INFO 6105 Data Science Engineering Methods and Tools

Course Information

Course Title: INFO 6105 Data Science Engineering Methods and Tools
Course Number: INFO6105
Term and Year: Spring 2024
Credit Hour: 4
CRN: 39506
Course Format: On- ground (Traditional)
Location: Ryder Hall 157
Lecture Hours: Saturday 13:00 - 16:30 ET
Student Hours (in person): Saturday 12:30-13:00 & 16:30-17:00 ET | By Appointment
TA Student Hours (via Zoom): Tuesdays and Fridays 10:00-12:00 ET

Instructor Information

Full Name: Hong Pan, Ph.D., Professor Pan (He/His/Him)
Email Address: hong.pan@northeastern.edu
TA: Goutham Kanahasabai
Email Address: kanahasabai.g@northeastern.edu

Instructor Biography

Hong was born to a family of educators. He attended Shanghai Jiao Tong University, where he studied Biomedical Engineering, and then joined Purdue University in the U.S. for his PhD program in Electrical and Computer Engineering. After obtaining his PhD, he first joined Cornell University Medical College as a faculty member where he conducted and overseen technical, analytic and engineering aspects of human in vivo functional and molecular neuroimaging research and trained multidisciplinary students, research fellows and clinician scientists; and then moved to Harvard Medical School as a faculty member where he further his invention to innovation technology transfer journey in data science applications for medical imaging.

He led data science efforts as the subject matter expert on 20+ federal and institutional projects, translating AI/ML algorithms and advanced statistics capabilities for developing statistical, data-driven diagnostic tools for guiding the treatment of brain disorders, created best practice approaches for optimized data acquisition, data science solutions for biomarker discovery, automated analytics and informatics pipelines based on functional neuroimaging methodology, resulting in 4 patents and a spin-out startup, earned Mass General Brigham Excellence in Innovation Award 2 times and Brigham and Women's Hospital's Pillar Award in Research & Innovation, with 60+ journal publications.

Course Prerequisites

Graduate level INFO 5100 Minimum Grade of B- or Graduate level CSYE 6200 Minimum Grade of B-
The instructor's further advice is as follows: Understand object-oriented programming in Python or R

Course Description

Introduces the fundamental techniques for machine learning and data science engineering. Discusses a variety of machine learning algorithms, along with examples of their implementation, evaluation, and best practices. Lays the foundation of how learning models are derived from complex data pipelines, both algorithmically and practically. Topics include unsupervised learning (clustering, dimensionality reduction, recommender systems) and supervised learning (parametric/nonparametric algorithms, support vector machines, kernels, neural networks, deep learning). Based on numerous real-world case studies.

Standard Learning Outcomes

By completion of the course, you will:

1. Gain a basic understanding of statistical machine learning methods and know how to apply these basic techniques in real-world problem solving for various tasks such as prediction, regression , grouping, etc.
2. Be able to understand the fundamentals of study designs, use graphical and other means to explore data, build and assess basic statistical machine learning models, employ a variety of formal inference procedures, and draw appropriate scope of conclusions from the analysis.
3. Be able to write clearly, speak fluently, and construct effective visual displays and compelling written summaries, to communicate statistical findings and results.

Required Tools and Course Textbooks

Textbooks:

- **An Introduction to Statistical Learning** by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. The second edition of this book, **with Applications in R (ISLR)**, was released in 2021; The **Python edition (ISLP)** with additional team member Jonathan Taylor, was published in 2023. Each edition contains a lab at the end of each chapter, which demonstrates the chapter's concepts in either R or Python, and is available online for free at https://statlearning.com . You can also purchase a hard copy from Springer.
- **Multivariate Data Analysis with R** by Nick Fieller (2011) with Appendices of Machine Learning: Clustering Analysis, Tree-based Methods, Neural Networks, Kohonen Self-organizing Maps. https://drive.google.com/file/d/1noCA4MQnvxaAHVEglYrgR-L6lteBmWWW/view?usp=sharing
- **Univariate and Bivariate Statistics:** https://openstax.org/details/books/introductory-statistics and **Learning Statistics with R** https://learningstatisticswithr.com/ , **Python Edition of Learning Statistics with R** https://ethanweed.github.io/pythonbook/landingpage.html

