Machine Learning in Heliophysics

Monica Bobra
Stanford University
mbobra@stanford.edu

NASA SDO / AIA

What is the state of machine learning in heliophysics today?

What does a machine learning data set look like?

What types of observations produce the best data sets for machine learning?

What infrastructure do we need to get the most out of large data sets?

What is the state of machine learning in heliophysics today?

- NASA, NSF, and NOAA opened solicitations for machine learning research
- Conference series called *Machine Learning in Heliophysics*
- Books
  - Machine Learning Techniques for Space Weather (Camporeale et al., 2018)
  - Statistics, Data Mining, and Machine Learning for Heliophysics (Bobra and Mason, 2020)
- Research accelerator called Frontier Development Laboratory
- Interdisciplinary collaborations with computer scientists and statisticians
- An increased emphasis in the machine learning community on science applications

What is the state of machine learning in heliophysics today?

Two main reasons machine learning has become so popular:

1. Algorithms can unearth statistical relationships in large data sets that massively improve event detection and prediction
2. Scientific Python provides a rich and mature ecosystem of software packages
   - Machine learning, e.g. Pyro, statsmodels, scikit-learn
   - Parallel computing, e.g. Dask, Modin, Ray
   - Heliophysics, e.g. SunPy, SpacePy, HelioPy, PlasmaPy

Bingham et al. 2018, Seabold and Perktold, 2010, Pedregosa et al. 2011
Rocklin et al. 2015, Petersohn et al. 2020, Liaw et al. 2018
The SunPy Community et al. 2020, Morley et al. 2011, Stansby et al. 2020, PlasmaPy Community et al. 2018

What does a machine learning data set look like?

An example set of features that describe solar active regions:

- Current density [derived from SDO/HMI magnetic field maps]
- Total magnetic flux [derived from SDO/HMI magnetic field maps]
- Filament length [derived from SDO/AIA images of the chromosphere]
- Sigmoid tilt angle [derived from Hinode/XRT images of the corona]

What does a machine learning data set look like?

	current density (mA/m^2)	total magnetic flux (Mx)	filament length (Mm)	sigmoid tilt angle (^{o})	…	class label
t_1	0.7	2×10^{22}	11.0	86		Yes
t_2	-0.1	5×10^{21}	2.8	10		No
t_3	.2	8×10^{21}	3.1	-20		No
…						
t_n	0.8	8×10^{22}	14.3	-35		Yes

What does a machine learning data set look like?

	current density (mA/m^2)	total magnetic flux (Mx)	filament length (Mm)	sigmoid tilt angle (^{o})	…	class label
t_1	0.7	2×10^{22}	11.0	86		Yes
t_2	-0.1	5×10^{21}	2.8	10		No
t_3	.2	8×10^{21}	3.1	-20		No
…						
t_n	0.8	8×10^{22}	14.3	-35		Yes

What does a machine learning data set look like?

	current density (mA/m^2)	total magnetic flux (Mx)	filament length (Mm)	sigmoid tilt angle (^{o})	…	class label
t_1	0.7	2×10^{22}	11.0	86		Yes
t_2	-0.1	5×10^{21}	2.8	10		No
t_3	.2	8×10^{21}	3.1	-20		No
…						
t_n	0.8	8×10^{22}	14.3	-35		Yes

What types of observations produce the best data sets for machine learning?

1. **Continuous observations**
   - Regular cadence
   - Few gaps or missing data
2. **Consistent observations**
   - Sit-and-stare mode
3. **Observations that span a long period of time**
   - To obtain a sufficiently large sample of flaring active regions
   - To understand how flaring activity depends on solar cycle

What types of observations produce the best data sets for machine learning?

Current data sets that satisfy these three characteristics:

- SDO to predict solar flares
- GNSS to predict ionospheric scintillation
- ACE to predict properties of the solar wind

How well do these predictions perform compared with traditional methods?

