

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 2025

Credit Hour: 4

CRN: 40534

Course Format: On-ground (Traditional)

Location: TBD

Lecture Hours: Tuesday 18:10 - 21:40 ET

Student Hours (in person): Saturday 17:40-18:10 & 21:40-22:10 ET | By Appointment

TA Student Hours (via Zoom): TBD

Instructor Information

Full Name: Hong Pan, Ph.D., Professor Pan (He/His/Him)

Email Address: hong.pan@northeastern.edu

TA: TBD

Email Address: TBD

Instructor Biography

Hong came from a family of educators and pursued his academic journey by studying Biomedical Engineering at Shanghai Jiao Tong University before earning a PhD in Electrical and Computer Engineering from Purdue University in the U.S.

With over 25 years of research career as a faculty member at Cornell University Medical College and then at Harvard Medical School, Hong has been at the forefront of data science efforts. He is a subject matter expert on over 20 federal and institutional projects, and his expertise in AI/ML algorithms and advanced statistics has been instrumental in developing statistical, data-driven diagnostic tools for guiding the treatment of brain disorders. His professional standing in the field is solidified by four patents, a successful startup, and prestigious awards, including the Mass General Brigham Excellence in Innovation Award and Brigham and Women's Hospital's Pillar Award in Research & Innovation.

In 2023, Hong began a new chapter by transitioning to a full-time data science and statistics teaching role at Simmons University. There, he can impart his knowledge and expertise to the next generation of professionals in the field.

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 to 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. It lays the foundation of how learning models are derived from complex data pipelines, both algorithmically and practically. The course content will equip you with the skills and knowledge necessary to tackle real-world data science challenges. 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. 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 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, demonstrating the chapter's concepts in either R or Python. It is available online 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>, and **An Introduction to Multivariate Analysis with R** by Brian Everitt & Tolsten Holthorn (2011) <https://drive.google.com/file/d/1Q9cKsm0SVI84WoxjzUoP2mE3A9YzYYVV/view?usp=sharing>
- **Univariate and Bivariate Statistics:**
 - (a) **Introductory Statistics at OpenStack:** <https://openstax.org/details/books/introductory-statistics> Referred to as **IS**
 - (b) **Statistics and Probability at Khan Academy:** <https://www.khanacademy.org/math/statistics-probability> Referred to as **SP**
 - (c) **Learning Statistics with R** <https://learningstatisticswithr.com/>, **Python Edition of Learning Statistics with R** <https://ethanweed.github.io/pythonbook/landingpage.html> Referred to as **LSR**

Required Tools:

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

Course Schedule

	January							W04	February							W08	March							W13	April						
	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
W01	5	6	7	8	9	10	11	W05	2	3	4	5	6	7	8	W09	2	3	4	5	6	7	8	W14	6	7	8	9	10	11	12
W02	12	13	14	15	16	17	18	W06	9	10	11	12	13	14	15	W10	9	10	11	12	13	14	15	W15	13	14	15	16	17	18	19
W03	19	20	21	22	23	24	25	W07	16	17	18	19	20	21	22	W11	16	17	18	19	20	21	22	W16	20	21	22	23	24	25	26
W04	26	27	28	29	30	31		W08	23	24	25	26	27	28		W12	23	24	25	26	27	28	29		27	28	29	30			
																W13	30	31													
	Class Begins			Homework Due Date					Spring Break					Midterm Exam			Final Project Presentation				Final Exam										

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 Activities

1. Homework Assignments

There will be **12 Homework Assignments**, assigned one 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 2 scores will be dropped.**

2. Quizzes

There will be **3 Midterm Exams (60 minutes)** and **9 Weekly Quizzes (15 minutes)** at the beginning of class time to assess students' understanding of concepts presented in the class. Students should ensure

adequate preparation before starting the Midterm 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 midterm 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 as well.

3. Attendance and Participation

There will be **in-classroom practices and exit tickets**, which will be submitted at the end of class. **The lowest 2 scores will be dropped.**

4. Final Project

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

5. Final Examination

- The final exam will be comprehensive and cover the entire course's material.
- The final exam will be closed notes and closed books in pen-n-paper format.

