CS-401/L/501 - Programming for Data Science

Spring 2024 - Syllabus and Information

General Course Information

Course Description: Intensive programming course with a focus on solving problems in data science, specifically focusing on big data and dealing with different data formats. Students are introduced to data mining and machine learning algorithms with a focus on being able to use programming packages. Data mining and machine learning focus on algorithms that automate the process of discovering patterns in, and devising models for, large datasets.

Instructor: Mr. Luke B. May

Credit Hours:

- CS-501: 3.0
- CS-401/L: 3.0 + 1.0 = 4.0

Prerequisites: Graduate student or a grade of C or better in **CS 202**, **CS 303**, and **MATH 132**.

Course Website: https://cs.indstate.edu/~lmay1/courses/ (**bookmark this site**)

Required Texts: None

Required Materials for Class:

- Bring your laptop to class (or be prepared to use the classroom lab computers)
- Bring headphones or earbuds to class so that you may watch tutorial videos during lab time.
- Paper and a pen or pencil.

Required Software:

- Python 3
  - Packages (`numpy` , `pandas` , `matplotlib` , Beautiful Soup, Scikit-learn)
- Code Editor - Sublime Text Editor (you may use others like Kate, Notepad++, or VS Code, etc.)
- SFTP Client - FileZilla (transfer files to and from the server)
- Anaconda (Jupyter Notebook)

- **Optional and Recommended For Windows Users** - Windows Subsystem for Linux 2 (WSL2) - Installs the Linux kernel in Windows, providing access to a Linux terminal environment. There will need to be some installation overlap as Windows and Linux programs are treated separately, so you will need to install some things in both environments (Python for example). https://docs.microsoft.com/en-us/windows/wsl/install-win10

Optional/Recommended Texts and/or Resources:

- Probability and Statistics
  - *Think Bayes: Bayesian Statistics in Python* (2nd Edition) by Allen B. Downey
- Data Science
  - *Python Data Science Handbook: Essential Tools for Working with Data* (1st Edition) by Jake VanderPlas
  - *Python Data Cleaning Cookbook: Modern techniques and Python tools to detect and remove dirty data and extract key insights* (1st Edition) by Michael Walker
- Machine Learning
  - *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd Edition) by Aurélien Géron
  - Scikit-Learn: https://scikit-learn.org/stable/index.html

Review Materials:

- Learn X in Y minutes - Python 3: https://learnxinyminutes.com/docs/python3/
- *Automate the Boring Stuff* by Al Sweigart: https://automatetheboringstuff.com/

MS Teams Group: `cs401lm` (Course number, then a lowercase LM for Luke May)

MS Teams Join Code: Can be found on the Canvas course Home screen.

CS Server: `cs.indstate.edu`

Instructor Information and Office Hours

Name: Mr. Luke B. May

Email: Luke.May@indstate.edu

Office Location: Root Hall (RO), A-178

- Root Hall Offices Map: http://cs.indstate.edu/info/where.html.

Official Instructor Office Hours (RO A-178):

- **T/Th 12:30pm-1:30pm** or by appointment (online meetings with MS Teams is the preferred meeting method).
- Class related questions should be directed to your **MS Teams** class chat, and to ensure a prompt response use an @mention to your instructor so they get a notification (e.g. `@May` then press tab). Most of the time you can just message me on MS Teams and I will get back to you fairly quickly.

Instructor Appointments: https://cs.indstate.edu/~lmay1/schedule/

CS Server (cs.indstate.edu):

- Instructor Username: `cs401lm`
- Instructor/Class Directories
  - `/u1/h5/lmay1`
  - `/u1/class/cs401lm`

Final Exam / Project Due Date

Graduate students will have an additional final project worth 10% of their overall grade. ALl of the other graded items (including the final exam) will account for 90% of the final grade.