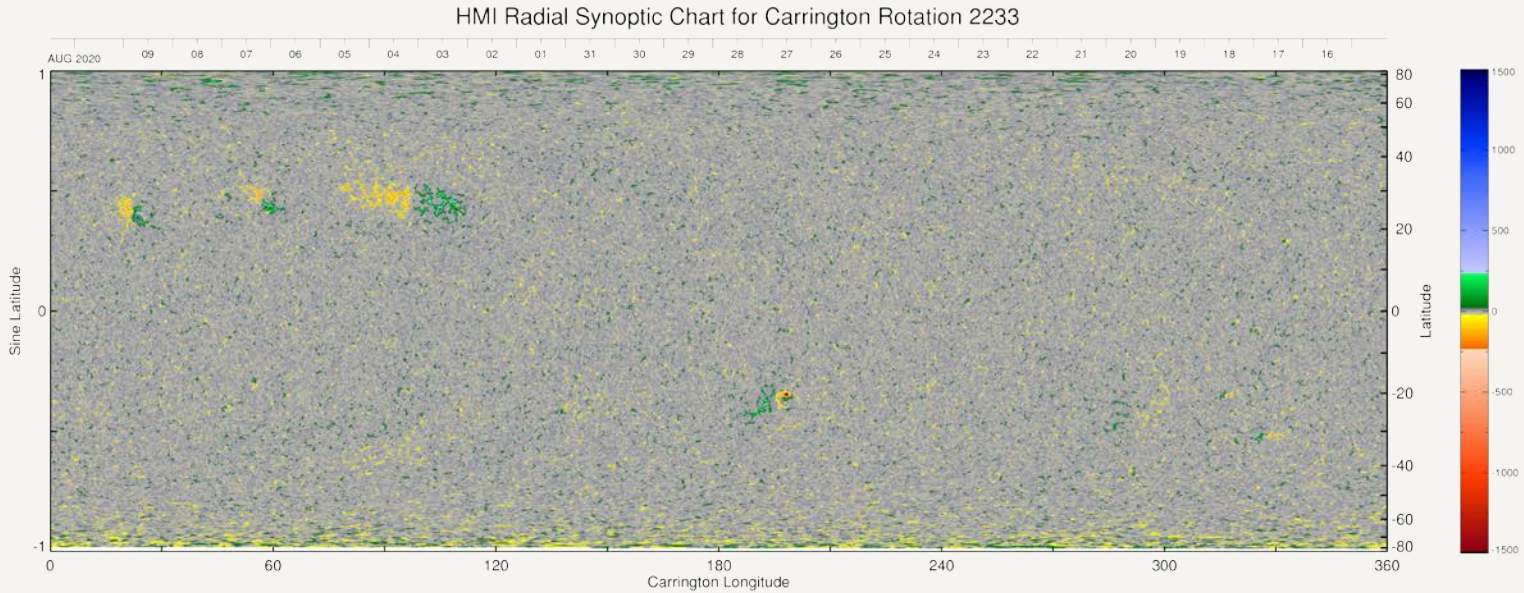
- Equally well or better than traditional methods
- Research does not use consistent validation metrics and data sets
- Operational validation metrics and data sets are not all publicly available

e.g. Bobra et al. 2015, Bobra et al. 2016, Jonas et al. 2018
e.g. McGranaghan et al. 2018
e.g. Heidrich-Meisner and Wimmer-Schweingruber 2018

What types of observations produce the best data sets for machine learning?

New data sets that satisfy these three characteristics:

1. Take surface magnetic field maps from different perspectives (L4 and L5)



What types of observations produce the best data sets for machine learning?

New data sets that satisfy these four characteristics:

1. Take surface magnetic field maps from different perspectives (L4 and L5)
2. Take simultaneous and identical measurements from multiple perspectives with a constellation of satellites
   - Terrestrial weather prediction benefits from having multiple copies of the same instrument, taking the same data, at the same times, from different vantage points

