ATMS 517 (Online)

SYLLABUS

Data Science for the Geosciences (4 HRS)

Satellite image of the 2010-10-26 low pressure over the United States, NOAA, 2010

Instructor

Prof. Alicia Klees

Teaching Assistant Professor
Rm 3070, Natural History Building
1301 W. Green Street
Urbana, IL 61801
aklees@illinois.edu

Online Office Hours: TBD or by appointment (my schedule is quite flex this semester!)

Location and Times

Online, synchronous class session (you only attend ONE per week):
OB1: Tuesdays, 2:00-3:20 PM CT
OB2: Thursdays, 7:30-8:50 PM CT

Prerequisites

There are no prerequisites for this course, although prior familiarity with programming and/or quantitative analysis is helpful.

Course Description

Global-scale data collection efforts via remote sensing satellites, large multinational and multi-observational research campaigns with aircraft and ground-based observations, and the assimilation of station and model generated data into large, publicly available, multidimensional datasets provide a highly resolved window into the complexity of Earth system processes. For example, with well over 20 petabytes (PB) of data and counting, NASA's Earth observation data and the data observed and curated by NOAA's Earth System Research Laboratory will require new data science tools to unlock the underlying insights that can improve weather and climate forecasts to help feed the world, optimize green energy production, predict natural disasters, save lives, and more.

In this course, you will learn to use state-of-the-art programming tools in the Python language, and how to apply the best practices for analyzing data in the geosciences. Topics include the nature of data (collection, components, ethical use), data science principles, basic statistics, basic programming in Python, and using the Pandas, Xarray, and statistical packages in Python to carefully process, identify statistical relationships in, run linear regression models on, and visualize, a variety of datasets.

Learning Outcomes

Upon completing this course, you will be able to …

- Identify various characteristics of data, including the potential sources of errors, the process by which data were collected, and the importance of metadata

- Outline and complete the steps of the data science process, from careful dataset selection, to making ethical considerations, to conducting exploratory data analysis, to developing and analyzing (reproducibly) a scientific question, to visualizing data, using the NumPy, Pandas, Xarray, statistical, and visualization packages in Python

- Formally process, including cleaning, merging, smoothing, and/or interpolating, data using Python tools

- Analyze correlations and predict trends in data using simple machine learning models like linear regressions in Python

- Plot data in easy-to-interpret 1, 2 and 3-D visualizations using the Matplotlib package in Python

Texts

Required Textbooks

1. *(Free)* VanderPlas, J. (2016) *Python Data Science Handbook: Essential Tools for Working with Data*. Sebastopol, CA: O'Reilly Media. ISBN-10: 9781491912058, https://jakevdp.github.io/PythonDataScienceHandbook/
2. DeCaria, A. J. (2016) *Python Programming and Visualization for Scientists.* Madison, WI: Sundog Publishing. ISBN-10: 0972903380, https://sundogpublishingstore.myshopify.com/products/python-programming-and-visualization-for-scientists-2nd-ed (Brand-new 2021 edition; 2016 edition is fine!)

Supplementary Textbooks

1. (Free) Wilks, D., 2011. *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Oxford; Academic Press, 2011. Access through Univ. Illinois library:
2. (Free) VanderPlas, J. *A Whirlwind Tour of Python:* https://jakevdp.github.io/WhirlwindTourOfPython/

Course Structure

Asynchronous recorded lectures assignments, and other assessments are to be completed per the schedule outlined in the course calendar. There will be assignments and weekly timed quizzes to reinforce the course topics, as well as weekly in-module exercises to help you practice applying the new material. During our weekly synchronous class meetings, you will participate in discussions and activities, engage with additional, helpful examples of the material, and have your questions, and particularly challenging concepts from the week, addressed. As a crucial component of this course, you will define, plan, and conduct a semester-long data science project and present your results to our class. Your project should integrate several of the course topics into a cohesive approach to solving a geoscience-based problem of your choice, and it will be an opportunity for you to start building an employment portfolio showcasing your data science projects.

Computational Resources

Students will have access to Google Colaboratory for Python based assignments, or students may install the Anaconda Python build on their own computers (recommended). Course assistance and feedback will be available through interfaces on Moodle, and our course website will be available at https://learn.illinois.edu.

Assessments and Grading

Weekly Quizzes	10%
Weekly In-Module Exercises	20%
~ Weekly Homework Assignments	40%
Project	30%

Quizzes

Each week, you will take a quiz that tests your comprehension of the principles and coding techniques covered on the homework assignment and the week's lecture material. The deadline for submitted answers will be at 11:55 PM CT Mondays.

Lectures & In-Module Exercises

Each week, you will watch multiple lectures on your own time, and then complete a set of in-module exercises. These exercises are graded on completion and will evaluate your comprehension of new material. Equally as importantly, they will give you the opportunity to practice new coding techniques before using them on the more complicated (and graded in part on accuracy) homework assignments. The deadline for these in-module exercises is 11:55 PM CT Mondays. I will strive to give you quick feedback on these in-module exercises, so it is in your best interest to submit them ASAP once you've worked through the week's lessons.

Homework Assignments

There will be weekly homework assignments throughout the semester, with due dates always clearly posted. Solutions will be posted in a timely fashion past the due date. Your homework will be graded based on both completion and accuracy.

Data Science Project and Presentation

The purpose of the data science project is two-fold 1) to build your data science toolkit with an application relevant to your research and/or desired field of employment, and 2) to begin creating a data science project portfolio. As the semester progresses, I will provide a separate, detailed breakdown of the expectations for the project and the milestones for components of your project, but, to broadly summarize, you will:

- carefully select an interesting dataset in your field
- conduct a thorough exploratory data analysis (using a variety of statistical tools) on the data and formulate a scientific question
- process your data
- analyze your data using statistical techniques
- produce meaningful visualizations of your results
- interpret your results and make conclusions regarding your scientific question

What you will submit is a Jupyter notebook containing the code for, visualizations of, and descriptions of the above, as well as a screen-captured presentation of your work that I—and your peers—will review!

