
- Panos et al. (2018) used a k-Means clustering algorithm to recognize typical shapes of Mg II line profiles observed during 33 solar flares (hundreds of thousands of data samples).
- The authors found correlations between certain profile shapes and the GOES Soft X-Ray (SXR) time derivatives at the front of fast-moving flare ribbons.
- Panos and Kleint (2019) recently investigated the possibility of predicting solar flares based on the appearance of certain types of profiles. The same approach was introduced by Woods et al. (AGU 2019)



Appearance of cyan line profiles (cyan histogram) correlates with the GOES SXR time derivative (orange curve). Credits: Panos et al. (2018)

Example 4.2: Clustering of Relative Running Differences (RRD) of SDO/AIA EUV emissions

- We utilized 10-minute observations of the quiet Sun at the disk center by SDO/AIA. Observations in different channels are aligned.
- We define RRDs as RRD(t) = I(t)/I(t-1) – 1. RRDs do not show dependence on global structures
- Seven clusters are selected using a k-Means clustering algorithm. These correspond to seven "quantum states" of SDO/AIA EUV emission.

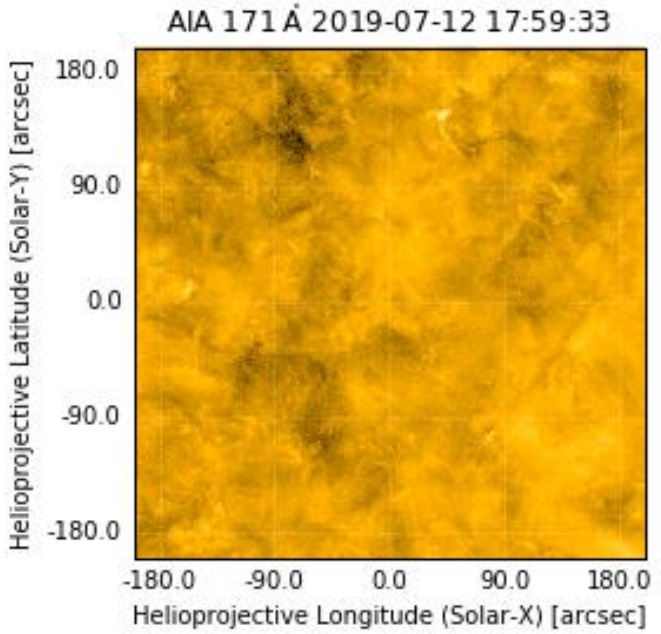


- Idea: Now one can consider evolution of the discrete number (quantum state) instead of a six-dimensional vector. It makes the problem much easier. Preliminary conclusions:
  - The behavior of quantum states is Markovian.
  - The system fully "forgets" the quantum state where it was previously in about 48 seconds.

Left: Clusters of SDO/AIA relative running differences of the quiet Sun.

Right: Illustration of the quiet Sun domain used in the study.

Credits: Sadykov et al. (AGU 2019)