Wednesday, May 8, by 11:59pm

Course Outline

- Course Tools
  - Anaconda and Jupyter Notebooks
    - Installation, Configuration, and Workflow
    - Python 3 conda environments
    - Managing Python Packages
- Data Science Introduction
  - Capabilities and Limitations
  - Ethics, Biases, and Misinformation
- Math Topics/Review
  - Probability and Counting

    - Statistics - Distributions and Regressions
    - Vectors and Matrices
        - Matrix Multiplication, Dot Products, and Cross Products
        - Eigenvectors and Eigenvalues
- Python Topics/Review
    - Functional Programming
    - List Comprehensions
    - Iterators and Generators
        - `itertools` Package
    - Simple Data Structures
        - `list`s, `tuple`s, `set`s, `frozenset`s, and `dict`ionaries
        - `collections` Package
    - Regular Expressions
    - Data Science Packages and Tools:
        - `numpy`, `pandas`, `matplotlib`, Beautiful Soup, Scikit-Learn
- Collecting Data for Data Science
    - Scientific/Experimental
    - Surveys
    - Data APIs
    - Data Scraping
        - File Scraping in Linux: `grep`, `sed`, `awk`, etc.
        - Web Scraping: Beautiful Soup
- Cleaning Data and Data Consistency
    - Projects (subject to change)
        - Weather Data
        - COVID-19 Data
        - Census Data
        - Author Identification
        - Product Review Data Scraping Analysis
- Presenting Data (`matplotlib`)
    - Projects (subject to change)
        - Weather Data
        - COVID-19 Data
        - Census Data
        - Author Identification
        - Product Review Data Scraping Analysis

- Analyzing and Classifying Data
    - k-Nearest Neighbor
    - Decision Trees and Random Forests
    - Naive Bayes
    - Logistic Regression
    - SVM and ANN intro
- Support Vector Machines (SVMs)
- Dimensionality Reduction
- Deep Learning and Artificial Neural Networks (ANNs)
    - Intro to Scikit-Learn (potentially others)
- Big Data Analysis Projects (subject to change)
    - Author Identification
    - Temperature on COVID-19
- Machine Learning Projects (subject to change)
    - Character Recognition
    - Media Recommendation System
    - Mouse Movement Analysis
        - Users vs Bots or Artificial Algorithms
        - Unique User Identification

Learning Outcomes

Students will be able to:

1. Install and use an appropriate set of software tools to work on data science projects
2. Demonstrate an understanding and make use of functional programming techniques
3. Demonstrate an understanding of when to use programming concepts such as comprehensions, iterators, generators, regular expressions, and common data structures for data science
4. Demonstrate an understanding of many of the ethical concerns around data science and how to operate within solid ethical boundaries
5. Demonstrate an understanding of common data collection methods, and be able to implement simple file and web scraping techniques
6. Demonstrate an understanding of data cleaning techniques and how to perform data cleaning on simple data sets
7. Demonstrate an understanding of data analysis and classification techniques and when and how to implement them

8. Demonstrate an understanding of machine learning strategies and techniques (such as SVMs and ANNs) and how they work at a high level
9. Demonstrate understanding of Dimensionality Reduction at a high level, and why it is important to data science

---

Grading and Assignments

Student Responsibilities

- Consume assigned lecture material before lecture: books, videos, articles, etc.
- Attend lecture if in-person or or synchronous class, otherwise watch lecture material in a timely manner.
- Take and submit quizzes using the online quiz system.
- Complete labs following the instructions provided on the course website.
- Take the exams.
- Confirm that assignment grade sheets match Canvas grade book entries.

Course Grade Breakdown

Grade Item	Percentage of Total Grade
Quizzes	30%
Labs	30%
Midterm Exam	20%
Final Exam	20%

Quizzes: Each quiz will have questions of type *multiple choice*, *select all that apply*, and/or *short answer*. If not otherwise indicated, each answer will be worth 1 point. Quiz data is randomized on a per student basis. You may complete the quizzes at any time between when they are assigned and when they are due. In some cases, quizzes may also be timed. In this case, once you begin the quiz you will be required to complete it entirely before the time expires, at which point you will be locked out of the quiz. Once you begin, the timer starts (no exceptions), so make sure you have a stable internet connection before you begin. **Individual quizzes are not equally weighted**. The sum of all quiz points from each of the quizzes will be your final quiz score.

Labs: Labs are generally activities that require interaction with a computer system. They generally result in the creation of digital media of some kind. All labs will be turned in by submitting your work via the CS Server using your provided CS User Accounts.

