605.448 – Data Science

Syllabus

Instructor(s) Contact

Stephyn G. W. Butcher
Cell: 202-257-5937 (emergencies)
Instructor E-mail: EN685.648+sbutcher@gmail.com

Andrew Stewart
Cell: 202-505-3633 (emergencies)
Instructor E-mail: EN685.648+astewart@gmail.com

Course E-mail: EN685.648@gmail.com

Please use the course email unless you wish to contact a specific instructor.

We will make every effort to respond to your email **within 24 hours or earlier**. If an issue is urgent, please indicate "urgent" within the subject line of the email and I will respond as soon as is practical. Additionally, one of us will check Blackboard once a day, every day of the week (and so should you!).

Course Description

This course will cover the foundations of the emerging field of data science. We will discuss the Data Science Process, Data Acquisition, Probability and Inference, Visualization, Exploratory Data Analysis, Regression, Model Evaluation and a variety of Machine Learning techniques. Students will gain direct experience in solving the programming and analytical challenges associated with data science through short assignments and a larger project.

We will emphasize breadth over depth because we only have 14 weeks to cover a huge multi-disciplinary field.

Prerequisites

There are no official prerequisites for this course. You will have an easier time of it if you know Python already. Additionally, it is highly recommended that you do not take this course with any other course as the workload is quite heavy.

Course Goals

Data Science is a multi-disciplinary, emerging field in flux. The goal of this course is to introduce you to the foundational topics and skills of a Data Scientist. If you want to know about TensorFlow, watch YouTube. At the end of the course, you should be able to determine a problem appropriate to Data Science, produce a plan for solving the problem and execute that plan.

Course Objectives

By the end of the course, you will be able to:

- Identify a problem appropriate to Data Science and formulate a proposed solution in cooperation with various stakeholders and team members.
- Explain the role of Probability plays in Data Science and apply Statistical Inference where appropriate.
- Obtain data appropriate to the problem, process and verify it understanding the importance of provenance and reproducibility.
- Explore data sets for errors and structure.
- Build, interpret and improve on statistical models such as Linear Regression and Logistic Regression.
- Build, interpret and improve on machine learning models such as k-Nearest Neighbors, k-Mean clustering and Recommender Systems.
- Clearly communicate results using various modes of communication including Visualization.

Course Structure

The course materials are divided into modules which can be accessed by clicking Course Modules on the left menu of Blackboard. A module will have several sections including the overview, content, readings, discussions, and assignments. You are encouraged to preview all sections of the module before starting. All modules run for a period of seven (7) days.

You should check Blackboard every day. Additionally, make sure that the email on file with the Registrar and associated with your student account is either forwarded to an email you check everyday or is an email you check every day. (My understanding is that you should use your JHU email and have it forwarded if you don't check it every day. See http://my.jhu.edu/ .)

Textbook

Required

There is no required textbook for this course because no such textbook exists. There are quite a few "popular" books on Data Science published by O'Reilly but none are suitable for a graduate level course in Data Science. *There is a heavy emphasis on the lectures, labs, and problem sets as a result.*

However, the notes that have been produced for this course over the last 6 years are slowly reaching "book" status. Throughout the semester you will have access to drafts of the book including the associated Jupyter Notebooks. This "book" will be referred to as *Fundamentals of Data Science* or *Fundamentals* throughout the semester.

Optional

I will post a suggested reading list in the General Discussion Forum.

Required Software

Linux

Linux is not exactly required but you should be very familiar with Linux-like operating systems and commands. A significant amount of data acquisition, viewing and cleaning can be accomplished at the command line using Linux commands.

If you use Windows, you should consider installing a VM on your machine and popular Linux flavor. You *might* be able to get away with Cygwin.

Python

The official language of the course is Python 3.7. There are a large number of libraries we will be using this semester (either directly or indirectly). Additionally, you will need to use Jupyter to view course materials and complete course assignments. All of these are provided by the Anaconda distribution of Python which I suggest you install if you can.

https://www.continuum.io/downloads

The default for Anaconda should be Python 3. Use python –version or python3 –version to check.

If for some reason you do not want to install Anaconda (like me, you may have a set up you like or need for work), I suggest running a VM with the Anaconda distribution even if you are already using a Linux-like system (Linux or MacOS).

Technical Requirements

You should refer to Help & Support on the left menu for a general listing of all the course technical requirements as they related to *Blackboard*.

