UNIVERSITY OF MINNESOTA
School of Physics & Astronomy

Proposed Syllabus for a Graduate Course *Big Data in Astrophysics*
AST 8581 / PHYS 8581 / CSCI 8581

The Course:
This course will be a part of the new NSF research training program *Data Science in Multi-messenger Astrophysics* and of the proposed graduate (M.S. and Ph.D.) minor degree (*Data Science in Astrophysics*). The course is designed to appeal to graduate and advanced undergraduate students in both Physics/Astrophysics programs and in Statistics, Computer Science, and Engineering programs. It will assume minimal pre-requisites, and it will provide the necessary background to students who will pursue data-analysis research projects in the field of astrophysics.

This course will introduce key concepts and techniques used to work with large datasets, in the context of the field of astrophysics. In the first 4 weeks of the course the focus will be on the modern approaches to creating and manipulating databases, with the focus on SQL that is heavily used for astrophysics survey data. The remaining part of the course will focus on a range of machine learning techniques for processing data: classification algorithms (supervised and unsupervised learning), clustering algorithms, regression problems, recommender systems, graphic models and others. The course will dedicate about 2 weeks to each algorithm type: the algorithms will first be introduced in 1-2 lectures, and the emphasis will then be placed on team projects in which the students will apply the algorithms (and already available packages) to astrophysical data sets to answer specific astrophysics questions. The course will assume familiarity with basic concepts in astrophysics, but it will include brief reviews as needed to demonstrate the use of modern data analysis techniques.

For graduate students in Physics and Astrophysics programs, this course will provide a systematic introduction to some of the most modern data analysis techniques used in astrophysics today. For graduate students with traditional data science backgrounds (Computer Science, Statistics, Engineering), this course will provide immersion into domain-specific data analysis problems with numerous examples of fundamental problem-specific challenges in applying the modern data analysis techniques.

Proposed Textbooks:
Primer: The primary text for this course will be a primer that will be prepared to cover the diverse topics of this course. The primer will include instructors' notes as well as selected papers.

Supplementary Textbooks:
Statistics, Data Mining, and Machine Learning in Astronomy; A Practical Python Guide for the Analysis of Survey Data, Z. Ivezic, A.J. Connolly, J.T. VanderPlas, and A. Gray, Princeton University Press, 2014.
Mining of Massive Datasets, J. Leskovec, A. Rajaraman, and J. Ullman, Cambridge University Press, 2014.
Introduction to Data Mining, P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, 2019.

Number of credits: 4

Pre-requisites: Multivariable calculus (e.g. MATH 2263) and linear algebra (e.g. MATH 2243); or instructor consent. Suggested– familiarity with astrophysics topics such as star formation and evolution, galaxies and clusters, composition and expansion of the universe, gravitational wave sources and waveforms, and high-energy astrophysics.

Students and Scheduling: The course will be a part of the new research training program *Data Science in Multi-messenger Astrophysics* and of the proposed graduate (M.S. and Ph.D.) minor degree (*Data Science in Astrophysics*). As such, the course will appeal to both students in Physics/Astrophysics programs and in Statistics, Computer Science, and Engineering programs. It will assume minimal pre-requisites, and it will provide the necessary background to students who will pursue research projects in this field. The course will be offered every Spring, starting in the Spring 2021. The course will be cross-listed between AST, PHYS, and CSCI, and is expected to be taught by faculty in all three programs, as well as in STAT.

Tentative Schedule:

Week	Topic
1	Intro to Big Data, techniques for handling it
2	Astrophysics Datasets
3	Databases - Overview
4	Databases - SQL
5-6	ML - Intro to Classification Algorithms Project Examples: * Separating stars and galaxies in SDSS data * Identification of glitches in LIGO data * Identification of periodicities in the solar corona using SDO/AIA EUV data correlated with radio bursts using spacecraft radio detectors * Classification of plasma wave modes in the solar wind using spacecraft electric and magnetic field data
7-8	ML - Intro to Clustering Algorithms Project Examples: * Supernova hunters using Zooniverse data * Identifications of boundaries (such as shocks) in solar, solar wind and planetary satellite data
9-10	ML - Intro to Regression Algorithms Project Examples: * Removal of environmental contamination from LIGO data
11-12	ML - Intro to Recommender Algorithms Project Examples: * Broker system for looking for a particular type of objects/events in LSST data
13-14	ML - Intro to Graphic Models Project on Graphic Models (TBD)
15	Special Topics (e.g. text mining, Zooniverse)