CAS PY 191S 1310: RISE Practicum in Data Science

Course Description: This six-week course is for rising senior high school students and gives students the opportunity to learn formal statistical thinking that can be used to carry out meaningful research as a budding computational scientist; specifically, we emphasize computational statistics as a lively field where new statistical methods are created nearly every day. Making use of data from the neurosciences, environmental science, physics, economics, and urban planning, students will learn essential statistical and coding skills. This includes a wide set of computational modeling approaches, uncertainty quantification skills, and frameworks for reproducible scientific computing. Additionally, students will also learn valuable perspectives in scientific rigor and ethics as a data-informed practitioner. To earn a certificate of completion for the course, students will work in groups to develop, complete, and present a computational project. In one of four areas of focus with support of the teaching staff:

1) **Time Series Analysis and/or Prediction:** Apply statistical signal processing and forecasting techniques to analyze and predict temporal data.
2) **Categorical and Count Data Analysis:** Perform statistical analyses of experimental or observational datasets that don't lend themselves to time-based analyses, particularly those involving categorical data, counts, proportions, or contingency tables. Conduct comparative analyses and inference using hypothesis tests, regression modeling (including generalized linear models), and estimation methods in frequentist and/or Bayesian frameworks.
3) **Scientific Software Package Development:** Implement a statistical method or algorithm in Python as a reusable and maintainable software package. Include thorough documentation, robust unit tests, and comprehensive user tutorials.
4) **Teaching a Statistical Technique:** Create an interactive and pedagogical Jupyter Notebook lesson designed to explain a specific statistical concept, complete with narrative explanations, practical examples, and hands-on exercises for learners.

Reference Text: All lectures will be self contained and thus no textbook is required. However I will pull heavily from the following texts:

Kass, R. E., Eden, U. T. & Brown, E. N. *Analysis of Neural Data*. (Springer New York, 2014). doi:10.1007/978-1-4614-9602-1.

Igual, L. & Seguí, S. *Introduction to Data Science*. (Springer International Publishing, Cham, Switzerland, 2024).

Montgomery, D. C., Jennings, C. L. & Kulahci, M. *Introduction to Time Series Analysis and Forecasting*. (John Wiley & Sons, Nashville, TN, 2024).

Brownlee, J. *Introduction to Time Series Forecasting With Python*. (Jason Brownlee, 2017).

Required Equipment: A computer running a recent version of Windows or Mac is required for this course. All software used in this course is either freely available (open source) or will be made available by Boston University. We will install all the software required on the first day of our lab.

Location:

"CDS" refers to the Center for Computing & Data Sciences Boston University, 665 Commonwealth Ave, Boston, MA 02215. This is the large glass "Jenga-like" building. B62 is in the basement down one flight of stairs in the center of the building or down one flight on the elevator.

Activity	Location
Morning Lecture (Tuesday - Friday): 9:30 to 11:30 am	CDS B62
Computational Lab Lecture and Lab (Tuesday - Friday): 1 - 5 pm	CDS B62

Teaching Staff:

Patrick F. Bloniasz (pblonais@bu.edu**)** is the morning lecturer for the Practicum and a Ph.D. candidate in computational neuroscience, working in the labs of statistician Dr. Emily P. Stephen and biomedical engineer Dr. Anna Devor. His research focuses on the development and application of statistically rigorous methods for neural data analysis, specializing in stochastic (point-process) modeling, measurement theory for biophysical signals, and the design of novel estimators for spectral, temporal, and connectivity features in electrophysiological recordings. He currently has a National Science Foundation grant developing statistics to bridge local network information (i.e., spiking data) to electrocorticography (ECoG) during propofol anesthesia-induced unconsciousness.

Eugene Pinsky (epinsky@bu.edu) is the Computational Lab Instructor. Dr. Eugene Pinsky first joined the faculty of Boston University in 1986, after earning his doctorate at Columbia University. He was an assistant professor of computer science in the College of Arts & Sciences until 1993, when he left BU and gained extensive industry experience designing computational methods to analyze and monitor market risk at multiple trading and investment firms, including Bright Trading, F-Squared Investments, and Harvard Management Company. He has also applied his expertise in data mining, predictive analytics, and machine learning to video advertising at Tremor Video. Dr. Pinsky's areas of expertise include pattern recognition, clustering, regression, prediction, factor models, support vector machines, and other machine learning algorithms and methods for data analysis; multi-dimensional statistical data analysis of large time-series data; data mining and predictive analytics to uncover patterns, correlations, and trends; algorithmic trading, pricing models, financial modeling, risk, and portfolio analysis; Python, R, C/C++, VBA, Weka, MATLAB, and MySQL; Big Data technologies and visualization (Hadoop/Hive, AWS, Tableau); design and implementation of software tools for quantitative analysis; and professional curriculum and course development.

Tharunya Katikireddy (tkatiki@bu.edu) is a Teaching Fellow for the Practicum and a master's student in Applied Data Analytics at Boston University. Her work spans statistical modeling, machine learning, and deep learning, with research focused on currency trend analysis using time-series decomposition and Benford's Law, as well as real-time Morse code decoding with LSTM networks. She is co-authoring a symbolic Python textbook that leverages visual metaphors to simplify programming education. Previously, she worked at Salesforce, where she optimized data pipelines and analytics dashboards to enhance operational workflows. Her broader interests include model interpretability, ethical AI, and efficient architectures for generative systems.

