ATMS 597 / GEOL 593
SIMLES: Statistical Inference and Machine Learning for Earth & Environmental Sciences

Instructor: Cristian Proistosescu (Assistant Professor of Atmospheric Sciences & Geology)
E-mail: cristi@illinois.edu

Logistics:
MWF 9-9:50 am. Room: NHB 2020.
4 Credit Hours. Letter Grade.
Office hours: TBD.

Course description: This course will provide students with data science tools directly applicable to graduate research in Earth and Environmental Sciences. Students will learn how to use data in order to produce estimates, draw inferences, and make predictions. The main topics covered will be: 1. Inferential statistics: parameter estimation and hypothesis testing; 2. Machine learning approaches to prediction; and 3. Model-data fusion.

Emphasis will be placed on mixing *conceptual understanding*, *practical applications*, and *domain expertise*. Methods will be introduced using idealized synthetic data to help build intuition, then applied to observations and measurements. With a few pedagogical exceptions, we will use ready-made tools available in the high-level programming python language and its libraries. Emphasis will *not* be placed on either rigorous mathematical treatment, or algorithm implementation (i.e. tool development).

Inclusivity: The effectiveness of this course is dependent upon the creation of an encouraging and safe classroom environment. Exclusionary, offensive or harmful speech (such as racism, sexism, homophobia, transphobia, etc.) will not be tolerated and in some cases will be subject to University harassment procedures. We are all responsible for creating a positive and safe environment that allows all students equal respect and comfort. I expect each of you to help establish and maintain and environment where you and your peers can contribute without fear of ridicule or intolerant or offensive language.

Prerequisites: Linear algebra and calculus. Basic programming experience in either Python (preferred), Matlab, or R. A previous course on statistics or probabilities is helpful but not required. Contact instructor if you have any questions about prerequisites.

Course Objectives:

- **Factual**: You will acquire fundamental knowledge of mathematical, statistical and machine-learning methods for the analysis of earth and environmental science data. You will become familiar with the terminology, the statistical procedures, and expected outcomes.
- **Procedural**: You will learn how to apply specific methodologies to the analysis of earth and environmental data, including the practical, hands-on procedures for managing data and implementing these methods and approaches in a high-level programming language. You will be able recognize and remove errors associated with data or the implementation of your procedures ('debugging').

- **Conceptual**: You will develop an understanding of available tools and when to best or appropriately apply them You will cultivate a first-order understanding of the motivations, advantages, and disadvantages for different procedures and how uncertainties in the underlying data and methods potentially propagate through your analyses.
- **Metacognitive**: You will recognize the potential and limitation of statistical and data analytical methods with respect to the constraints from the underlying physical, deterministic processes you seek to explore. You will be able to identify reasonable (and unreasonable) conclusions from your analyses, based on *domain expertise*. You will develop an enhanced recognition of how potential biases – including both methodological as well as cognitive – enter into statistical analyses of both deterministic and stochastic systems and inference based on these results.

Topics Covered:

Fundamental concepts:

1. Basics: Probabilities and random variables, Monte Carlo, correlation & covariance, regression.
2. Inference: Parameter estimation & hypothesis testing.
3. Prediction: loss function & optimization; training, testing, and validation; bias variance trade-off; regularization; introduction to neural nets.
4. Model data fusion: Hierarchical Bayes and Kalman filters

Survey of select topics:

5. Unsupervised Learning: Principal Component Analysis & Empirical Orthogonal Functions; Cluster Analysis.
6. Resampling methods: Cross validation, bootstrap.
7. Space-time methods: EOF analysis; Spectral Analysis.
8. Select machine learning approaches: Decision trees, random forests, Recurrent Neural Nets, Convolutional Neural Nets.

Format:

Data science methods are best learnt by doing. Thus, the class will have a flipped flavor: you will use assigned reading and example code (provided as jupyter notebooks) to gain a hands-on understanding of the material.

Class-time will be devoted to a mixture of activities.

- Lectures: In depth discussions of select aspects of the material.
- Paper discussions: Student-led discussions of research papers showcasing earth and environmental-science applications.
- Discussions of class projects: each week one or more of the student groups will have a chance to talk through their course-project and get feedback from colleagues and the instructor. Such feedback and discussions are essential, since any data analysis project involves subjective choices that need to be carefully thought through.

Assessments:

Assignments: 40%

Participation & Discussions: 20%
Class Project: 40%

You will work in groups of 3-4 student on a semester-long class project. The project will involve applying one or more of the methods introduced in the class to a dataset directly applicable to your research area. At the end of the course you will give a brief presentation, and submit a written report.

Readings:

No single good textbook exists. A variety of course notes and example Jupyter notebooks will be provided by the instructor.

The most relied-on textbook will be:

- James, Witten, Hastie, Tibshirani: An introduction to statistical learning
    - (freely available online - https://www.statlearning.com/)

Selected reading will be assigned from several other textbooks. All of these are freely available either online or through the UIUC library:

- Hastie, Tibshirani, Friedman: The elements of statistical learning : data mining, inference, and prediction. [link]
- Ross, Introduction to probability models [UIUC library link]
- Emile-Geay: Data Analysis in the Earth & Environmental Sciences [link]
- Von Storch and Zwiers: Statistical Analysis in Climate Research [UIUC library link]
- UW Objective Analysis Course notes (provided by instructor)