Office / contact info
CCDS 1536
bcleary@bu.edu

Course
CDS DS 596

Title
Foundations of Biological Data Science

Description
This course establishes a foundation in applied statistics and data science in biology for those interested in pursuing data-driven research. Students will develop fundamental and transferable computational and statistical skills for critically thinking about and using data in biology. The course will develop the foundations of and illustrate major methods applied in modern biological problems and data sets. Data science topics will include data wrangling, exploration and visualization, data resampling, clustering, dimensionality reduction, manifold learning, optimal transport, random forests, kernel methods, and latent space models. The course will explore application of these methods in the context of gene regulatory networks, genotype to phenotype mapping, chromatin structure analysis, single-cell biology, and quantitative biological imaging. The python programming language is extensively used to explore methods and analyze data.

Prerequisites
{CDS DS 110, 120, 121, and 122 (or equivalent courses)} and {CAS BI 105 or BI 108 or by permission of instructor}.

Learning outcomes
Students will develop foundational skills for performing and critically analyzing data-driven research in biology.

Assignments and grading structure
Final grades will be determined on the basis of three mid-term assignments (25% each), a final project (15%), and in-class participation (10%).

Each of the assignments will involve exploratory analysis of a published dataset using techniques covered in class. Students will present their analysis in a brief written report accompanied by a jupyter notebook with computational work. Reports will include factual findings as well as a critical analysis of the data, methods, and interpretation of results.

The final project will require students to develop a proposal for a data-driven research project in biology. Proposal topics could include the development of a new computational approach, application of existing methods to new or under-explored data, or analysis of multiple methods applied to common, benchmark datasets.

Course policies
Lectures will involve group discussion of the presented material, and in-person attendance is expected. In exceptional circumstances we will accommodate remote attendance with a hybrid class (in-person and zoom).

The BU Academic Code of Conduct is here: https://www.bu.edu/academics/policies/academic-conduct-code/. All students are required to familiarize themselves with this code, its definitions of misconduct, and its sanctions. Students should especially familiarize themselves with the section on plagiarism.

All written and computational work in this course must be original to you. If you consult outside texts, or other forms of assistance, cite these sources in the proper format. This pertains to all external sources (books, journals, lectures, web sites, AI). We are required to report all suspected cases of plagiarism to the Academic Dean for review.

Academic integrity in computing coursework has some special aspects. Please review the examples of plagiarism as provided by the BU Computer Science department.

Logistics

Lectures will take place Monday and Wednesday afternoons (2:30-3:45) with discussion sections Wednesday mornings (11:15-12:05) used to review material from the previous week and address outstanding questions. In-class examples of code will be presented in python and computational assignments are expected to be completed in jupyter notebooks. Unless otherwise noted assignments are due by 5pm EST.

Outline of lectures

The preliminary schedule of topics and assignment deadlines is as follows:

- Lecture 1: Course introduction
- Lecture 2: Intro biology
- Lecture 3: Exploratory data analysis
- Lecture 4: Data wrangling
- Lecture 5: Genomic technologies
- Lecture 6: Data resampling and bootstrap testing
- Lecture 7: Permutation and multiple hypothesis correction
- Lecture 8: Single-cell technologies
- Assignment 1 due (2/16/23)
- Lecture 9: Clustering
- Lecture 10: Differential expression
- Lecture 11: Dimensionality reduction
- Lecture 12: Manifold learning
- Lecture 13: Optimal transport
- Lecture 14: Lineage tracing technology and methods
- Lecture 15: RNA velocity
- Assignment 2 due (3/22/23)
- Lecture 16: Random forests
- Lecture 17: Kernel trick
- Lecture 18: Vector field inference (Euler's method)
- Lecture 19: Perturbation screening
- Lecture 20: Regularized regression
- Lecture 21: Latent space models
- Lecture 22: Imaging technologies 1 (morphology and functional imaging)

- Assignment 3 due (4/19/23)
- Lecture 23: Imaging technologies 2 (gene expression measurements)
- Lecture 24: QTLs and heritability
- Lecture 25: Structural causal models
- Lecture 26: Final review
- Final project due (5/6/23)