MACHINE LEARNING APPLIED TO ASTEROID TAXONOMY BASED ON REFLECTANCE SPECTROSCOPY: AN OBJECTIVE METHOD. Sydney M. Wallace[1,2], Thomas H. Burbine[2], Daniel Sheldon[3], and M. Darby Dyar[2]. [1]Harvey Mudd College, 301 Platt Blvd., Claremont, CA 91711(smwallace@hmc.edu), [2]Department of Astronomy, Mount Holyoke College, 50 College St., South Hadley, MA 01075, [3]College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA 01003.

Introduction: The Bus-DeMeo (B-DM) asteroid taxonomy [1,2] has revolutionized our ability to organize asteroid populations by making it possible to group similar objects together. This classification was initially based on slope values and principal component scores that were computed for the Small Main-belt Asteroid Spectroscopic Survey (SMASSII). There are 26 main classes defined on the basis of specific features. Since B-DM was developed, several enhancements have been added. It has also become apparent that visual inspection of data is often necessary for correct classification, even though this introduces subjective judgments into the process.

In this project, we apply more modern machine learning classification algorithms to the task of asteroid taxonomy using two common machine learning (ML) classification tools: logistic regression and *k*-nearest neighbors (*k*-NN), with a goal of developing a modern classification algorithm that does not require human intervention.

Data: We obtained a total of 686 asteroid spectra from several research groups (see Acknowledgments) including the 371 spectra used in the original classification papers [1,2]. Data from all sources were resampled in 0.05 μm steps over a range from 0.45 to 2.5 μm. We then used a batch version of the B-DM Taxonomy Classification Web tool from Stephen Slivan (http://smass.mit.edu/busdemeoclass.html) to determine the B-DM class for each of the spectra obtained from non-MIT sources.

One limitation of our methods (and the original taxonomy) is the varying number of spectra per class (**Figure 1**). In this project, we dropped the classes with fewer than 4 representatives because the ML methods do not do well with so few examples for training.

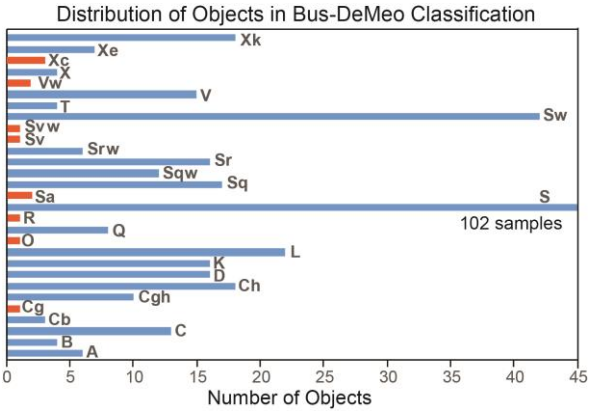

Figure 1. Distribution of 371 objects used in the original Bus-DeMeo taxonomy [1,2]. Orange bars represent classes with too few examples for subsequent training.

Data Analysis: Data analysis was undertaken using an in-house tool written in Python and utilizing the SciKit-learn library [3]. We used two types of classification algorithms. *Logistic regression* (LR) is a classical technique that predicts the probability that an input value belongs to a particular class. Like linear regression, it is a parametric linear model that estimates coefficients for each dimension of the input; however, its prediction is a categorical class label instead of a real number. *K-nearest neighbor* (kNN) classification is a non-parametric technique used to classify new samples based on their similarity to samples in known classes from a training data set. The parameter *k* controls how many samples are used to predict the label of a new sample. The class is assigned on the basis of the most common class of the *k*-nearest neighbors.

Results: First, we determined if the ML methods could accurately predict the Bus-DeMeo classes with >4 examples. Results are shown in **Table 1**, along with the average accuracy of each method and the classes to which wrong matches were assigned. Second, we tested that classifier by applying it to the two groups of asteroid spectra we obtained from colleagues and a suite of meteorite spectra obtained from RELAB [4].

Discussion: With 67% and 78% overall classification accuracy, respectively, the LR and kNN methods did a remarkably good job of classification, though with slightly different results. A majority of mismatches inTable 1 occurred when matching to asteroids in similar classes with very subtle distinctions. However, there is no evidence that many of the B-DM classes are mineralogically distinct. Nor is there any way to test the robustness of the B-DM taxonomy without sample return from each object. It is possible that some of the subdivisions in the B-DM taxonomy are statistically unsupported, and that a more simplified taxonomy might be more appropriate. In the long run, the optimal asteroid taxonomy would be tied directly to the meteorite classes that are (to at least some extent) derived from them, to allow the underlying mineralogical bases for the differences among classes to be understood.