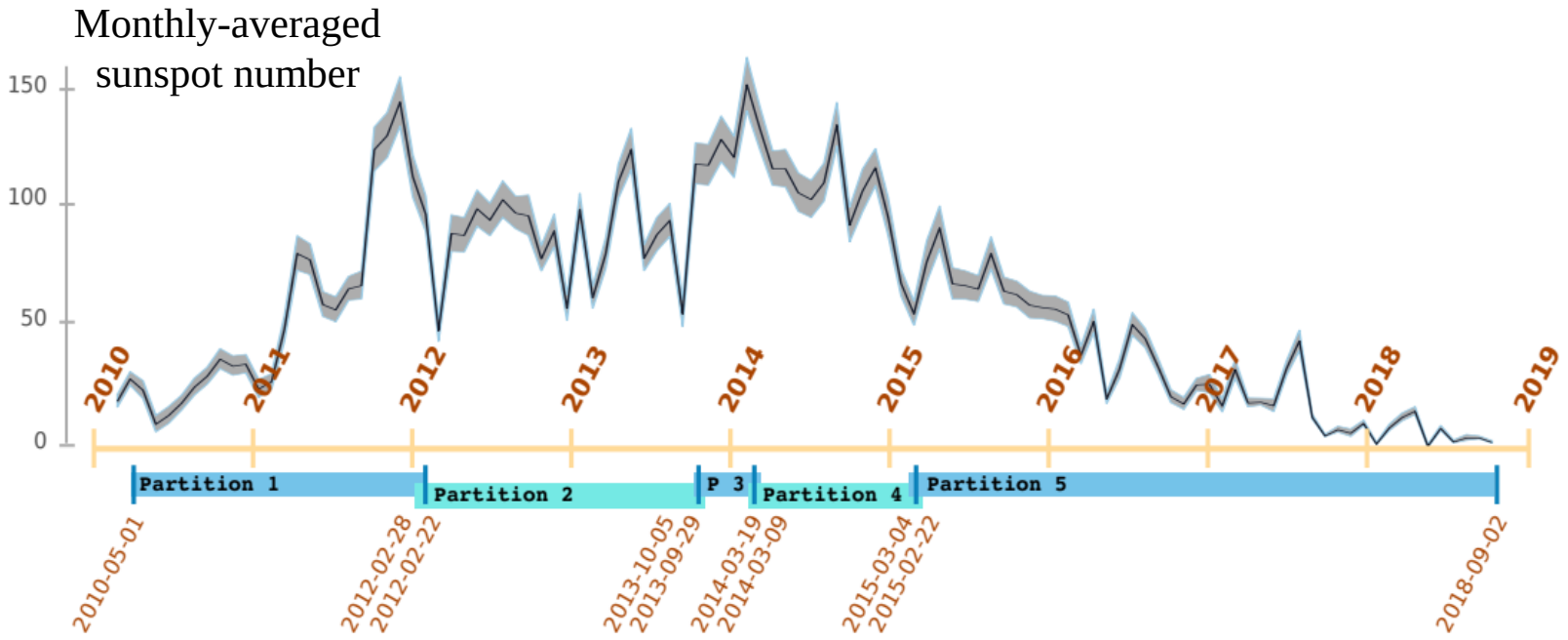


Cluster 5: Creation of homogeneous ML-ready datasets and related quality standards

- Data preparation is probably the most important phase in any machine learning attempt. If the data is not prepared and cleaned properly, the results will be unreliable (“Garbage in – garbage out”).
- The data preparation phase is usually very time- and effort-consuming. ML-ready datasets are highly valuable for the community.
- The field spans beyond the “standard” procedures to prepare the data (search for outliers and corrupted data, calibration, and instrument degradation corrections).

Example 5.1: Space Weather ANalytics for Solar Flares (SWAN-SF)

- Currently it is almost impossible to compare the scores from a variety of flare prediction efforts (as previously discussed).

