#### Machine Learning, Statistics, and Data Mining for Heliophysics

Monica Bobra<sup>1</sup> and James  $Mason^2$ 

<sup>1</sup>Stanford University <sup>2</sup>University of Colorado at Boulder

November 24, 2022

#### Abstract

We present a book entitled "Machine Learning, Statistics, and Data Mining for Heliophysics," an online and open source book available at helioml.org. This book includes a collection of interactive Jupyter notebooks, written in the programming language Python, that walks the reader through the process of applying machine learning, statistics, and data mining techniques on various kinds of solar and space physics data sets to reproduce published results. We consider this book to be a living document with frequent updates. Please contact us if you'd like to submit a chapter!

# We present an online book about machine learning,

## statistics, and data mining

for heliophysics.

Title Page What's in this book?				
		A closer look at Chapter 6	Click the interact button to run the code – no setup necessary!	
Title × +	Notebook	× +		
$\leftrightarrow \rightarrow$ C $\triangle$ S helioml.org	$\bigcirc$ $\bigcirc$ $\checkmark$ $\leftarrow$ $\rightarrow$ $\bigcirc$ $\bigcirc$ $\bigcirc$ $\bigcirc$ $\lor$	ot Secure   helioml.org/06/1/notebook	🔨 🔥 e 🕁 🚱 S 🏹 🍇 💿	🖋 🕐 🔚 🎦
A list of chapters, each of which replicate a published result	C CORLE SIDEBAR  Machine Learning, Statistics, and Data Mining for Heliophysics  Jy Monica Bobra and James Mason  Contributions by Andrés Asensio Ramos, Mark Cheung, Carlos José Diaz Baso, David Fouhey, Richard Galvez, Meng Jin, Andrés Muñoz-Jaramilio, Brandon Panos, Alexandre Szenicer, Rajat Thomas, and Paul Wright.  This is a book about machine learning, statistics, and data mining for heliophysics.  This book includes a collection of interactive Jupyter notebooks, written in Python, that explicitly shows the reader how to use machine learning, statistics, and data minining techniques on various kinds of heliophysics data sets to reproduce published results.  The contents of this book are licensed for free consumption under the following  Coretive Commons Attribution-NonCommercial-NoDerivatives 4.0 International  Two into 2018 - 01 https://doi.org/10.5281/zenodo.112825 0.20 2019-02-22 https://doi.org/10.5281/zenodo.112825 0.20 2019-02-22 https://doi.org/10.5281/zenodo.1412824  fyou'd like to atte the evolving book, instead of a specific version, use the following DU: https://doi.org/10.5281/zenodo.1412824   (	Title Acknowledgements Table of Contents 1. Whys and Whats 2. Predicting Coronal Mass Ejections 3. Enhancing SDO Images 4. Differential Emission Measurements	<ul> <li>COGGLE SIDEBAR</li> <li>Download</li> <li>Interact</li> <li>Sections of Jupyter notebook</li> </ul> Analyzing the behavior of a single spectral line using unsupervised learning by Brandon Panos <sup>1,2</sup> 1 University of Applied Sciences and Arts Northwestern Switzerland, Bahnhofstrasse 6, 5210 Windisch, Switzerland 2 University of Geneva, CUI-SIP, 1205 Geneva, Switzerland In this chapter we will explore how the spectral line associated with once ionized Magnesium (Mg II) behaves during a solar flare. To do this, we will make use of one of the most celebrated classical machine learning algorithms	ON THIS PAGE      I. INTRODUCTION     WHAT DOES OUR     INSTRUMENT SEE?     WHY IS THE MG II K-     LINE AN EXCELLENT     SOURCE OF     INFORMATION?      K-MEANS     ALGORITHM     FINDING THE     OPTIMAL NUMBER OF     GROUPS k     APPLYING K-MEANS     TO AN IRIS     OBSERVATION      S. APPLICATION OF K-     MEANS TO IRIS DATA     OUR OBSERVATIONS:     FINDING QUIET SUN     CENTROIDS AND     MAKING CENTROID     MASKS     NEW OBJECTIVE:     APPLICATION OF THE     CENTROIDS FOUND     IN THE QUIET SUN TO     A FLARING     OBSERVATION.      FINDING CENTROIDS     FINDING CENTROIDS     RELATED TO THE     FLARING SUN AND     MAKING NEW     CENTROID MASKS     IS THE CENTROID     MASK SOURCED     FROM ONE FLARE     CONSISTENT WITH A

### paper.

## Why did we write this book?

## To teach readers how to create replicable results.

"To help ensure the reproducibility of computational results, researchers should convey clear, specific, and complete information about any computational methods and data products that support their published results in order to enable other researchers to repeat the analysis, unless such information is restricted by non-public data policies. That information should include the data, study methods, and computational environment."

– National Academies of Sciences, Engineering, and Medicine report on Reproducibility and Replicability in Science (2019)

## To show readers how to use modern analysis techniques on heliophysics data.

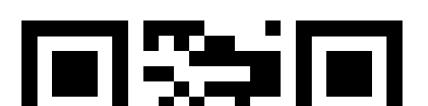
"...scientific research in many disciplines is becoming more and more dependent on the careful analysis of large datasets. This analysis requires a skill-set as broad as it is deep: scientists must be experts not only in their own domain, but in statistics, computing, algorithm building, and software design as well."

. . . . . . . . . . . .

– Jake VanderPlas, author of The Python Data Science Handbook (2016)

We consider this book a living document. Please contact us if you'd like to submit a chapter!

**By** Monica Bobra (Stanford University) and James Mason (University of



### Colorado at Boulder)

With contributions by Andrés Asensio Ramos, Mark Cheung, Carlos José Díaz Baso, David Fouhey, Richard Galvez, Meng Jin, Andrés Muñoz-Jaramillo, Brandon Panos, Alexandre Szenicer,

Rajat Thomas, and Paul Wright

Take a picture to

load the book