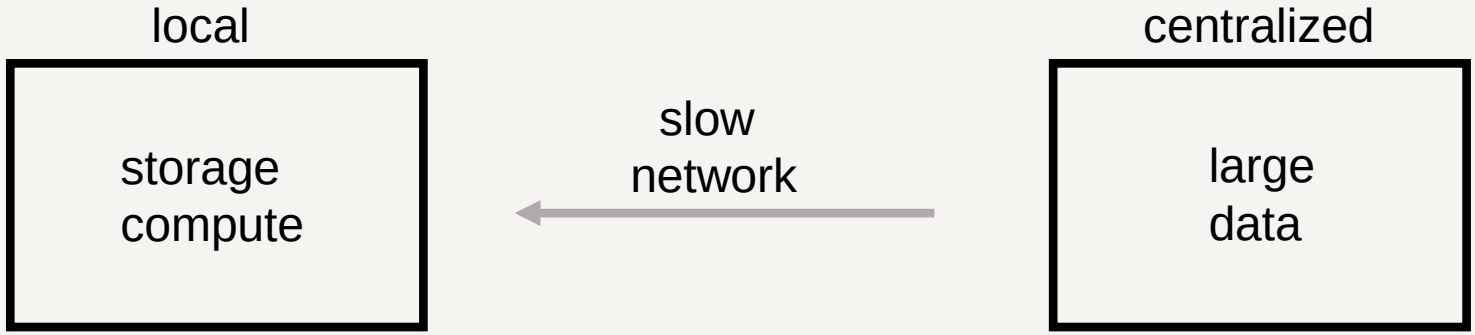
Observing the Sun from different perspectives and multiple perspectives is the most meaningful way to improve flare prediction using machine learning models.

What infrastructure do we need to get the most out of large data sets?

Solar and space physicists struggle to analyze large data sets.

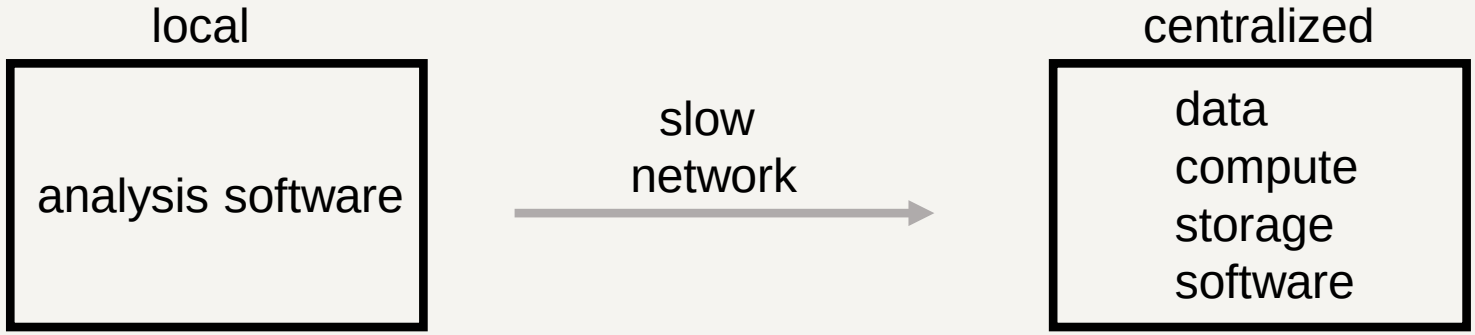
1. It is difficult to download large data sets
2. It is difficult to store large data sets
3. Some researchers lack adequate computational resources

What infrastructure do we need to get the most out of these data sets?



Bauer et al. 2019, Robinson et al. 2020

What infrastructure do we need to get the most out of these data sets?



- SDO science platform
- Vera Rubin Observatory science platform [funded by NSF]
- National Optical Astronomy Observatory science platform [funded by NSF]
- Pangeo science platform for earth science [funded by NSF, NASA, and NCAR]

Dubois-Felsmann et al. 2019, Fitzpatrick et al. 2014, Odaka et al. 2020
Barnes, Cheung, and Bobra 2019
Bauer et al. 2019, Robinson et al. 2020

Conclusion

The best way to improve space weather prediction with machine learning techniques:

1. Take long, continuous, and consistent data sets
   - Continuing long-term observations that span a solar cycle or more
   - Taking new, identical observations from multiple vantage points
2. Ensure easy access to these data
3. Ensure adequate computational resources to efficiently analyze these data