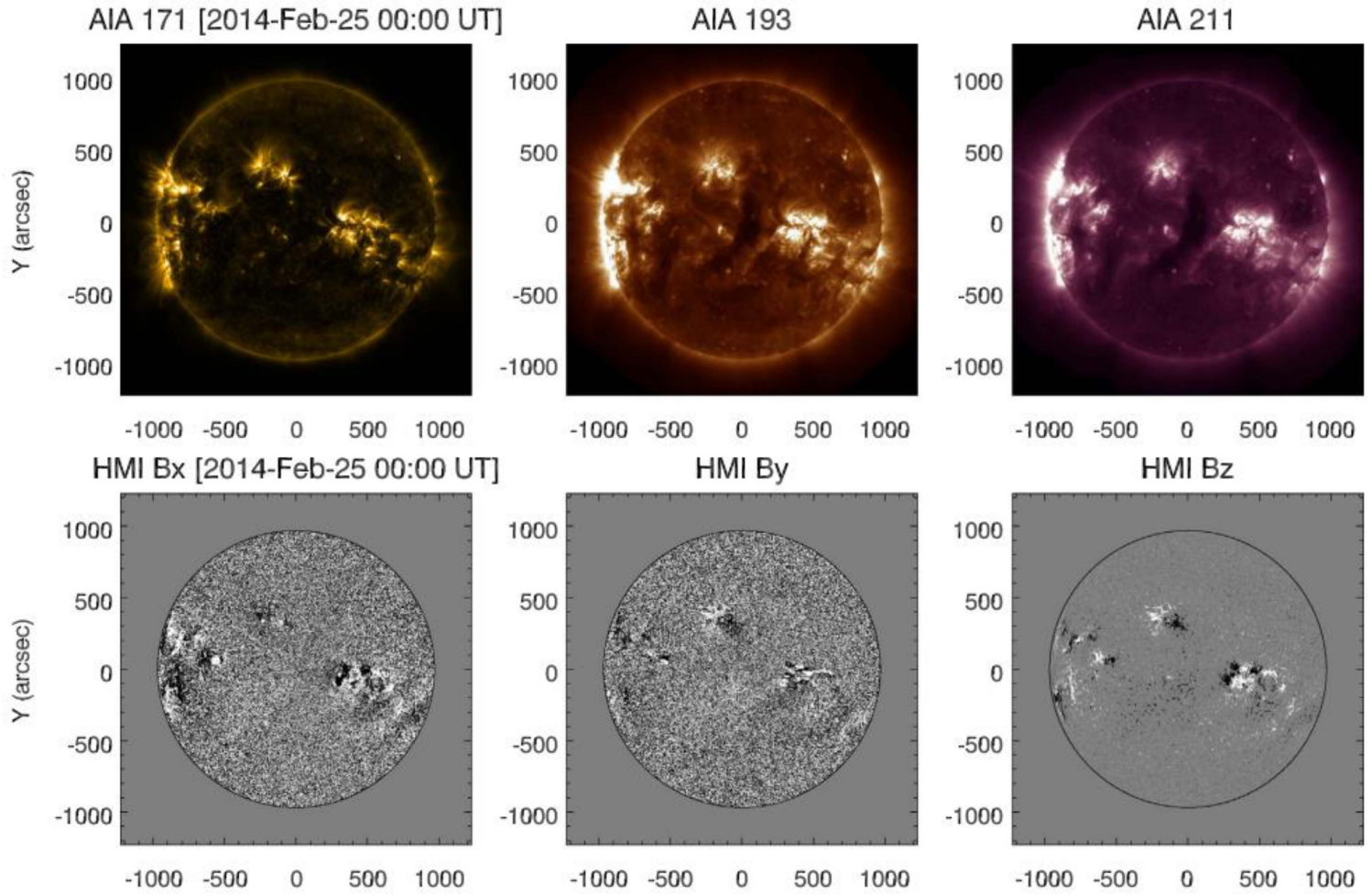


Data partitioning in a SWAN-SF data set. Credits: Ahmadzadeh et al. 2019

Solution: build open-accessible data sets for flare prediction purposes (SWAN-SF). The data set is properly separated and contains time series, which is important for some deep learning (LSTM) algorithms and other methods in applications to flare forecasting.

