# CS4315, Machine Learning and Big Data (3-1)
# Syllabus for Summer 2021

## Catalog description

A survey of methods for process large amounts of data and classifying and analyzing it using machine-learning methods. Big-data topics examine the obstacles to processing including managerial obstacles, problems of data consistency, problems of data accuracy, data-reduction methods, and big-data distributed processing methods. Topics on machine learning include concept learning, decision trees, Bayesian models, linear models, neural networks, case-based reasoning, genetic algorithms, sequence learning, and assessment techniques. Students will do projects with software tools on military data.

## Instructor

Prof. Neil Rowe, ncrowe@nps.edu, (831) 656-2462 and 373-1732. More information about research projects and software mentioned in class is at http://faculty.nps.edu/ncrowe.

## Textbook and notes

The required textbook is I. Witten, E. Frank, M. Pal, and C. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, fourth edition, Morgan Kaufmann, 2016. The third edition is sufficient, but the fourth edition is better. There are some good videos explaining about Weka and data mining at www.youtube.com/user/WekaMOOC.

I am told you may be able to get the book for free from O'Reilly Books. To access O'Reilly (formerly Safari) you need to go through the Navy MWR Digital Library in order to get a free O'Reilly account; go to www.navymwrdigitallibrary.org. To get started, log in to the digital library with a CAC (or register for a DS Logon account). Once you've logged into the library, find the O'Reilly section and use the 'Visit' button to register for an O'Reilly Learning account. A shortcut is https://www.oreilly.com/self-registration/dod-mwr-libraries-navy/. After your O'Reilly account has been established, you can access it from outside the Digital Library at https://learning.oreilly.com or downloading the O'Reilly app.

## Lectures

Wednesdays 1300-1600 PST on Zoom. Quizzes will be done at 1500 during class time in the days listed on the schedule.

## Software

As a data scientist, you need to be familiar with a variety of tools, so we will not use R in lectures. In particular, you will need to run these open-source tools, which will require downloading them if you use a home computer.

- WEKA (https://waikato.github.io/weka-wiki/downloading_weka/). This is a simple machine-learning tool that does most of the work for you, as it is intended for people that are not programming experts. The only major challenge is in creating input in a format that it will accept, since it is picky.

- The Python programming language (http://python.org/downloads). You may be able to run this through the Jupyter AMS site; otherwise, it is easy to download. Many complex investigations require a full programming language, even though you won't need much of it at a time. Python has a lot of software available, more than R.
- SQL Developer (https://www.oracle.com/tools/downloads/sqldev-downloads.html). This is a GUI (graphical user interface) to the Oracle XE Database running on an AWS (Amazon Web Services) server. Oracle downloads require creating an account on Oracle.com, for this you can use any email and then set the password. Each of you will get your Oracle XE (AWS) credentials (login/password) separately via email with instructions on how to connect (and it will also be documented in Sakai). Once connected you can load CSV files into Database tables, use SQL queries to clean up the data and save it back into a CSV file.
- Jupyter notebook (Jupyter Hub on AWS) credentials will also be sent by email to each of you.
- Optionally, a text editor that supports macros. Simple fixes to data can often be quickly done with macros. If you have no preference, we recommend the Emacs editor (https://www.gnu.org/software/emacs/download.html) over Windows and Mac proprietary text editors since it is easier to implement macros.

You need these on a site or computer that you can use for class assignments. If correctly installed on Windows, they should appear in the list of programs when you left-click the Start button; on a Mac, they should appear in the Launchpad. Students generally have more success installing them on home computers than work computers because of the software restrictions in most organizations. On the class project, however, you can use whatever software you want.

## Grading

Grading is based on two homework assignments and three quizzes, and a class project. They are weighted approximately equally. The homework assignments are substantial, so pace yourself and do not expect to do the homework only on the night before it is due. Grades are assigned relative to the rest of the class so you are not penalized when a question is difficult for everyone. Overall, a median grade in this course is between an A- and B+. You will only get below a B if you do not appear to be doing the required work.

Homework should be submitted electronically as a single digital document with your name at the top of the document, all material for each problem together (with no appendices), and without the text of the problems (just your answers). Homework should be submitted in a format which is either text (TXT), Microsoft Word DOCX, or Adobe Acrobat PDF; Word format is preferred. Excel and Jupyter Notebook formats are not acceptable formats for homework.

Quizzes are open-book and open-notes, but you may not communicate with anyone during them. You may also not use the Internet during the quiz except to get to the quiz site. The expected class median is 70%.

Generally, homework questions are written so quoting or paraphrasing something written somewhere else will not answer the question. Homework must be done individually without communicating with anyone except the instructor. You can look up or discuss general knowledge of the subject with other people, but you cannot discuss the homework questions with anyone besides the instructor. Homework submitted after the due date but before solutions are made available is subject to a 15% penalty; homework will not be accepted after solutions are made available.

The class project is a data-science investigation of a data set you choose. Kaggle.com has a lot of data sets, and you can find others online in such places as government sites. The project should involve at

least 200 rows of data with 10 attributes per row, and should require at least 10 hours of work. You will write a report of at least 1000 words on this project. Your report should describe the dataset, what methods you used to analyze it, what results you got, and what you conclude from the results.

## Schedule

By 7/14: Read Chapter 1 and sections 5.1, 5.2, 5.3, and 5.8 in both editions of the textbook
By 7/21: Read Chapters 2 and 3, and sections 4.4, and 4.5 in both editions; the videos at www.youtube.com/user/WekaMOOC provide additional information about Weka
By 7/28: Read sections 4.1 and 4.3 in both editions
On 7/28: Quiz 1 (sections 1-3) in class
By 8/4: Read sections 6.1 and 6.2 in both editions
On 8/4: Homework 1 due in class
By 8/18: Read section 4.2 in both editions
On 8/18: Quiz 2 (sections 4-7) in class
By 8/25: Read section 7.2 introduction, "The Maximum Margin Hyperplane", "Nonlinear Class Boundaries", and "Multilayer Perceptrons" including "Backpropagation"
By 9/1: Read sections 4.7 and 4.8 in both editions, and section 7.1 in the 3rd edition and 8.1 in 4th edition
On 9/8: Homework 2 due in class
On 9/15: Quiz 3 (sections 8-11) in class (last day of class)
On 9/17: Class project due

## Course Outline

Part 1: Introduction and data setup
Part 2: Data manipulation (wrangling)
Part 3: Introduction to machine learning
Part 4: Logical learning
Part 5: Decision graphs
Part 6: Bayesian models
Part 7: Neural networks
Part 8: Case-based reasoning
Part 9: Sequence learning
Part 10: Evolutionary learning
Part 11: Further directions and conclusions