Exams: Exams will be a combination of a quizzes and labs. Exams can be taken online at your leisure on the day of the exam. In-class sections may be required to take the exam in-person at the discretion of the instructor.

- Quiz portion - **The quiz portion is timed** and covers similar (if not the exact same) questions from previously assigned quizzes.
- Lab portion - The lab portion of the midterm is a lab activity that you should reasonably be expected to complete in about an hour or two. For the final the lab portion will likely be a larger project, and may take 3-4 hours, but you will have ample time to work on it. On the midterm you will usually have the whole day to complete the lab portion. For the final, you will usually have around a week.

A quiz or a lab will usually occur once a week, depending on the speed at which we cover the material. Start work early!

Late Work

Late quizzes will not be allowed to be turned in for credit. We will be using them as learning tools and going over the solutions immediately after they are due. I will offer amnesty if you miss one quiz, if and only if, you get a B- or better on all the other assignments. In this case I will replace the quiz percentage score with the lowest between your midterm and final, minus a 10% penalty.

Late labs will be accepted up to 4 days late, with a 10% penalty for each day late (weekends and holidays count as late days). If you do miss a lab, even if you miss the 4 day late work acceptance window, I highly recommend you still attempt to complete it on your own. The material builds on itself, so completing the previous labs will be beneficial to each new lab. Additionally, showing effort on the labs can often help me be more sympathetic when I grade, so do not just skip an assignment or problem if you are confused.

DO NOT MISS EXAMS - Late exams are generally not accepted. A rare exception may be made for extenuating circumstances evaluated at the discretion of the instructor. Some such circumstances may include a serious medical absence with supporting documentation, house fire/flooding, etc. On the rare occasion in which exam make-ups are permitted, in the case of a midterm exam, then the final exam will count double and be registered as your midterm grade as well. In the case of missing the final exam with justifiable reason, you may have an additional day or two to complete it, but because the semester is ending there isn't much more to be done. In extremely rare cases of serious issues, at the discretion of the instructor, the student can get an

INCOMPLETE for their grade, finishing when they are able. This requires the involvement and approval of several departments and is very unlikely to occur.

Students should begin assignments as soon as possible, preferably the day they are assigned. This should give you time to get help in case you have a problem, which is very common. Many of the assignments require deep thought and problem solving skills, which can take “time on the calendar”, not just “time on the clock”. That means spending 2 hours on 3 consecutive days may be more productive than trying to spend 6 hours at once on the assignment. This of course depends on personal characteristics and differs from one person to the next, so you may want to try out different strategies.

Graduate Students and Undergraduate Honors Students

Graduate students and undergraduate honors students will have to complete an additional end of term project. This project will not be required for other students, but may be completed for extra credit. The project will be significant in scope, but students will have a minimum of 2 full weeks to complete it. Details will be provided as we near the term's end. Grading for this project will be part of your final exam grade (20% of total grade), such that the final exam grade will be (30% multiple-choice, 30% standard final lab, 40% additional project).

Course Policies

This course follows standard CS course policies. In particular check the CS course policies related to - cheating/plagiarism, attendance, missing exams. See http://cs.indstate.edu/info/policies.html for details. The below policies are for this particular course.

Attendance and Illness Policy

Attendance is expected for in-person and synchronous online sections, however, it will not be directly tracked beyond assignment submission and electronic communications activity (chat/email). For in-person and synchronous online sections you should arrive to class or login to the meeting prior to the scheduled start time. Late arrivals are disruptive to those who arrived on time.

If you feel ill in any way, please do not come to class. There will be no attendance penalty. All in-person lecture material is recorded via Zoom and accessible through the Canvas Zoom tool. All course announcements and assignments are available via the course website.

If you are experiencing something that may prevent the completion of your assigned work, please email your instructor as soon as you are aware of the situation. Assignments will not be extended nor will they be allowed to be made up unless the situation is sufficiently debilitating and there is medical documentation to corroborate the condition. Any and all other issues preventing the completion of student work will be examined and evaluated at the discretion of the instructor. Most, if not all, will not warrant policy exceptions.

