PDAT 611G: Big Data Management *Course Syllabus*

Welcome: to this (online, mostly asynchronous) PDAT 611 course! I am glad you are here. By the end of the (eight-week) term, you will have taken real steps towards understanding how we think about big data, and some practical ways that we approach it to explore real world problems.

Instructor: Oluremi Abayomi
Email: oabayomi@truman.edu
Availability: M 1:00 pm – 4:00 pm,
F 8:00 am – 1:00 pm.

Academic Success Mentor: Andrea Magg
Office: Kirk 107
Phone: 660-785-7403
Email: amagg@truman.edu
Availability: By appointment

Zoom Link for weekly check-ins: https://zoom.us/j/3082808488

Communication: The best way to contact me is through e-mail. All messages may be answered within 24 hours on weekdays and within 48 hours on weekends. I check email multiple times each day. In order to make sure your e-mail stands out from the rest, make sure your subject line starts with PDAT 611.

Catalog Description: An exploration of techniques used to manage and prepare very large data sets, focusing on tools needed for future classes in the PDAT program.

Prerequisites: Successful completion of PDAT 610 Intro to Data Science, or both STAT 220 Fundamentals of Data Science AND STAT 250 Statistical Computing.

Common Objectives: A successful student will:

- Understand techniques and tools useful in managing large data sets.
- Demonstrate proficiency using R with tidyverse packages for data cleaning, managing tidy data sets, and other processing, while maintaining a reproducible data trail.
- Prepare a large data set for use in a distributed setting such as Hadoop or Spark.
- Use the techniques learned on assignments involving real data.

Materials:

Online Textbook Resources:

Garrett Grolemund and Hadley Wickham, *R for Data Science*, (free online! $33 for a hard copy) http://r4ds.had.co.nz/
Hadley Wickham, 2014. *Tidy Data*. http://vita.had.co.nz/papers/tidy-data.html
Other articles will be available via open-source links.

Online Experience: This course is offered as a primarily asynchronous experience, where students work towards set deadlines to master material, complete assignments, and demonstrate competence. **We do have weekly check-ins (Thursdays at 7 – 8 PM), and while these are optional, previous students have found them very helpful.** As a three-credit graduate class, you should plan on spending about 10-18 hours each week on this course, depending on your background, and it's recommended to spend at least an hour each day, rather than marathon weekly sessions.

Evaluation: Each module will include one or two discussion posts (50 points each) and a short assignment or two (100 points each). Assignments and discussions will be graded holistically according to the following scale. Basically, doing OK is a B and doing well is an A. Anything less than that can be redone. Most assignments may be resubmitted for up to a Check mark in the grading scheme below. One larger project is assigned as homework in Week 7 and combines points for the last two weeks.

Grade	Mark	of 50	of 100
Exceeded Expectations. Excellent. Gosh, Wow!	✓++	54	108
Good. Did all that was hoped for.	✓+	48	96
Met Expectations; acceptable work	✓	44	88
Flawed	✓-	38	76
Didn't "Get it". Unacceptable. Please Re-Do.	✓--	32	64 %
Didn't do it at all. Boo.	✓---	0	0

Of course, Blackboard doesn't like checkmarks, so they turn into point equivalences. Notice that a ✓++ is pretty rare, but most assignments get a check or check-plus, and students who acceptably complete all assignments earn a grade of a B.

Substantiative Interaction: Truman policy and federal regulations require that students demonstrate that they are academically engaged in the courses they take. **You must meet this requirement within the first and a half calendar week of the semester, beginning at 12:00 am on Monday, October 18, 2021, and ending 11:59 pm on Friday, October 22, 2021.** Failure to do so, or to provide an explanation of an extenuating circumstance by that date and time will result in your removal from the course. Under certain circumstances, removal could impact your scholarship eligibility or financial aid. **For the purposes of this class, establishing academic engagement requires, at a minimum, students are required to complete the *M1 discussion: Tidy Data* requirement in your Module 1 folder, no later than 11:59 pm on Saturday, October 23, 2021.**

Grading: There are eight modules in this course. The following table below shows the deadlines for each of the modules. Notice that modules have 2, 7 and 8 have slightly different forms of due dates.