Example 5.2: Machine learning dataset prepared from the SDO/AIA mission

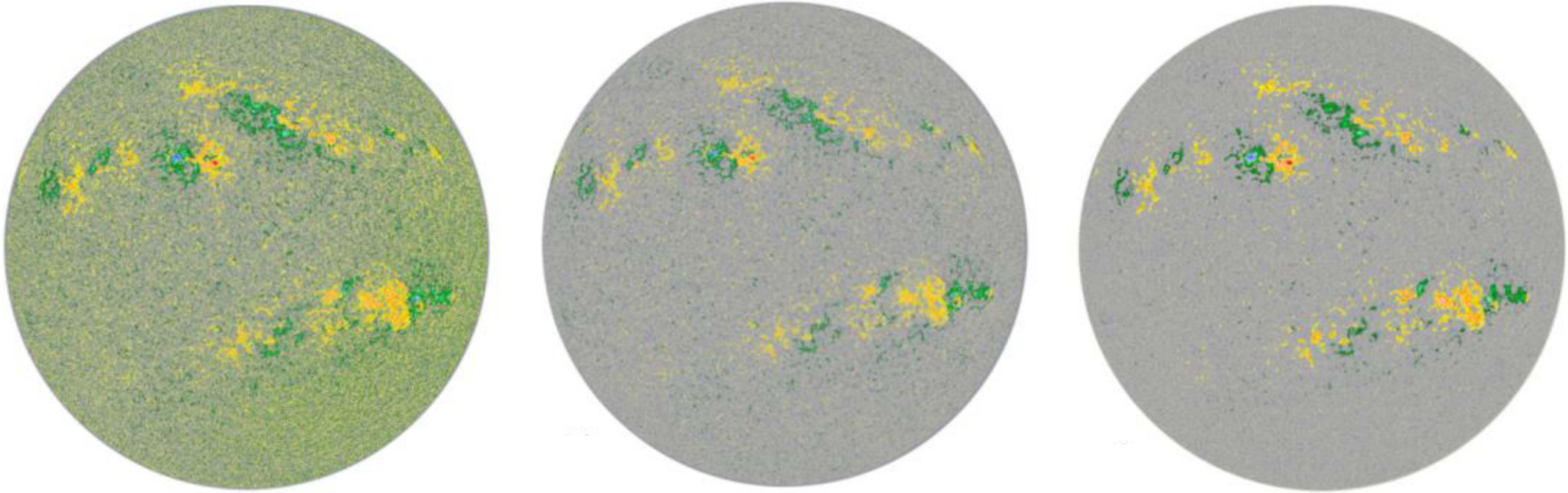
- The dataset contains SDO AIA images and HMI magnetograms with 2 min and 12 min cadence correspondingly, with a size of 512x512 pixels. SDO/EVE data is delivered every 10 seconds.
- The images are synchronized to keep the same solar disk size and rotation phase, and the SDO/AIA and SDO/EVE data are corrected for degradation of the instruments.
- The data set delivered a variety of results in FDL 2018&2019.



Credits: Galvez et al. 2019

Example 5.3: Super-resolution homogeneous magnetic field maps

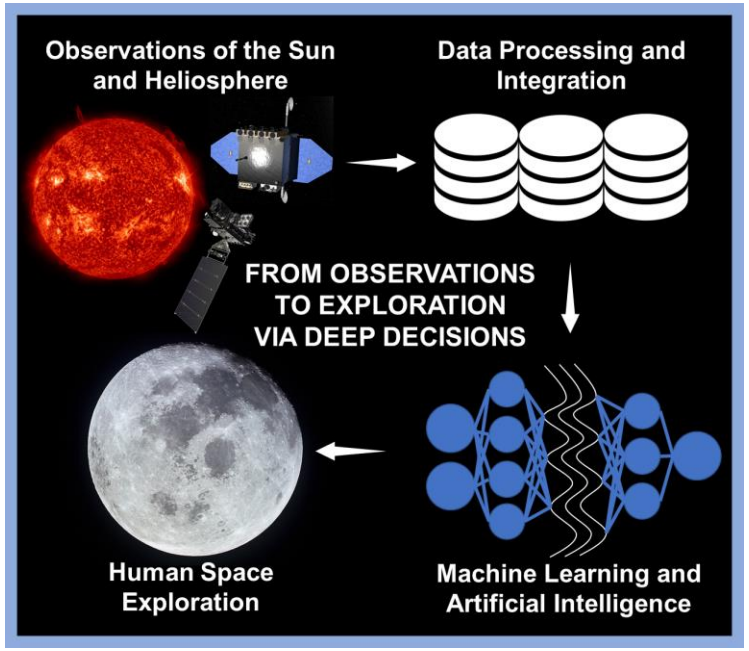
- An attempt to super-resolve magnetic field maps and create a homogeneous dataset using deep learning architectures spanning from Wilcox solar observatory data to Hinode/SOT maps.
- The work started in FDL 2019.

