CSCI 453: Lg Scale Data Analytics Viz

Jingnan Xie

2022 Spring

**E-mail:** jingnan.xie@millersville.edu  
**Web:** cs.millersville.edu/~jxie

**Office:** Roddy 142  
**Office Hours:**  
MF: 11:00AM – 11:55AM, 2:00PM – 2:55PM  
W: 1:00PM – 2:55PM

**Lecture:**  
MF 12:00PM – 12:50PM  
W 11:00AM – 12:50PM

**Classroom:** Roddy 140

---

**Course Description**

A practical introduction to data analytics, visualization, and blending theory. Students will learn about and apply various clustering algorithms and techniques for dealing with noisy data, use a distributed data analytics framework, complete laboratory assignments using version control, and enforce reproducibility by having all science easily sharable. Students will become familiar with modern data analytics methods and explore real-world data sets. Visualization of results will be a large component of the course through interactive and static frameworks.

**Course Prerequisites**

This course needs a solid foundation in Statistics and Linear Algebra. Topics like normal distribution, standard deviation, variance, covariance, matrix, matrix multiplication, determinant, vector and vector space, etc. are NOT covered in this course. If you forget some of the topics, please review them as soon as possible.

**Course Outcomes**

1. Create reproducible, explainable data science workflows
2. Use modern distributed Map-Reduce frameworks, such as Apache Spark, to analyze data
3. Implement parallel clustering methods
4. Develop strategies for overcoming common imperfections in real-world datasets
5. Apply visualization techniques to multi-dimensional data
6. Apply gained skills to extract insights from multi-dimensional, real-world datasets
Textbook

Textbooks are NOT required, but recommended. They are *Practical Data Science with Hadoop® and Spark: Designing and Building Effective Analytics at Scale*, by Ofer Mendelevitch, Casey Stella, Douglas Eadline, and *Data Analytics with Spark Using Python*, by Jeffrey Aven.

Course Policies

Responses

I will respond to emails within 24 **weekday** hours unless an exception is noted through email, D2L, or in class. Please note that this means if you email me over the weekend or the night before an exam or assignment submission, I am not guaranteed to respond. **Start labs when they are assigned.**

Announcements

I will frequently send announcements through email. I will also post new/additional material on D2L. Read over it by the date indicated on the announcement. Under inclement weather, due dates may be pushed back or changed at my discretion, so please pay attention to all announcements.

D2L (Desire2Learn)

**Link to D2L** I will primarily use D2L as the grade portal for classes. The submission portion will also be leveraged for homework assignments. Lecture material and notes are accessible through D2L also.

Office Hours

I hold office hours for your benefit. Please do not hesitate to show up to office hours! If you find that my office hours do not fit your schedule, let me know so we can arrange for a time that does work. Office Hours are handled using Zoom Waiting Room. So If you cannot join the meeting immediately, please just wait. I will be there.

My Expectations of Students

- Arrive prepared and on time for class
- Engage in active discussion during lecture
- Ask questions. Do not hesitate to clarify a concept
- Do your best work and be confident in your abilities
- Check email frequently for announcements and additional information
- Adhere to Millersville University’s Academic Honesty Guidelines
Lecture
Attendance of the lecture is mandatory. I encourage everyone to make an active attempt toward participating. There are times where many examples throughout the lecture are better suited to be done with pen and paper. Please refrain from using your computer for any other reason than note-taking for the class. I also do not expect to see any mobile phones in use during class. I will ask you to leave my class if I observe misuse of technology.

Laboratory
Attendance of any laboratory component is also mandatory. Assignments will be given and will most likely take more time to complete than the lab period. You are permitted to leave the lab period if and only if you have completed the assignment.

Grading Policy

- **10%** of your grade will be determined by your attendance.
- **35%** of your grade will be determined by laboratory assignments.
- **35%** of your grade will be determined by the final project.
- **20%** of your grade will be determined by two midterm exams (10% for each).
- Up to **2%** will be added to your grade at the professor’s discretion based on active participation during lecture and laboratory periods.