Module #	Module Name on Blackboard	Due date of all assessments in the module
1	Why is Big Data Important	11:59 pm on Sunday 10/24/2021
2	Expanding Tidyverse	11:59 pm on Sunday 11/7/2021
2	Expanding Tidyverse	11:59 pm on Sunday 11/7/2021
3	Web Scraping and JSON	11:59 pm on Sunday 11/14/2021
4	Databases	11:59 pm on Sunday 11/21/2021
Thanksgiving Break: Monday – Friday, November 22 – 26, 2021.		
5	Distributed Computing & Pig (Hadoop)	11:59 pm on Sunday 12/5/2021
6	Spark: Local and Cluster	11:59 pm on Sunday 12/12/2021
7 &8	Cloud Computing & Next Steps	11:59 pm on Saturday 12/18/2021

Course Content Modules:

This PDAT course is designed to be completed as modules, with each module taking approximately one week to complete. That said, you may find some modules longer or trickier than others, so we recommend that you consider working ahead early in the term, so that you have flexibility at the end.

1: Why is Big Data Management Important?
2: Expanding Tidyverse:
3: Web Scraping and JSON
4: Databases: object-relational mapping, SQL, noSQL
5: Distributed Computing; using pig w/Hadoop
6: Sparklyr, Apache Spark through R
7: Google Cloud, AWS, and other cloud-based solutions
8: Exploring other Big Data Issues and Opportunities.

Health and Illness: If you are sick, isolated, quarantined away from your computer, or otherwise unable to do your work, we can figure it out, but let me know ASAP.

Technical Requirements: Members of the class must have regular access to a reliable computer, a webcam/microphone, and the Internet. Try to create a workspace where you can work for a couple of hours in peace every day. It is a good idea to also locate a backup site (for instance, a library, coffee shop, or a friend's house) in case of technical failure. The class principally uses Blackboard for the course content, but also uses R, which runs best on your own machine. We will also explore software tools specifically designed for enormous datasets (such as Hadoop and Spark). While having your own computer is essential, we will also use Truman's Data Science Server, both using the virtual machines located at http://view2.truman.edu, but also by directly accessing the server at http://fire.truman.edu

Important Contacts: Various offices that provide services to online students are identified at the One Stop Services page on online.truman.edu. Should you need to consult with administrators that oversee this department and course, here is the contact information for those individuals:

Statistics Department Chair
Dr. Scott Alberts
Violette Hall 2132
660-785-7649
salberts@truman.edu

Dean, School of Science and Mathematics
Dr. Timothy Walston
Magruder Hall 2004
660-785-4248
samdean@truman.edu

Hopefully your experience with this class is positive. When and if you feel a complaint about this or another course is required, however, the procedure for lodging a complaint can be found on the University's Report a Complaint page. Students taking an online course from outside of the state of Missouri should follow the complaint procedure offered here. **Students are always asked to address their complaint to the professor of the course first when possible, then take their concerns to the Department Chair if the matter cannot be resolved with the faculty member.**

Student Survey of Instruction: You will be asked to complete a survey regarding instruction in this course at the end of the term. The survey is anonymous, and the professor will not see the results until after grades have been completed. This feedback is very important and helps the professor continuously improve the course. It also helps the University make decisions about our overall curriculum. Please be sure to participate in this survey opportunity.

Learner Support: The University provides a range of both academic and student support services to ensure your success. These offices can advise you on learning strategies, point you toward valuable services, and help you troubleshoot technical problems as they arise.
The Center for Academic Excellence provides advising services for students in their first year for most departments, as well as tutoring services. The Center is located in PML 109 and it may be reached at 660-785-7403.
Counseling Services are available on campus at McKinney Center. Appointments may be scheduled by calling (660) 785-4014. An after-hours crisis line is also available at 660-665-5621.
The IT Service Center has combined the IT Call Center, Help Desk and Telephone Services into a one-stop location to serve you. You will find the following services and more when you stop by PML 203 or call 660-785-4544. You may submit a customer support ticket at this web address.

Netiquette and Civil Behavior: Taking a course that leverages technology presents communicators with a challenging task. It is important to remember several points of etiquette that will improve our ability to communicate with each other in our class:

1. Read first, write later. Read the entire set of post(s) or comments before commenting yourself to prevent repeating commentary or asking questions that have already been answered.
2. Avoid language that may come across as strong or offensive. Language can be easily misinterpreted in written electronic communication. Review email and discussion board posts before submitting. Humor and sarcasm may be easily misinterpreted by your reader(s). Try to be as professional as possible.
3. Follow the language rules of the internet. Do not write using all capital letters because it will appear as shouting. Use emoticons if necessary, to convey nonverbal feelings. ☺
4. Keep attachments small. If it is necessary to send pictures, change the size to 250kb or less if possible.
5. No inappropriate material. Do not forward virus warnings, chain letters, jokes, etc. to classmates or instructors. The sharing of anything considered pornographic is forbidden. Also, trying to sell classmates on your product or solicit any kind of business interaction is inappropriate in the classroom context.