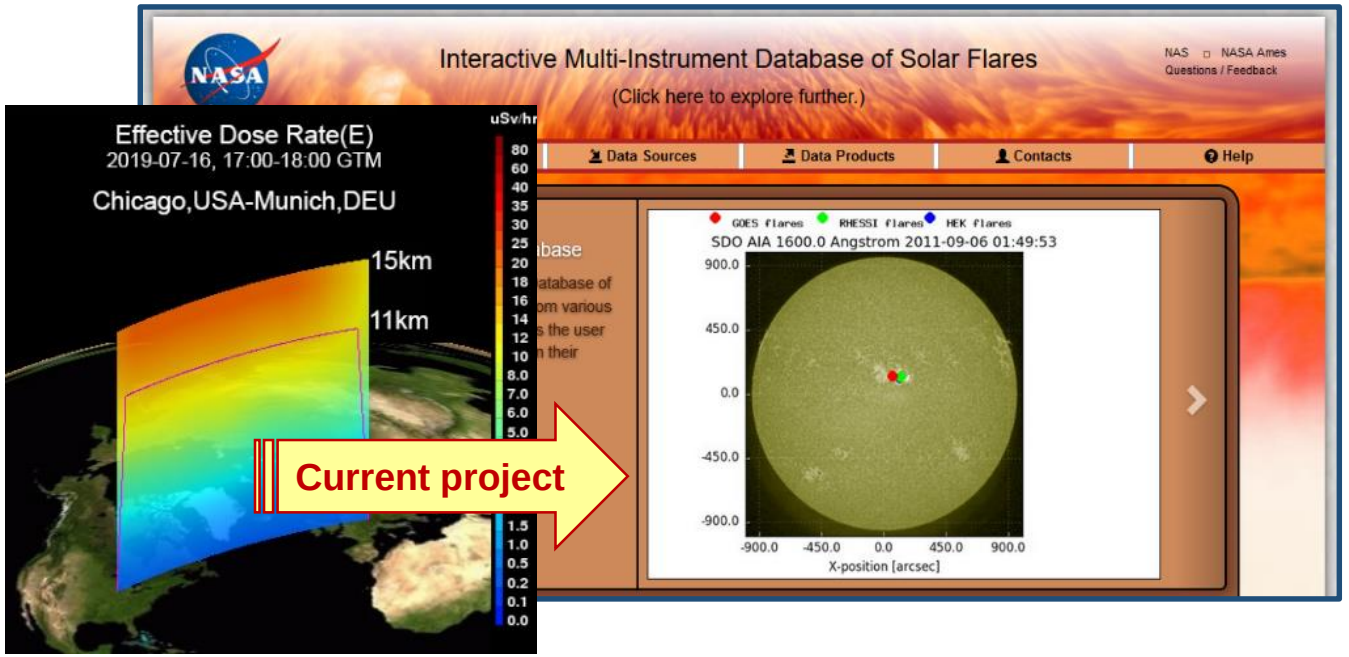


Left: SOHO/MDI image. Center: corresponding SDO/AIA image. Right: super-resolved SOHO/MDI image using physics-based loss function. Credits: Jungbluth et al. 2019

Ongoing projects related to creation of ML-ready datasets

- We all wish to have reproducible and traceable results and open-access high-quality data in our community. The best practice is to start doing it for our projects.
- Two projects with involvement of NASA Ames / BAERI resulted in such data:
  - “Machine Learning Tools for Predicting Solar Energetic Particle Hazards” (NASA ESI)
  - “Interactive Database of Atmospheric Radiation Dose Rate” (NASA)





Conclusion

- Machine learning has many applications in Heliophysics. It has already moved beyond Space Weather forecasting and far beyond flare prediction.
- Research attempts are growing in number and receiving strong attention from the community.
- The research attempts can be subdivided into several categories based on the problems which they address.
- It is critical to continue development of the field:
  - We should address very important challenges which are impossible to solve if no machine learning is applied.
  - We must gain more understanding of machine learning, its limitations and pitfalls.