Required Tools:

- R and RStudio, or Python and Spyder (or Jupyter Notebook or other IDEs of your choice)

Topics Covered

1. **Univariate and Bivariate Statistics:**
    a. Descriptive Statistics
    b. Sampling and Experiment Design
    c. Probability, Random Variables, Sampling Distributions of Sample Statistics
    d. Inferential Statistics

2. **Multivariate Data Analysis:**
    a. Interdependence: Principal Component Analysis (PCA), Multidimensional Scaling (MDS), Correspondence Analysis, Canonical Correlation Analysis (CCA), Latent Class Analysis (LCA), Clustering Analysis
    b. Dependence: Discriminant Analysis (LDA, QDA), Logit/Logistic Regression, Multiple Regression, Multivariate Analysis of Variance (MANOVA) and Covariance, Conjoint Analysis, Structural Equations Modeling (SEM)
3. **Statistical Machine Learning:**
    a. Classification: From Logistic Regression to Generative Models (LDA, QDA, Naive Bayes, K-Nearest Neighbors)
    b. Resampling Methods (Cross-Validation and Bootstrap)
    c. Tree-Based Methods
    d. Support Vector Machine and Neural Networks

Course Calendar

Week	January 2024								Week	February 2024								Week	March 2024								Week	April 2024							
	Su	Mo	Tu	We	Th	Fr	Sa			Su	Mo	Tu	We	Th	Fr	Sa			Su	Mo	Tu	We	Th	Fr	Sa			Su	Mo	Tu	We	Th	Fr	Sa	
		1	2	3	4	5	6		4					1	2	3	3	8						1	2	2	13		1	2	3	4	5	6	6
1	7	8	9	10	11	12	13		5	4	5	6	7	8	9	10	10	9	3	4	5	6	7	8	9		14	7	8	9	10	11	12	13	13
2	14	15	16	17	18	19	20	20	6	11	12	13	14	15	16	17	17	10	10	11	12	13	14	15	16	16	15	14	15	16	17	18	19	20	
3	21	22	23	24	25	26	27	27	7	18	19	20	21	22	23	24	24	11	17	18	19	20	21	22	23	23	16	21	22	23	24	25	26	27	
4	28	29	30	31					8	25	26	27	28	29				12	24	25	26	27	28	29	30	30		28	29	30					
																			31																

X	HW Assigned	X	HW Due Date	X	Project Presentation	X	Module Exam
X	Lecture Only	X	Weekly Quiz		Course Final Review	X	Final Exam

Course Activities

1. Homework Assignments

There will be **6 Homework Assignments**, assigned 1 week before the due date, focused on applying theory learned in the class to analyze a data set in R or Python. Assignment submissions should be in a single **PDF** file. The R or Python code used to generate your results should be appended to the end of your assignment. The *lowest score* will be dropped.

2. Quizzes

- There will be **3 Module Exams (60 minutes) and 9 Weekly Quizzes (15 minutes)** at the beginning of the class time, to assess students' understanding of concepts presented in the class. Students should ensure adequate preparation before starting the module exams and the weekly quizzes. Please note that it won't be possible to do well on the exams and quizzes without reviewing the course materials. The *lowest module exam score* may be replaced by the final exam score (if the final exam score is higher), and the *lowest weekly quiz score* will be dropped.
- There will be **In-Classroom Practices and Exit Tickets** which will be submitted at the end of the class time. The *lowest 3 scores* will be dropped.

3. Final Project

The project is open ended and the topics can be chosen by students. In this project, students will frame and solve problems using quantitative capabilities of Statistical Machine Learning with R or Python. Students will draft a formal proposal and submit for approval by the Teaching Team (5%), then carry out the project, write a project report, and prepare a 2-minute presentation to be presented in the classroom in the final week of the course (95%).

4. Final Examination

- The final exam will be comprehensive and will cover material from the entire course.
- The final exam will be closed notes and closed book, and in pen-n-paper format.

5. Reflective Journal

Keep a personal journal of critical reflections: To reflect on one's own individual journey throughout the learning process, to log important moments of growth and key learning during this process, to reflect on personal development or change in relation to learning, including lessons learned about self, the way of learning, and any accomplishments or challenges. A link to the live google doc of your reflective journal shall be included at the end of each Assignment submission.