I will grade on a ten-point grading scale. I will round your grades. An 89.51 will be classified as a A-

<table>
<thead>
<tr>
<th>Grade</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>≥ 93</td>
</tr>
<tr>
<td>A-</td>
<td>≥ 90</td>
</tr>
<tr>
<td>B+</td>
<td>≥ 87</td>
</tr>
<tr>
<td>B</td>
<td>≥ 83</td>
</tr>
<tr>
<td>B-</td>
<td>≥ 80</td>
</tr>
<tr>
<td>C+</td>
<td>≥ 77</td>
</tr>
<tr>
<td>C</td>
<td>≥ 73</td>
</tr>
<tr>
<td>C-</td>
<td>≥ 70</td>
</tr>
<tr>
<td>D+</td>
<td>≥ 67</td>
</tr>
<tr>
<td>D</td>
<td>≥ 63</td>
</tr>
<tr>
<td>D-</td>
<td>≥ 60</td>
</tr>
<tr>
<td>F</td>
<td>&lt; 60</td>
</tr>
</tbody>
</table>

You must attempt all examinations and assignments to pass the course.

Labs
Labs will have a grading turnaround time of one (1) week from the due date. Lab grades can only be contested for up to one week from when the grade is posted on Desire2Learn. Three grace days will be provided throughout the semester without penalty. Otherwise, no late submissions will be permitted.

Exams
Exams will be graded by the next class (e.g. Tuesday-Thursday class with a test on Thursday will have the tests returned the immediately following Tuesday. Monday-Wednesday-Friday class with a test on Friday will have the tests returned the immediately following Monday). You do not get to keep your exams. Failure to return an exam will result in an updated grade of zero (0). I will hand them out in class and go over any answers, but I will collect them at the end of the class period.
University Policies

Academic Dishonesty Policy

Copying or extensive collaboration on assignments is not permitted and may result in failure of the course and expulsion from the University. You may discuss approaches to solving a problem, as long as the discussion remains above the level of detail expected for the course. You may also seek aid in resolving compiler messages. However, if you copy a code fragment verbatim, you are likely committing academic dishonesty. If you copy a code fragment and rename variables, you are likely committing academic dishonesty. *Obtaining a solution on the Internet or elsewhere and submitting it as your own work is plagiarism and will result in severe disciplinary measures.* Be sure you can explain every line of every program you submit. **Writing code is no different than writing a paper — if it was not your original idea, then you should not submit it as your own work.**

Title IX

Millersville University and its faculty are committed to assuring a safe and productive educational environment for all students. In order to meet this commitment, comply with Title IX of the Education Amendments of 1972, 20 U.S.C. §1681, et seq., and act in accordance with guidance from the Office for Civil Rights, the University requires faculty members to report to the University’s Title IX Coordinator incidents of sexual violence shared by students. The only exceptions to the faculty member’s reporting obligation are when incidents of sexual violence are communicated by a student during a classroom discussion, in a writing assignment for a class, or as part of a University-approved research project. Faculty members are obligated to report to the person designated in the University Protection of Minors policy incidents of sexual violence or any other abuse of a student who was, or is, a child (a person under 18 years of age) when the abuse allegedly occurred.

Information regarding the reporting of sexual violence, and the resources that are available to victims of sexual violence, is available at [http://www.millersville.edu/sexualviolence/index.php](http://www.millersville.edu/sexualviolence/index.php)

Counseling Resources

Students sometimes face mental health or drug/alcohol challenges in their academic careers that interfere with their academic performance and goals. Millersville University is a caring community and resources are available to assist students who are dealing with problems. The Counseling Center (717-871-7821) is an important resource for both mental health and substance abuse issues. Additional resources include: Health Services (871-5250), Center for Health Education & Promotion (871-4141), Campus Ministries, and Learning Services (717-871-5554).
Planned Course Schedule

Week 01, 01/17 - 01/21: Welcome and Overview, Jupyter Notebook
Week 02, 01/24 - 01/28: Python Basics
Week 03, 01/31 - 02/04: Intro to Data Analytics and Map + Reduce
Week 04, 02/07 - 02/11: Intro to Numpy and Pandas
Week 05, 02/14 - 02/18: Numpy and Pandas
Week 06, 02/21 - 02/25: Review, Exam 1
Week 07, 02/28 - 03/04: Data Cleaning
Week 08, 03/07 - 03/11: Data Visualization – Matplotlib
Week 09, 03/14 - 03/18: Data Visualization – Seaborn
Week 10, 03/21 - 03/25: PCA and Clustering Algorithms
Week 11, 03/28 - 04/01: PCA and Clustering Algorithms
Week 12, 04/04 - 04/08: Review, Exam 2
Week 13, 04/11 - 04/15: Basic overview of Apache Spark
Week 14, 04/18 - 04/22: Final project – Proposal
Week 15, 04/25 - 04/29: Final project – Implementation
Week 16, 05/02 - 05/06: Final project – Presentation

Subject to change