Student Coursework Requirements

This course requires a lot of programming. I cannot reliably predict how long it will take you to complete assignments because a single bug, typo, fat-fingered command may make a 15 minute assignment into a 2 hour debugging session. As students with some programming background, I expect you to understand this. When I say "this should take an hour", that means—if you don't have a typo, have all the libraries installed, etc. Plan for this.

Additionally, I cannot provide you with everything you will need to know. For example, I cannot provide tutorials on every library. You will have to be self-reliant and resourceful, using Stackoverflow, library documentation (such as it is) and other such resources on the Internet to answer questions you may have. This is a key skill for Data Scientists to possess. You may also post questions as well as share particularly relevant references in the Tools discussion forum on Blackboard.

Finally, because there is no formal textbook, I have higher expectations of students for self-directed investigation and curiosity. This is a key characteristic of a good data scientist.

The workload breaks down as follows:

1. Reading the chapters/notebooks (every week).
2. Watching the lectures (every week).
3. Completing the Lab (every week).
4. Completing the Assessment (every week).
5. Completing the Group Discussion (2 week cycle).
6. Completing any Class Participation items such as Class Discussions, posting to "In the News", posting questions to "Muddy Point" posts, etc (weekly)
7. Completing Problem Sets (2 week cycle).

The only exception is the first week/module.

This course will consist of four basic student requirements: Preparation/Participation, Labs, Assessments, Problem Sets, and Group Discussions.

1. **General Preparation and Participation**

You are, of course, expected to read the specified materials each week and watch the lectures. Additionally, you are responsible for **all posts on Blackboard**.

Throughout the semester, there will also be various opportunities for you to contribute to the course.

1. There will be a Discussion forum “In the News”. You are expected to contribute links to articles about Data Science and read links posted by your classmates. *Do not post links about Machine Learning or Artificial Intelligence*.

2. There will be Group Discussions. These Group Discussions will run two (2) weeks. The first week involves preparation, usually something from your Lab, and some coordination.

   During the second week you will read your group member’s submissions and make a series of comments. The first set of comments will be by Sunday. The second set of comments will be in response to your group mates comments to your initial post.

   Typically, Group Discussion posts are due on Sundays and Tuesdays.

   During every Group Discussion there will be a Group Leader responsible for overseeing the discussion. Find the directions on Blackboard.

3. There will be a Muddy Point discussion that runs two weeks. During that two week period, you must ask at least one question directly pertaining to the material, which I will answer.

2. **Labs**

The Lab is the chief vehicle of hands on exploration of the topics for each Module. The Lab is due by Tuesday midnight. A Key will be posted on the following Wednesday. *It is your responsibility to compare your answers with the Key and make sure you understand why you may have been wrong.*

Labs are *self-checks* but will also be randomly monitored for correctness and completeness. They must be turned in by 11:59 ET on **Tuesday** for the specified Module.

3. **Assessments**

Assessments will test your conceptual understanding the Module’s material. There will always be 10 questions (true/false, multiple choice, etc.) covering the Modules content or possibly content from previous modules to reinforce important concepts.

The assessments are timed and show questions one at a time. There is no backtracking. Answer carefully. They must be taken by 11:59 ET on **Tuesday** for the specified Module.

4. **Problem Sets**

While the Labs will guide you towards understanding the concepts and tools covered in the Module, each Problem Set will be your opportunity to demonstrate your understanding. They replace a mid-term or final in the course so you should assign the highest priority to them. They are basically practical examinations.

Problem Sets will be offset from the material they cover. For example, Problem Set 1 (it covers Modules 1 and 2) will be assigned during Module 3 and due at the end of Module 4. Every Problem Set covers two weeks of material and you have two weeks to do it.

This offset schedule permits two things. First, you get experience and an opportunity to ask questions through the Labs along with ample time for feedback before you have to apply the technique to a graded problem. Second, you have two weeks to work on the assignment as you see fit (recognizing you'll need to fit it in with the regular work for the Module) rather than trying to finish Lab *and* a Problem Set on the same topics. This also means that you will not see your first Problem Set until Module 3, 2 weeks into the course.

The Problem Set is due by 11:59 PM ET on **Tuesday** for the specified Module.

Grading

Assignments are due according to the dates posted in your Blackboard course site. You may check these due dates in the Course Calendar or the Assignments in the corresponding modules.

Assignments will be graded in one to two weeks. Generally, it takes me or the grader as long to grade N students' work as it took one student to complete their submission (so, 1 week to grade if 1 week to do; 2 weeks to grade if 2 weeks to do, etc.). Assessments will be graded automatically (and immediately).