Class Policies

- **Attendance & Absences** – This is an on-campus class. Class attendance is **mandatory** and will be recorded, and punctuality is expected. Statistics can be a challenging course at times and attending class is essential for your success in this course. If you have to miss a class, please email me as soon as possible.
- **Reading Assignments** are specified in the Course Schedule, to help check your understanding in advance, and help form quality questions to be asked and discussed during class meeting time.
- **Homework assignments** will be posted, on Assignment Day specified in the Course Schedule, on Canvas. If you need an extension due to illness, email me BEFORE the homework due date. The homework is meant for you to practice solving problems. **Do not search for homework solutions online**.
- **Late Policy**: The assignment due dates are created intentionally to help you manage time effectively, and for you to receive timely formative feedback to facilitate learning. It is expected that you are turning in your assignments by the due dates. Any late assignments are not guaranteed to receive timely feedback. **Assignments more than a week late without an official accommodation will result in a 0**.
- **Grading Policies:** While cooperative learning via group discussion is encouraged (and the final grades will not be curved, for the purpose of promoting peer learning), you should write your answers independently. Exam problems will often be similar in nature to assigned homework problems. Therefore you are personally responsible for knowing how to do each homework problem (even if you worked in a group on the homework). **So it is important that you understand how to solve the homework problems!**

- **Laptop Requirement:** Students should have a personal laptop. We will use laptops in the classroom to write Python or R programs. You will need a laptop in quizzes as well. Please have your laptop **FULLY CHARGED** before coming to the classroom every Saturday!

Grading Scale

The course grade will be based on
- Class participation (5%)
- 3 Module Exams (15%)
- 9 Weekly Quizzes (10%)
- 6 Homework Assignments (25%)
- Final project (20%)
- Final exam (25%)

Final grades will be assigned according to the following ranges:

A 93.50-100%	B+ 86.50-89.49%	C+ 76.50-79.49%	D+ 66.50-69.49%	F <60%
A– 89.50-93.49%	B 83.50-86.49%	C 73.50-76.49%	D 63.50-66.49%	
	B– 79.50-83.49%	C– 69.50-73.49%	D– 60.00-63.49%	

Statement of Support

Take care of yourself. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding substance abuse, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress.

All of us benefit from support during times of struggle. There are many helpful resources available on campus and an important part of the college experience is learning how to ask for help. Asking for support sooner rather than later is almost always helpful.

If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support.

Tips for Success:

1. **Three Simple Rules for Success** (that can benefit anyone who wants to be better in life):
    a. **Know the text**: Complete the reading assignments before class meeting time
    b. **Have a head full of ideas**: Bring questions to the classroom & willing to participate
    c. **Show up on time**: Coming in a few minutes early liberates you, allowing you time to get comfortable and composed before you need to be at your very best
2. **Learning** statistics **by doing** statistics:
    a. **Conceptual understanding** over memorizing
    b. **Experimenting** over being perfect
    c. **Process** over product
3. **Learning statistics is like learning a new language**: **Practice makes perfect!**
4. **Time** commitment and management (**at least 10.5 hours per week outside of class**) and **practice regularly** (**at least 15 minutes per day** will make a big difference within the short period of a semester).
5. ***"The secret of getting ahead is getting started." – Mark Twain***

End-of-Course Evaluation Surveys

Your feedback regarding your educational experience in this class is particularly important to the College of Professional Studies. Your comments will make a difference in the future planning and presentation of our curriculum.

At the end of this course, please take the time to complete the evaluation survey at **https://neu.evaluationkit.com**. Your survey responses are **completely anonymous and confidential**. For courses 6 weeks in length or shorter, surveys will be open one week prior to the end of the courses; for courses greater than 6 weeks in length, surveys will be open for two weeks. An email will be sent to your Husky Mail account notifying you when surveys are available.

Academic Integrity

A commitment to the principles of academic integrity is essential to the mission of Northeastern University. The promotion of independent and original scholarship ensures that students derive the most from their educational experience and their pursuit of knowledge. Academic dishonesty violates the most fundamental values of an intellectual community and undermines the achievements of the entire University.