Work Ethic

This course should give you the tools for achieving competency over the given topic, but you should be doing much more than the assigned material in order to be successful. Employers expect more than minimum level of course work completion from a potential hire. Personal projects can help crystallize difficult concepts, and solidify your skills. These projects are excellent additions to portfolios, too, which are a critical component to most successful job interviews.

If you take this class seriously you should be spending between 1-2 hours per credit hour on course work (not including lectures). Generally the students who get A's in their CS courses (and have an easy time finding jobs) are the students who spend the appropriate amount of time on the course outside of the classroom. Not everyone will need to spend this much time and not all weeks will be the same, but you should plan on putting in whatever time it takes.

Note - In MOST cases (the overwhelming majority), your classes should be more important than your part-time job. You should think of your course work like you would think of work assigned by a high-paying employer.

Course Website and Announcements

The majority of this course will be run through the course website linked at the top of this document. Bookmark that page. The course website contains announcements, a schedule of due dates, course assignments, lecture materials, and even links to exams and projects. You should check this site daily to ensure that you do not miss assignments or content.

Announcements regarding the course will be posted under the *Announcements* section of the course website. Announcements may also be made during class (if applicable), via MS Teams, or via your ISU sycamores email account. You are responsible for being aware of announcements however they were communicated, so regularly check the course website, MS Teams, and your email. The *Announcements* section of the course website should be the most comprehensive list of any and all course activity, so check it regularly.

Canvas Course Management Software

This course uses the course management software called Canvas (https://indstate.instructure.com/). You should see this course listed under your courses for the current term in Canvas. Canvas is only used for 2 purposes in this course (potentially only 1).

1. Grade book (all courses)
    - Your grades on assignments and exams will be emailed back to you once they are graded, then the grades will be entered in Canvas. Go to this course in Canvas, then click on *Grades*. All course content (lecture material, assignments, tutorials, due dates, etc.) is kept and maintained on the course website. You are ultimately responsible for your own grade, so make sure these grade values match to reduce the risk of clerical errors.
2. Recorded lectures (only if your course has a concurrent in-person section)
    - In Canvas click on your calendar and locate the course meeting you would like to join, or click any course for recorded lectures. To join one live (for synchronous online sections or in-person sections), just click the *Join* button to join synchronously. If you want to see a past recording of a lecture (asynchronous people), click the *Cloud Recordings* tab, then select the recording. The passcode will be copied automatically to you clip-board. Paste in the passcode when prompted.

Device Usage Policies

Laptop Required for Course: Regular Usage

For the purposes of this course, it will be assumed that you are in compliance with the mandatory laptop policy of the University. You will be expected to bring your laptop and be ready to use it for every class period. Usage of the laptop must conform to the provisions of this course as laid out in this syllabus as well as the Code of Student Conduct.

- Exception: If you are comfortable using the CS lab computers effectively enough to complete all class activities, you will not be required to bring your laptop. You should disinfect the lab keyboards and mice before and after each use.

I encourage you to use your computer during class if you are using it to follow along with the examples that are being discussed. You should not check social media or work on other courses, other projects, etc. during class. Do not consume or share any inappropriate material at any time. Be professional so that you may become a professional.

Cell Phone Usage Policy: Occasional Usage

Turn all audible notifications off. You may only use cell phones for things related to the course work or for urgent communications. During lab coding time you may use headphones to listen to music **IF AND ONLY IF** no one else can hear it (no exceptions). If I can hear your headphones that means you are being disruptive, and if I have to ask you to turn down a device more than once, you may be asked to leave that day's lecture.

Headphones and Earbuds: Occasional Usage

You are expected to bring headphones or earbuds to class in case you need to watch tutorial videos during lab time without distracting others. You may listen to music during lab time provided it isn't audible to others. If you have to be asked twice to turn down the music you may be asked to leave.

Professionalism and Conduct

Instructors and students are your colleagues in this academic setting; treat them with kindness and respect. Any software platforms used for the class (MS Teams, Zoom, CS server accounts etc.) will be considered an extension of the classroom, so all policies on classroom conduct apply. Be courteous and professional. Harassment of any kind will not be tolerated and will result in severe consequences. Do not share explicit material (of any kind), and do not share content if you think there is a reasonable likelihood that it may offend someone else. Common sense "Not Safe for Work" rules apply.