Grading Scale

A > 93.4 > A- > 90.00 > B+ > 86.67 > B > 83.4 > B- > 80.00 > C+ > 76.67 > C > 73.4 > C- > 70.00 > D+ > 66.67 > D > 63.4 > D- > 60.00 > F

Extra Credit

Extra credit opportunities may be available at the discretion of the instructor. Opportunities may include but are not limited to additional data analyses, a relevant essay topic of your choice, additional exam questions, etc. Extra credit opportunities will not exceed 3% of the course total points.

Course Policies

Contesting Assignment Grades

Balancing timely feedback with accurate grading is always challenging. I may make mistakes and I may miscommunicate on the finer details of the assignment grading breakdown. Therefore, I encourage you to petition me with a regrading request if you feel your grade does not reflect your work quality or if there is a discrepancy between the grading breakdown and your grade. All regrading requests must be made via email with a specific request (e.g., Please re-examine question 2, as I believe my code for the smoothing of my data was actually correct.) rather than a general request (e.g., Please regrade assignment 2, I feel that I deserve a better grade based on the exemplary quality of my work.)

Reporting Suspicions of Cheating

Our primary goal is to ensure a fair, respectful, and stimulating learning environment. Witnessing cheating is demoralizing and demotivating. If you suspect or have evidence of cheating in ATMS 517, we want to know. Students suspected of cheating will be referred to the College of LAS on a case dependent basis.

Student Accommodations

To obtain disability-related academic adjustments and/or auxiliary aids, please contact the course instructor, aklees@illinois.edu, and/or the Disability Resources and Educational Services (DRES) as soon as possible. To contact DRES you may visit 1207 S. Oak St., Champaign, call 333-4603 (V/TTY), or e-mail a message to disability@illinois.edu. Please also inform us if you are under evaluation by DRES. We can likely implement accommodations prior to your completed evaluation.

Academic Integrity

You are encouraged to work with your peers on the homework assignments and in-module exercises in this class, *but the work you turn in must clearly be your own.* ***If you do work with your peers on a homework or in-module assignment, please include all their names on the assignment.*** ***You are welcome to utilize Stack Overflow and other great online resources for coding hints, but you must***

make the code you submit your own and understand it. In the event that you plagiarize another person's work without citation or permission, your case will be submitted to the College of LAS in accordance with LAS protocols. The Student Code at the University of Illinois is a document that specifies your rights and responsibilities. The current version of this document is available on the web at: http://admin.illinois.edu/policy/code/.

COVID-19 Policies

While this particular course meets solely online, please be aware of the following COVID-19 related policies for any on-campus activities you may have. In order to implement COVID-19-related guidelines and policies affecting university operations, instructional faculty members may ask students in the classroom to show their Building Access Status in the Safer Illinois app or the Boarding Pass. Staff members may ask students in university offices to show their Building Access Status in the Safer Illinois app or the Boarding Pass. If the Building Access Status says “Granted,” that means the individual is compliant with the university’s COVID-19 policies—either with a university-approved COVID-19 vaccine or with the on-campus COVID-19 testing program for unvaccinated students.

Students are required to show only the Building Access Screen, which shows compliance without specifying whether it was through COVID-19 vaccination or regular on-campus testing. To protect personal health information, this screen does not say if a person is vaccinated or not. Students are not required to show anyone the screen that displays their vaccination status. No university official, including faculty members, may ask students why they are not vaccinated or any other questions seeking personal health information.

All students, faculty, staff, and visitors are required to wear face coverings in classrooms and university spaces. This is in accordance with CDC guidance and University policy and expected in this class.

Please refer to the University of Illinois Urbana-Champaign’s COVID-19 website for further information on face coverings. Thank you for respecting all of our well-being so we can learn and interact together productively.

Course Topics

COMPUTATIONAL ENVIRONMENTS

Jupyter Notebook
Remote Access to Data and Computers
Cluster Computing
Parallel Processing

PYTHON BASICS

Invoking Python
Python Help, Debugging Tools and Tips
Basics of Syntax + Using Packages

Data Objects
Data Collections
Reading in Data (text, csv, netcdf)
Plotting: 1D
Plotting: 2D
Basic Mathematical & Other Built-In Python Operations
Flow Control: If statements, For loops, While loops
Numpy Arrays: Creation, Indexing, Modification, Math, Comparisons, Characteristics

CONCEPTUAL DATA SCIENCE BASICS

Data Types, Collection & Errors
Key Components of Data
Where to Find & How to Select Data
Background and Overview of Data Science Process
Ethical Considerations When Using Data
Reproducibility in Data Science
Review of Basic Statistics
Exploratory Data Analysis
Data Processing Techniques (conceptually)
Data Analysis Techniques (conceptually)
Good Practices for Creating Visualizations
Good Practices for Interpreting Visualizations

PROCESSING AND ANALYZING SCIENTIFIC DATASETS IN PYTHON

Basics of Xarray for Netcdf/spatial data
Basics of Pandas for CSV/database-type data
Quality Controlling Data
Handling Missing Data
Decoding Time
Merging Data
Grouping Data
Smoothing Data
Interpolating Data
Correlations
Evaluating Distribution of Data
Anomalies & Other Transformations
Linking Temporal and Spatial Data
Univariate Linear Regression (Including Evaluation!)
Multivariate Linear Regression
More Advanced Machine Learning Techniques (as time permits)
Visualizing and Quantifying Uncertainty in Data