On Late Submissions…

I do not accept late submissions for a grade without prior consultation, except in the case of extreme emergencies (the birth of a child, incapacitating illness, etc). The following are not legitimate reasons: work, taking other classes, weddings, family reunions, holidays, anniversaries, business travel, work deadlines, etc., etc. *without prior arrangement.*

If there is a problem, do not wait until five weeks have passed, contact one of the instructors immediately before you miss an assignment deadline.

Most of the assignments in this class have very specific due dates and some require coordination/participation with groups of other students. These all need to be submitted in a timely manner. For example, if you do not post, there can be no discussion.

A **grade of A** indicates achievement of **consistent excellence and distinction** throughout the course—that is, conspicuous excellence in all aspects of assignments and discussion in every week.

A **grade of B** indicates work that **meets all course requirements** on a level appropriate for graduate academic work. These criteria apply to both undergraduates and graduate students taking the course.

I cannot stress this enough, merely working hard is not grounds for an A. You may very well work very hard for a B. That happens. I once worked very hard for a C in the Advanced Microeconomics (actually, I worked too hard and did not ask enough questions!).

Many students work hard instead of working smart. Ask questions.

Specifications Grading

This class does not use the traditional 100 point scale, I do not weight or average things. There is no point gaming. You must perform with excellence in all areas of the class in order to get an “A”. This means you cannot plan to do better on tests as compared to problem sets or better at the end of the semester as opposed to the beginning of the semester.

Here are the categories of work:

Labs	12
Assessments	13
Problem Sets	5
General Participation	Group Discussions, In the News, etc.

A few assignments are binary (pass/fail). I merely note if you turned them in or if they had an acceptable level of effort (an incomplete Lab might be a 0, for example).

Everything else will be scored on a 4 point scale:

4 – Excellent. You completed the assignment in a timely manner, demonstrating a thorough understanding of the technique, tool or concept and conducted an excellent exploration of its use. If it is a discussion, your post was substantive, did not just quote other materials, and contributed to the on-going discussion. **You went *above what was required, asked for, or expected.*** Sometimes an A is doing exactly what is necessary but not more.

3 – Acceptable. You completed the assignment in a timely manner, **you did exactly what was requested**, demonstrating a sufficient understanding of the technique, tool, or concept. There may have been minor deficiencies. If it was a discussion pot, the post contributed to the discussion but it may have been a reference to other materials or perhaps even slightly off topic. You may have done more too much in the hopes that something was correct. That is, if you say 4 things hoping one of them is right, that’s not an A, that’s a B. Succinctness is appreciated. The “exceeds” must be relevant.

1,2 – Unacceptable. You either did not complete the assignment, it was not timely or **you did what was minimally required**. There are significant areas of confusion. A lack of exploration or curiosity about the concept, tool or technique. If it was a discussion post, it may have been off topic.

0 – Oops. You did not submit the assignment on-time or post on-time.

Basically the only way you can get a 0 is by not doing something on-time.

Grades

Final Grades are based on the counts of scores.

For an A, you must at least achieve:

Labs	10 of 12 submitted with a 1 (“P”).
Assessments	10 of 13 submitted with a 4 (“A”) with only 1 “0”.
Problem Sets	3 of 5 with a 4 (“A”) (the remainder must be 2 or higher).
General Participation	“A” level of participation (Group Discussions, In the News, etc.)

For a B, you must at least achieve:

Labs	10 of 12 submitted with a 1 (“P”).
Assessments	10 of 13 submitted with a 3 (“B”) or higher with only 1 “0”.

Problem Sets	4 of 6 with a 3 (“B”) or higher (the remainder must be 2 or higher).
General Participation	at least a “B” level of participation (Group Discussions, In the News, etc.)

Important – if you score a 2 or lower on the Problem Set, you must schedule a 1:1 Zoom session with one of the instructors to go over your submission so that any deficiencies are rectified.

So there are a few things to note:

1. Basically, you’re getting an “A”, “B”, “C”, “D”, “F” (4 through 0) or pass/fail (1/0) on all assignments. Point percentages don’t mean anything. So don’t see 3 out of 4 points as 75%, that’s not how it works. It’s just easier to encode numbers.
2. There is some leeway built into Assessments. They have 5 points and you can miss 2 questions and still get a 4 (“A”).
3. You can miss 2 Labs, 1 Assessment but no Problems Sets and still get an “A”. Be careful for when you use those.
4. There’s no slacking off once you hit 10 A’s or 10 B’s.
5. **THERE IS NO AVERAGING. Failure to get an “A” in all categories is a “B” or even a “C”.** You cannot do well in one area and still get an “A”.