The intentional or malicious use or modifications of systems, software, configurations/settings, to undermine another student's educational experience will not be tolerated and may warrant extreme academic consequences on par with plagiarism. Malicious tampering with user accounts, settings, or systems of students, instructors, or any other group or individual will be penalized severely. Unauthorized use/abuse of university resources is strictly forbidden and can result in extreme academic (and potentially legal) consequences. This could be considered anything outside of designated course work or faculty authorized activities. Examples of things to avoid: launching an email phishing campaign, mining cryptocurrencies, attempting to illegally access information systems, etc.

If you break the conduct policies, your access to required course software can and will be revoked. In that case, you will automatically forfeit any points on any assignments that required their use,

and, depending on the severity, you may be removed from the course with an F grade. Legal consequences may also apply, depending on the activity.

Official ISU Policy on Academic Integrity

As a student at Indiana State University you are expected to practice personal and academic integrity; commit your energies to the pursuit of truth, learning, and scholarship; foster an environment conducive to the personal and academic accomplishment of all students; avoid activities that promote bigotry or intolerance; choose associations and define your relationships with others based on respect for individual rights and human dignity; conduct your life as a student in a manner that brings honor to yourself and to the University Community; and discourage actions or behaviors by others that are contrary to these standards.

- Adopted by the Indiana State University Student Government Association April 17, 2002

Cheating and Plagiarism

Follow the standard CS Course Policies to determine what is and is not allowed on assignments.

ALL CODE not authored by you (even code from lectures and lab examples) **MUST HAVE A CITATION**!

Ask the instructor if you have doubts about what is considered cheating in this course or for a particular assignment. Copying work from external websites or tutorial videos is not acceptable without explicit permission from the instructor, or unless the assignment specifically instructs you to do so. For undergrads, a first offense will result in a zero grade on the assignment, and a second offense will result in failure of the course, and potential expulsion. For grad students, it’s an automatic course failure, and potential expulsion.

Artificial Intelligence

The developments around Artificial Intelligence (AI) synthesized text are in flux and the rules that are expressed in this syllabus may need to change on short notice; this may affect the contents of assignments, as well as their evaluation. Artificial Intelligence resources are widely available to generate text, images, code, and other media. The student assumes full responsibility for AI-generated materials; ideas must be attributed, and facts must be true. AI tools may only be used when expressly permitted by the instructor, and use must be open and documented.

Asking for Help

Make sure you ask for help sooner, rather than later, if you feel yourself falling behind or if you are struggling to understand any concepts. Addressing any problems as soon as possible will greatly improve your likelihood of success. **Do not wait until the end of the semester**; that will be far too late!

Your primary method of contact should be MS Teams course chats and/or direct messages to your instructor (@mention). You may use email, but I will usually respond quicker to MS Teams messages. MS Teams also offers an infinitely better experience when communicating code.

CS Unix Lab

Help is available via the CS Unix lab, where we have hired undergraduate and graduate students to act as tutors and to provide conceptual guidance to other students. **Student employees are not there to do your work for you**, they are specifically forbidden from doing so. They are there to help guide you through any concepts you may not be understanding, but they should not be doing your work. Do not expect them to help you indefinitely if you are not willing to put forth the appropriate effort. If you would like to schedule an in-person or online appointment, you may use the resources below:

- **Location**: Root Hall (RO), A-015 (basement, just west of west stairwell, first door on the left).
- **Wiki**: https://cs.indstate.edu/wiki/index.php/Unix_Lab_and_Help
- **Lab Worker Schedule**: http://cs.indstate.edu/info/labs.html

Grade Cutoffs

We try to design homework assignments and exams so that a standard cutoff for grades will be close to what you deserve. I make use of the generally accepted ISU grading scale used on Canvas:

Letter	Percent
A	94-100
A-	90-94
B+	87-90
B	84-87
B-	80-84
C+	77-80
C	74-77
C-	70-74

Letter	Percent
D+	67-70
D	64-67
D-	60-64
F	0-60

Our goal is that the different grades have the following rough meaning:

Grade	Meaning
A+/A	You are very well-prepared to use these skills in the real world.
A-/B+	You understand nearly everything and should be able to use this knowledge in other courses or a job.
B/B-	Most things you understand very well and a few you might not.
C+/C	Learned enough and have the minimum skills to move on in the subject.
C-/D+	You put some effort in, you understand many concepts at a high level, but you haven't mastered the details well enough to be able to use this knowledge in a practical way.
D-	You will normally not get an F if you attend 80% of the lectures, complete most of the assignments up through the end of the course, and get nearly half of the problems on the final exam correct.
F	Normally, students that get an F simply stopped doing the required work at some point.