As members of the academic community, students must become familiar with their rights and responsibilities. In each course, they are responsible for knowing the requirements and restrictions regarding research and writing, examinations of whatever kind, collaborative work, the use of study aids, the appropriateness of assistance, and other issues. Students are responsible for learning the conventions of documentation and acknowledgment of sources in their fields. Northeastern University expects students to complete all examinations, tests, papers, creative projects, and assignments of any kind according to the highest ethical standards, as set forth either explicitly or implicitly in this Code or by the direction of instructors.

Go to http://www.northeastern.edu/osccr/academic-integrity-policy/ to access the full academic integrity policy.

University Health and Counseling Services

As a student enrolled in this course, you are fully responsible for assignments, work, and course materials as outlined in this syllabus and in the classroom. Over the course of the semester if you experience any health issues, please contact UHCS.

For more information, visit https://www.northeastern.edu/uhcs.

Student Accommodations

Northeastern University and the Disability Resource Center (DRC) are committed to providing disability services that enable students who qualify under Section 504 of the Rehabilitation Act and the Americans with Disabilities Act Amendments Act (ADAAA) to participate fully in the activities of the university. To receive accommodations through the DRC, students must provide appropriate documentation that demonstrates a current substantially limiting disability.

For more information, visit https://drc.sites.northeastern.edu.

Graduate Student Resources

Free one-on-one communications support for courses and research on campus for graduate students:

The College of Engineering CommLab
offers in person and on-line workshops and peer-to-peer coaching for graduate-level writing and communication tasks associated with research and course work. View the website to schedule an appointment. Join the CommLab Team to view upcoming or past workshop materials.

Global Student Success and the International Tutoring Center
are dedicated to providing international and non-native English-speaking students, scholars, faculty, and staff with comprehensive English language and academic support. Sessions are offered in a one-on-one format in Boston or online. Group conversation tutorials are also available.
https://international.northeastern.edu/gss/tutoring/

The Northeastern Writing Center
is open to students, staff, faculty, and alumni of Northeastern and exists to help writers at any level, and from any academic discipline, in their written communication.
https://cssh.northeastern.edu/writingcenter/

The Northeastern University Library
has the expertise and tools to support you throughout the lifecycle of your research. The Northeastern University Library is at the hub of campus intellectual life. Resources include over 900,000 print volumes, 206,500 e-books, and 70,225 electronic journals. No matter where you are in the research process, or where you're located in the world, we're ready to help you and your team:

- Find resources to inform and enhance your projects
- Manage the details of your research and publishing processes
- Create graphics to clearly communicate your findings
- Fulfill funder requirements for data and research outputs
- Elevate the visibility and impact of your work

https://library.northeastern.edu/

24/7 Canvas Technical Help

For immediate technical support for Canvas, call 617-373-4357 or email help@northeastern.edu
Canvas Faculty Resources: https://canvas.northeastern.edu/faculty-resources/
Canvas Student Resources: https://canvas.northeastern.edu/student-resources/
For assistance with my Northeastern email, and basic technical support:
Visit ITS at https://its.northeastern.edu
Email: help@northeastern.edu
ITS Customer Service Desk: 617-373-4357

Diversity and Inclusion

Northeastern University is committed to equal opportunity, affirmative action, diversity, and social justice while building a climate of inclusion on and beyond campus. In the classroom, members of the University community work to cultivate an inclusive environment that denounces discrimination through innovation, collaboration, and an awareness of global perspectives on social justice.

Please visit http://www.northeastern.edu/oidi/ for complete information on Diversity and Inclusion

Title IX

Title IX of the Education Amendments of 1972 protects individuals from sex or gender-based discrimination, including discrimination based on gender-identity, in educational programs and activities that receive federal financial assistance.

Northeastern's Title IX Policy prohibits Prohibited Offenses, which are defined as sexual harassment, sexual assault, relationship or domestic violence, and stalking. The Title IX Policy applies to the entire community, including male, female, transgender students, faculty, and staff.

In case of an emergency, please call 911.

*Please visit **https://www.northeastern.edu/ouec** for a complete list of reporting options and resources both on- and off-campus.*

Syllabus Statement

This syllabus is not a contract. The instructor reserves the right to alter course requirements and/or assignments based on new materials, class discussions, or other legitimate pedagogical objectives.