Anything below this level of accomplishment will result in a C or lower, at the discretion of the instructors. As the semester unfolds, I may find it necessary to adjust both the assignments, criteria or both. I will assign “pluses” or “minuses” at my discretion.

I generally do not directly grade spelling and grammar. However, egregious violations of the rules of the English language will be noted without comment. Consistently poor performance in either spelling or grammar is taken as an indication of poor written communication ability that may detract from your grade.

Communication is very important in Data Science so I will tend to be a bit more picky about formatting, grammar and spelling. If your submissions look like a ransom note, however correct they might otherwise be, they will be counted as wrong.

Help & Support

You should refer to Help & Support on the left menu for a listing of all the student services and support available.

Policies and Guidelines

Assignments, due dates, materials can be changed at any (reasonable) time. Please log into Blackboard *at least once a day* to scan new discussions for such changes.

You are responsible for all discussions/content in the forums on Blackboard.

Academic Integrity

Academic Misconduct Policy

All students are required to read, know, and comply with the Johns Hopkins University Krieger School of Arts and Sciences (KSAS) / Whiting School of Engineering (WSE) Procedures for Handling Allegations of Misconduct by Full-Time and Part-Time Graduate Students available at:
https://ep.jhu.edu/wseacademicmisconductpolicy

This policy prohibits academic misconduct, including but not limited to the following: cheating or facilitating cheating; plagiarism; reuse of assignments; unauthorized collaboration; alteration of graded assignments; and unfair competition. You may request a paper copy of this policy at this by contacting
Mark Tuminello
Phone 410-516-2306
E-mail mtumine2@jhu.edu

Policy on Disability Services

Johns Hopkins University (JHU) is committed to creating a welcoming and inclusive environment for students, faculty, staff and visitors with disabilities. The University does not discriminate on the basis of race, color, sex, religion, sexual orientation, national or ethnic origin, age, disability or veteran status in any student program or activity, or with regard to admission or employment. JHU works to ensure that students, employees and visitors with disabilities have equal access to university programs, facilities, technology and websites.

Under Section 504 of the Rehabilitation Act of 1973, the Americans with Disabilities Act (ADA) of 1990 and the ADA Amendments Act of 2008, a person is considered to have a disability if c (1) he or she has a physical or mental impairment that substantially limits one or more major life activities (such as hearing, seeing, speaking, breathing, performing manual tasks, walking, caring for oneself, learning, or concentrating); (2) has a record of having such an impairment; or (3) is regarded as having such an impairment class. The University provides reasonable and appropriate accommodations to students and employees with disabilities. In most cases, JHU will require documentation of the disability and the need for the specific requested accommodation.

The Disability Services program within the Office of Institutional Equity oversees the coordination of reasonable accommodations for students and employees with disabilities, and serves as the central point of contact for information on physical and programmatic access at the University. More information on this policy may be found at http://web.jhu.edu/administration/jhuoie/disability/index.html or by contacting (410) 516-8075.

Mental Health Services

Mental health is important for both students and teachers. JHU has resources available for you, even in EP, even remote students.

If you are struggling with anxiety, stress, depression, or other well-being concerns, please consider contacting Johns Hopkins Student Assistance Program (JHSAP). If you are concerned about a friend, please encourage them to seek out counseling. JHSAP can be reached at 443-287-7000 or jhsap.org and JHSAP counselors are available at multiple campuses: East Baltimore, Eastern, Bayview, Washington DC and Columbia (as well as via video for remote students).

Please avail yourself of these resources should the need arise.

Disability Services

Johns Hopkins Engineering for Professionals is committed to providing reasonable and appropriate accommodations to students with disabilities.

Students requiring accommodations are encouraged to contact Disability Services at least four weeks before the start of the academic term or as soon as possible. Although requests can be made at any time, students should understand that there may be a delay of up to two weeks for implementation depending on the nature of the accommodations requested.

Requesting Accommodation

New students must submit a Student Request for Accommodation form along with supporting documentation from a qualified diagnostician that:

- Identifies the type of disability
- Describes the current level of functioning in an academic setting
- Lists recommended accommodations

Questions about disability resources and requests for accommodation at Johns Hopkins Engineering for Professionals should be directed to:

Mark Tuminello
Disability Services Coordinator
Phone 410-516-2306
Fax 410-579-8049
E-mail mtumine2@jhu.edu or ep-disability-svcs@jhu.edu

I reserve the right to make changes as I see fit throughout the course of the semester. Specific items in this Syllabus may be overridden by Announcements or General Posts in Blackboard. It is your responsibility to read and understand what is required of you and to keep track of any changes that may occur.