Academic Freedom

"Teachers are entitled to freedom in the classroom in discussing their subject, but they should be careful not to introduce into their teaching controversial matter which has no relation to their subject."

The preceding comes from the American Association of University Professors statement on academic freedom. Though the entire statement speaks to many issues, it is this portion on the conduct of the course that is most relevant. This means that faculty have the right to conduct their class in a fashion they deem appropriate as long as the material presented meets the learning objectives laid out by the entire faculty.
http://www.aaup.org/AAUP/pubsres/policydocs/contents/1940statement.htm

---

University Resources

Student Outreach and Well-being

For help with academic and/or personal issues contact **Sycamores Care** (https://www.indstate.edu/student-affairs/sycamores-care). At Indiana State, we care for your overall well-being, and we want to help you get the care, referrals, and answers you need to ensure your success.

Americans with Disabilities Act Policy

Indiana State University seeks to provide effective services and accommodation for qualified individuals with documented disabilities. If you need an accommodation because of a documented disability, you are required to register with Disabled Student Services within the Center for Student Success.

Center for Student Success:

- 1st floor Normal Hall
- (812) 237-2700
- https://www.indstate.edu/services/student-success/cfss/student-support-services/disability-student-services

Statement on Non-Discrimination, Harassment, and Sexual Misconduct

Indiana State University is committed to inclusive excellence. To further this goal, the university does not tolerate discrimination in its programs or activities. Indiana State University Policy 923 strictly prohibits discrimination on the basis of: race, color, national origin, gender, age, sexual orientation, gender identity or expression, disability, veteran status, or any other protected class. Title IX of the Educational Amendments of 1972 in particular prohibits discrimination based on sex in any educational institution that receives federal funding. This includes sexual violence, sexual misconduct, sexual harassment, dating violence, domestic violence, and stalking. If you witness or experience any forms of the above discrimination, you are asked to report the incident immediately to Public Safety: (812) 237-5555 or to the Equal Opportunity & Title IX Office: (812) 237-8954. If you witness or experience any forms of the above discrimination, you may report to the Office of Equal Opportunity & Title IX.

Office of Equal Opportunity & Title IX

- Rankin Hall, Room 426
- (812) 237-8954

- https://cm.maxient.com/reportingform.php?IndianaStateUniv&layout_id=10
- ISU-equalopportunity-titleix@mail.indstate.edu

Title IX of the Educational Amendments of 1972 prohibits discrimination based on sex, including sexual harassment. Sexual harassment includes quid pro quo harassment, unwelcome verbal or physical conduct, sexual assault, dating violence, domestic violence, and stalking. With respect to sexual discrimination, instructors, faculty, and some staff are **required by law** and institutional policy to report what you share with them to the Equal Opportunity & Title IX Office. You do, however, have the option of sharing your information with the following **confidential resources** on campus that are not required to share:

- ISU Student Counseling Center: (812) 237-3939; Gillum Hall, 2nd Floor
- Women's Resource Center/Victim Advocate: 812-243-7272 (24 hours a day); HMSU 7th Floor
- Victim Advocate: (812) 237-3829; HMSU 7th Floor
- Associate Dean of Students/Respondent Advocate: (812) 237-3829; HMSU 8th Floor
- UAP Clinic/ISU Health Center: (812) 237-3883; 567 N. 5th Street

For more information about discrimination and the support resources to you through the Equal Opportunity & Title IX Office, visit this website: https://www.indstate.edu/equalopportunity-titleix. Please direct any questions or concerns to: Assistant Vice President for Equal Opportunity and Title IX Director; (812) 237-8954; Parsons Hall 223; ISU-equalopportunity-titleix@indstate.edu.