Tejovan Parker (tejovanp@bu.edu) is a Teaching Fellow for the Practicum. Learn more about Tejovan at https://www.bu.edu/cds-faculty/profile/tejovan-parker/ and his personal website https://www.tejovanparker.com/.

Zhengyang Shan (shanzy@bu.edu) is the Teaching Fellow for the Practicum and a Ph.D. student in Computing & Data Sciences, working under the guidance of Dr. Aaron Mueller. Her research lies at the intersection of natural language processing and machine learning, with a growing focus on interpretability and bias evaluation. She is currently investigating causal and contrastive methods for analyzing demographic representations in large language models (LLMs), as well as developing personalization-aware retrieval techniques for long-context generative systems.

Kevin Quinn (quinnk@bu.edu) is a Teaching Fellow for the Practicum and a Ph.D. student in computing and Data Sciences, working with Dr. Mark Crovella and Dr. Evimaria Terzi. Working in the areas of machine learning and algorithmic data mining, he is most interested in the design and of interpretable machine learning models, specifically for unsupervised clustering problems. As a student interested in interdisciplinary science, he has also applied his work to studies of covid-19, climate data, and voting systems.

Class Policies:

To ensure that everyone gets the most out of their experiences in the course, please adhere to the following policies.

- Treat fellow students and staff respectfully
- Please come prepared for all lectures and labs.
- Do not use a phone during class. If you need to use your phone in case of emergency, please step outside of the classroom to do so.
- Please be on time for lectures and labs each day. If you are more than 10 minutes late, your instructor must notify the Summer Term Office and it will be marked against you. Do not leave class or lab early unless the entire class has been dismissed by the instructor. If you have a mandatory appointment, you must notify the Program Advisor ahead of time (24 hours).
- Feel free to ask questions during lectures and labs.
- If you have any accommodations that are required for accessibility and have not been addressed, please let the instructor know so that we can ensure you have a productive experience in the course.

- Please feel free to let the instructors know at any time during the course if you have feedback. An anonymous course evaluation survey will also be provided at the end of the course.

<u>Course Schedule:</u>

The following schedule is scheduled to change at the discretion of the morning or afternoon instructor based on pacing of students or trouble areas for a given cohort.

WEEK	DATE	(#) LECTURE	LAB
1	Introduction to Data Science & Python Fundamentals		
	July 1st	(1) Introduction to Data Science, Course Overview	Installing Software, Github Account Creation, ensure morning notebook runs in IDE
	July 2nd	(2) Programming fundamentals in Python for analysis	Loops, Github repo cloning, unit testing
	July 3rd	(3) Python Functions, Data Handling, and Documentation	Data Parsing and Scripting
	July 4th	NO CLASS	NO CLASS
2	Data Wrangling, Visualization, and Reproducibility		
	July 8th	(4) Object-Oriented Programming in Python and Package Development Intro	Software Packaging & More Testing Fundamentals
	July 9th	(5) Exploratory Data Analysis	Software Packaging & Testing Fundamentals
	July 10th	(6) Data Cleaning and Wrangling	Creating interactive data dashboard for simple visualizations
	July 11th	(7) Data visualization best practices	Visualizing Data for Real World Insights
3	Core Statistical Foundations		
	July 15th	(8) Introduction to Probability Theory	Exploring probability (random variables) through simulation and theory
	July 16th	(9) Point Estimation and Confidence Intervals	Introduction to analytic and bootstrap confidence intervals
	July 17th	(10) Introduction to hypothesis testing part I	Introduction to hypothesis testing part II

	July 18th	(11) Introduction to Linear Regression and Generalized Linear Models (GLMs)	Practice end-to-end data analysis – cleaning → EDA → inference → modelling – on a messy but instructive synthetic data set.
4	Time Series Analysis		
	July 22nd	(12) Foundations of Time Series Forecasting	Instructor Research Demo (Pinsky and Katikireddy) and Project Work
	July 23rd	(13) Data Preparation & Feature Engineering	Instructor Research Demo (Parker) and Project Work
	July 24th	(14) Time Series Visualization and Feature Engineering pt. 2	Instructor Research Demo (Shan) and Project Work
	July 25th	(15) Times Series Decomposition	Instructor Research Demo (Quinn) and Project Work **(Homework: draft abstracts)**
5	Predictive Modeling		
	July 29th	(16) Introduction to Supervised Learning	Project Work
	July 30th	(17) Introduction to Unsupervised Learning	Project Work
	July 31st	(18) Model Evaluation, Selection, and Validation	**Project Work (Abstracts due today @ 5pm)**
	August 1st	(19) Predictive Modeling End-to-end (Hack-a-thon)	Project Work
6	Project Communication & Special Topics		
	August 5th	(20) Effective Presentation Design & Storytelling	Anatomy of a good readme file (documentation) – **Posters due today @ 5 pm!**
	August 6th	(21) Crash Course – State Space Modeling or	Crash Course – Large Language Models (LLMs)
	August 7th	(22) Crash Course – Digital Text Analysis	Crash Course – TBD
	August 8th	Final Discussion **Poster Practice**	**Final Poster Presentation to the community 1- 3 pm**