DATA MINING, MANAGEMENT, AND CURATION

(SPECIAL TOPICS IN PHYSICS)
DATA 3401 (PHYS 5391)
Unique Number: 55771 (55780)
Summer 2020
https://www.cdslab.org/DMC2020U/

Instructor:	Amir Shahmoradi
office:	SEIR 365
e-mail:	a.shahmoradi@uta.edu
office hours:	Thursdays 1:00-3:00 pm

Class start/end:	June 8, 2020 – August 13, 2020
Lecture meeting times:	Tuesdays – Thursdays 10:30 – 12:20 pm
Lecture meeting place:	Teams Virtual Room
Lab meeting times:	Thursdays 1:00-3:00 pm
Lab meeting place:	Teams Virtual Room

Teaching Assistants:	NONE
office:	NONE
e-mail:	NONE
office hours:	NONE

COURSE OBJECTIVES / ACADEMIC LEARNING GOALS

This lecture and lab course will provide training in working with databases, including data mining techniques and principles and best practices in data management, storage, and curation. **Prerequisite**: DATA 1401, DATA 1402, **or** with the permission of the instructor.

The primary objective of this course is to study a variety of techniques for data mining, predictive modeling, and machine learning. Upon completing this course successfully, you will be able to understand the pros and cons of different data mining techniques, so that you can (i) make an informed decision on what approaches to consider when faced with real-life problems requiring predictive modeling, (ii) apply models properly on real datasets so to make valid conclusions.

COURSE SCHEDULE

The following is a tentative outline of topics to be covered:

- **Version Control Systems** (VCS): Principles of professional project management and collaborative programming with the use of Git. (1 week)
- **Quick introduction to SQL** as well as Python/MATLAB (1 week)
- **Data Pre-Processing**: Transformations, Imputations, Sampling, Outlier detection (1 week)
- **Modern Interactive Visualization** via Python and MATLAB (1 week)

- **Regression**: linear regression; multiple linear regression; logistic regression; Bayesian methods; dealing with sparse data; and (if we have enough time and appetite) ridge and lasso regression (3 weeks)
- **Ensemble Methods**, model averaging, bagging and random forests, bootstrapping (1 week)
- **Classification**: Bayes decision theory, Naïve Bayes; and (if we have enough time and appetite) Deep Learning and Tensorflow (1 week)
- **Clustering**: k-means; k-mediods; nearest neighbor; hierarchical clustering (1 week)

COURSE TEXTBOOKS

No textbook is required for this course. Online class lecture notes will be used as reference. However, a list of textbooks for those who are interested to self-educate themselves or go beyond class syllabus is provided below,

- Principles of Data Mining, Bramer
- Machine Learning: A Probabilistic Perspective, Kevin Murphy
- Pattern Recognition and Machine Learning, Bishop
- The Elements of Statistical Learning, Trevor Hastie, Robert Tibshirani, and Jerome Friedman (HTF)

COURSE LOGISTICS

Grading:
Weekly Homework: 33% (Assignments might not be weighted equally)
Weekly Quizzes: 33%
Final Project: 34%

Homework Policy:
There will be approximately one homework per week. Assignments will be due every Tuesday before the lecture begins and should be added to an online repository determined by the instructor. No late assignments will be accepted. No exceptions to the homework policy will be made without prior instructor approval.

Examinations:
There will be no midterm or final exams. Students will have to complete a project in place of the final exam, (possibly, in collaboration with their teammates who are determined randomly after the midterm).

Quizzes:
There will be weekly quizzes at the beginning of each lab session on Thursdays.

Attendance:
Regular attendance is expected. Any absence requires prior approval from the instructor, or compelling evidence of illness or an official letter from the university administration. Student attendance will be randomly checked.

Scholastic dishonesty: All students are responsible for upholding the University rules on scholastic dishonesty. Students who violate University rules on scholastic dishonesty are subject to disciplinary penalties, including the possibility of failure in the course and/or dismissal from the University. Since such

dishonesty harms the individual, all students, and the integrity of the University, policies on scholastic dishonesty will be strictly enforced.

Other matters: The University of Texas at Arlington provides, upon request, appropriate academic adjustments for qualified students with disabilities. Any student with a documented disability (physical or cognitive) who requires academic accommodations should contact the UTA's Office for Students with Disabilities as soon as possible to request an official letter outlining authorized accommodations. For visit https://www.uta.edu/disability/.