The ability to predict B-DM classes further degrades when applying ML models trained on the origi-

Table 1. B-DM Classification with ML Methods

Class	Logistic Regression		k-Nearest Neighbors	
	A	Mistakes	A	Mistakes
A	100		100	
B	100		100	
C	100		100	
Cgh	70	Ch(2), S(1)	90	Ch(1)
Ch	100		94	Xk(1)
D	100		100	
K	87.5	S(1), L(1)	94	Xk(1)
L	100		91	S(2)
Q	100		100	
S	96	L(1), Sw(2), Sqw(1)	98	Sr(1), Sq(1)
Sq	35	S(10), Xk(1)	71	S(3), K(1), Q(1)
Sqw	58	S(2), Sw(3)	92	Sw(1)
Sr	25	S(12)	69	S(5)
Srw	17	S(4), Sw(1)	33	Sw(4)
Sw	66	S(14)	95	Sqw(1), S(1)
T	0	Xk(3), L(1)	100	
V	100		100	
X	0	Xk(4)	75	Xk(1)
Xe	43	Xk(2), S(1), L(1)	57	Xk(2), Ch(1)
Xk	83	S(2), Ch(1)	89	T(1), Cgh(1)

Class = B-DM class, A = % accuracy, Mistakes = assigned misclassifications.

nal B-DM asteroids to other data sets (**Table 2**). However, applying *any* classifier to "unseen data" in this manner is rife with problems, particularly the assumption that the test data are drawn from the same population. Fortunately, this scenario has been studied in ML, using techniques such as calibration transfer to align datasets from different instruments [5,6]. Preprocessing methods such as better baseline removal, normalization, squashing and smoothing are also likely to improve matching accuracy, as shown with other types of spectroscopy [7,8]. This is a problem in need of rigorous exploration using a full array of ML tools.

Summary: This project assesses how well objective, principled ML methods approximate the results of the B-DM taxonomy. It is clear that they have great potential to do so. The automated methods have the advantage of being objective, easy to run, and lacking any need for human visual inspection. They also have the potential to objectively match any other independent classification scheme. Asteroid taxonomy can be greatly improved by using the growing number of new observations, linking to meteorite spectra with known mineralogies, and leveraging ML methods.

Acknowledgments: This project was supported by the RIS^{4}E and VORTICES nodes of the NASA SSERVI program, (grant NNA14AB04A) and MFRP grant NNX15AC82G. We thank F. DeMeo, R. Binzel, D. Polishook, S. Slivan, C. Thomas, M. Lucas, and J. Emery for generously providing us with asteroid spectra and assistance.

References: [1] Bus S. J. and Binzel R. P.(2002) *Icarus*, 158, 146-177. [2] DeMeo F. E. et al. (2009) *Icarus*, 202, 160-180. [3] Carey C. et al. (2017) *LPSC XLVIII*, Abstract #1097. [4] Wallace S. et al. (2019) this conference [5] Boucher T. et al. (2017) *J. Chemometrics*, 31, e2877. [6] Boucher T. (2018) Ph.D. thesis, UMass Amherst. [7] Carey C. et al. (2015) *J. Raman Spectr.* 10.1002/jrs.4757, 894-903. [8] Dyar M. D. et al. (2016) *Spectrochim. Acta B*, 126, 53-64.

Table 2. Accuracy When Applying B-DM LR Classifier to Other Datasets

Class	Near-Earth Asteroids		HARTSS		RELAB Meteorites	
	A	Mistakes	A	Mistakes	A	Mistakes
A	100				50	C(1)
B	60	C(2)	67	S(1)	55	Ch(20), Cgh(3), K(7), C(2), V(2), S(6), Sq(4), Q(2), Sr(1)
C	44	Ch(3), B(1), Xk(1)				
Cgh						
Ch	0	C(1)				
D	75	C(1)				
K						
L	37.5	S(1), C(2), Sw(1), Xk(1)				
Q	46	Sq(10), Cgh(1), S(4)			46	S(13), B(23), Sq(38), C(3), Ch(9), Xk(1)
S	77	Sq(2), Sw(2), K(1), B(1), Sr(1), Cgh(1)	40	Sr(1), K(1), L(1)	73	L(23), Q(1)
Sq	9	Q(1), S(16), Sr(2), Sw(1)	0	S(1)	4	B(7), S(68), Q(1), Ch(4), Xk(1)
Sqw	17	Sw(1), Sr(1), Sq(2), S(1)	0	Sw(2)		
Sr	11	S(8)			0	S(32), Q(13), Sq(3), B(1)
Srw	0	S(2), Sw(1)	0	Sr(2), S(1), Sw(1)	33	Sr(1), D(1)
Sw	53	S(6), Sqw(1)	44	S(3), D(1), L(1)	75	A(1), L(1)
T					0	Xk(1), C(1)
V	91	Q(1)	100		91	B(10), Q(7), Sq(2), S(2)
X	0	Xk(5), C(2), Xe(1)				
Xe	0	L(2)				
Xk	0	Cgh(1), Ch(1), S(2)				