May Institute on computation and statistics for mass spectrometry and proteomics

April 30, 1:30pm – May 2, 5:00pm 2025. **Applied Python for data analysis and beginners statistics**

Venue:
Northeastern University main campus, West Village H, first and second floor rooms: 110, 108, 210A/B, 212. Please see the annotated map.

Lead instructors:
Benjamin Gyori, b.gyori@northeastern.edu
Arzu Tuğçe Güler, a.guler@northeastern.edu
Jeremy Muhlich, jeremy_muhlich@hms.harvard.edu

Guest speakers:
Olga Vitek

Teaching assistants:
Nihira Golasangi, Haohang Yan, Thomas Lim

Description:
This course starts with an introduction to Python programming and then builds on that to introduce the most commonly used Python packages for data analysis. The schedule is organized such that Python modules are interleaved with modules on statistics shared across the R and Python courses. Overall, people taking this course can expect to gain a practical understanding of Python programming and the most widely used packages for data analysis, as well as fundamental statistical methods.

This course is strongly recommended for everyone who is planning to take the course on “Interpretation of proteomic experiments in the context of biomolecular networks” (May 7-9).

The course begins with a general introduction to Python, its fundamental data structures and built-in packages. It then covers topics necessary for Python applied to data analysis through specific packages that are most widely used in the community: data processing and analysis (pandas), data visualization (matplotlib, seaborn), and network construction and analysis (networkx).

Each topic will include a mix of lecture and hands-on examples with live coding.

References:
The course will combine lectures and practical hands-on exercises.

Target Audience:

Target audience are experimental scientists, bioinformaticians, computer scientists, data scientists, statisticians or engineers, who either (i) have no substantial background in Python and could benefit from a fast-paced introduction that then leads into data analysis packages or (ii) have some exposure to Python but would like to get a more structured introduction and learn about specific packages tailored to data analysis.

Course materials:

Applied Python for data analysis and beginners statistics

Tentative schedule (updated 1/30)

April 30, 2025

12:30 p.m. Registration
1:30 p.m. **Introduction to Statistics**, Olga Vitek
3:00 p.m. Refreshments
3:30 p.m. **Introduction to Python, setting up a Python environment, installing Python packages**, Jeremy Muhlich and Arzu Tuğçe Güler
5:00 p.m. Q&A and adjourn

May 1, 2025

9:00 a.m. **Python coding essentials: data structures and built-in packages**, Jeremy Muhlich and Arzu Tuğçe Güler
10:30 a.m. Refreshments
11:00 a.m. **Principles of statistical inference**, Olga Vitek
12:30 p.m. Lunch on your own
1:30 p.m. **Data processing with *pandas* and interacting with web services** Jeremy Muhlich and Arzu Tuğçe Güler
3:00 p.m. Refreshments
3:30 p.m. **Data visualization using *matplotlib* and *seaborn***, Jeremy Muhlich and Arzu Tuğçe Güler
5:00 p.m. Q&A and adjourn

May 2, 2025

9:00 a.m. **Experimental design. Class discovery and class prediction**, Olga Vitek
10:30 a.m. Refreshments
11:00 a.m. **Working with networks using *networkx***, Benjamin Gyori
12:30 p.m. Lunch on your own
1:30 p.m. **Github repositories**, Kylie Bemis (merged with Intermediate R)
3:00 p.m. Refreshments
3:30 p.m. **Practice & exercises**, Benjamin Gyori
5:00 p.m. Wrap-up and adjourn

Link to participant lunch order pickup

If you would like us to pick up your prepaid lunch order from Tatte or Anna's Taqueria, fill this form: https://forms.gle/5E8apFjvjW3qZhgQA

Course evaluation

Please help us improve the program in the future by filling in this form:
https://forms.gle/s6bbgAUGhdDy6GjG9