Note: The instructor reserves the right to remove posts that are not collegial in nature and/or do not meet the guidelines listed above.

Response Time and Feedback: Please feel free to contact me any time via email. I will respond within 24 hours if possible but on later than 48 hours. I try to respond to the “Ask the Instructor” discussion board as soon as possible but no later than 24 hours. Please reserve the “Ask the Instructor” forum for questions that are related to class since replies may be beneficial for other students as well. All communication needs to be conducted through Blackboard and Truman State University resources.

Academic Integrity: In this class, I will treat you and your colleagues as burgeoning professional data scientists and statisticians. Statisticians and other scientists are expected to live up to a high ethical standard. Besides normal class expectations of avoiding cheating, lying, or plagiarizing, statisticians must ensure that they are fair to their clients, their subjects, and even their data itself. In this class, plagiarism is a thing we learn about as we grow as writers. Poor citations are counted against your grade, but failing to acknowledge the work that was done by someone else is considered cheating. Cheating and lying is generally a poor choice and will result in an F in this course and referral for campus-wide action. http://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx

It is Truman's expectation that assignments submitted will be the original work of the student, using proper citation when building upon the work of others. In cases where the use of third-party code is allowed, its author and source must be properly cited.

As with any online course, the temptation to cheat or plagiarize can be great. Under time constraints and pressure to complete, finding an "answer" online may seem like an easy solution, but students should keep in mind that doing so not only violates Truman's Honor Code, it defeats the purpose of taking this course. Enrollment in this course indicates a student's desire or employment requirement to actually learn the material presented, with the assumption that the student will be expected to demonstrate a working knowledge of the associated skills in the workplace. Students who take the time to do the coursework themselves ensure that they are fully prepared to perform similar tasks in the workplace where they will be judged by their peers and supervisors. This is the place for students to make mistakes and learn to do it right themselves.

Students found to be in violation of this policy may be removed from the course, and may receive further sanctions, up to and including removal from the program and referral to the Office of Student Affairs for additional campus-wide sanctions.

Disability Services: To obtain disability-related academic accommodations students with documented disabilities must contact the course instructor and the Office of Student Access and Disability Services (OSA) as soon as possible. Truman complies with ADA requirements. For additional information, refer to the Office of Student Access and Disability Services website at http://disabilityservices.truman.edu/ or contact them at 660.785.4478 or e-mail them at studentaccess@truman.edu.

Title IX: Truman State University, in compliance with applicable laws and recognizing its deeper commitment to equity, diversity and inclusion which enhances accessibility and promotes excellence in all aspects of the Truman Experience, does not discriminate on the basis of age, color, disability, national origin, race, religion, retaliation, sex (including pregnancy), sexual orientation, or protected veteran status in its programs and activities, including employment, admissions, and educational programs and activities.

Title IX prohibits sex harassment, sexual assault, intimate partner violence, stalking and retaliation. Truman State University encourages individuals who believe they may have been impacted by sexual or gender-based discrimination to consult with the Title IX Coordinator who is available to speak in depth about the resources and options.

Faculty and staff are considered "mandated reporters" and therefore are required to report potential violations of the University's Anti-Discrimination Policies and potential incidents of sexual misconduct to the Institutional Compliance Officer. The counselors at University Counseling Services are NOT mandated reporters, learn more at https://ucs.truman.edu/ or contact them at ucs@truman.edu, 660-785-4014 (after-hours crisis counseling: 660-665-5621).
For more information on discrimination or Title IX, or to file a complaint contact Dr. Lauri Millot, Institutional Compliance Officer, Title IX and Section 504 Coordinator, Violette Hall 1308, (660) 785-4354, titleix@truman.edu. The institution's complaint procedure, complaint form and other material is available at http://titleix.truman.edu.

Emergency Procedures: Truman students, faculty, and staff can sign up for the TruAlert emergency text messaging service via TruView. TruAlert sends a text message to all enrolled cell phones in the event of an emergency at the University. To register, sign in to TruView and click on the "Truman" tab. Click on the registration link in the lower right of the page under the "Update and View My Personal Information" channel on the "Emergency Text Messaging" or "Update Emergency Text Messaging Information" link. During a campus emergency, information will also be posted on the TruAlert website http://trualert.truman.edu/ For more information and materials, see http://police.truman.edu/emergency-procedures/