6. Reflective Journal

Keep a personal journal of critical reflections: To reflect on one's journey throughout the learning process, to log essential moments of growth and key learning during this process, to reflect on personal development or change about 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 beginning of each Assignment submission.**

Class Policies

- **Attendance Policy: Students registered in MGEN courses (INFO, CSYE, and DAMG) are allowed a maximum of 2 absences per course, with 3 or more absences resulting in an automatic 'F' for that course.** Students are expected to inform their instructors of any absences in advance of the class; if a student is sick long-term or experiences a medical issue that prevents class attendance, it is strongly encouraged that they speak with their Academic Advisor (coe-mgen-gradadvising@northeastern.edu) to learn more about the Medical Leave of Absence. Should a student anticipate being unable to attend 3 or more classes, they should discuss their situation with their Academic Advisor to explore other types of leave in accordance with the University's academic and global entry expectations. International students should review the Office of Global Services webpage to understand their visa compliance requirements.
- **Reading Assignments** are specified in the Course Schedule to help you check your understanding and form quality questions for discussion during class meetings.

- **Homework assignments** will be posted on Canvas on Assignment Day, as specified in the Course Schedule. 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 to receive timely formative feedback to facilitate learning. You are expected to turn in your assignments by the due dates. Students must communicate with the faculty before the deadline if they anticipate work will be submitted late. Work submitted late without prior communication with faculty will not be graded. **Assignments over a day 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 to promote peer learning), you should write your answers independently. Exam problems will often be similar to assigned homework problems. Therefore, you are personally responsible for knowing how to solve each homework problem (even if you worked in a group on the homework). **So you must 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, and you will also need a laptop for quizzes. Please have your computer **FULLY CHARGED** before coming to the classroom every Saturday!

Grading Scale

The course grade will be based on

- Class participation (10%)
- 3 Midterm Exams (10%)
- 9 Weekly Quizzes (15%)
- 12 Homework Assignments (25%)
- Final project (20%)
- Final exam (20%)

Final grades will be assigned according to the following ranges:

A 95-100%	B+ 87-89.9%	C+ 77-79.9%	F \leq 69.9%
A- 90-94.9%	B 84-86.9%	C 74-76.9%	
	B- 80-83.9%	C- 70-73.9%	

Our Classroom Community

At Boston University, faculty and students work together to build a respectful, inclusive learning environment. We aim to create and maintain a positive and supportive classroom atmosphere where all members' diversity, backgrounds, and perspectives are valued and respected. The following guidelines will help us work toward this goal and clarify expectations for engagement in and with each other in this course.

1. **Cooperative Learning:** While cooperative learning via group discussion is encouraged (and the final grades will not be curved to promote peer learning), you should write your answers

independently. Exam problems will often be similar to assigned homework problems. Therefore, you are personally responsible for knowing how to solve each homework problem (even if you worked in a group on the homework). **So you must understand how to solve the homework problems!**

2. **During Class:** Cell phones may not be used during class. Laptop computers must be put away during class time, except for class activity time. Tablets (e.g., iPads) may be used only for note-taking if they are flat on the desk like a traditional notebook. Students may not use tablets to look at web pages, play games, etc. **Pencil-and-paper note-taking is encouraged, and the Cornell Note-Taking Method is recommended.**
3. **Communication:** The best way to contact me is through email. Please give me 48 hours to respond. After that time, please follow up if you have not heard from me in case your email was lost in the shuffle. **Emailing your professor or teaching assistants should be treated as professional communication.** Emails should have an appropriate greeting and ending; students should refrain from using any kind of “shortcuts,” abbreviations, acronyms, slang, etc., in the email text. Emails not meeting these standards may not be answered. Email questions must be sent a reasonable amount of time before a due date.

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 semester).**
5. **FACE: Focus - Attitude - Creativity - Effort:** All of these work together: When you improve one, the others will improve as well.
6. **A Growth Mindset:** Keep an open mind and be willing to try new things. We are all on this journey together!
7. ***“The secret of getting ahead is getting started.” – Mark Twain.***

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. Many helpful resources are available on campus, and **an essential part of the academic 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.

End-of-Course Evaluation Surveys

Your feedback regarding your educational experience in this class is particularly important to the College of Engineering. 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 Northeastern University 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.

MGEN Student Feedback

Students who would like to provide the MGEN unit with anonymous feedback on this particular course, Teaching Assistants, Instructional Assistants, professors, or to provide general feedback regarding their program, may do so using this survey: https://neu.co1.qualtrics.com/jfe/form/SV_cTIAbH7ZRaaWOKi

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>.

Library Services

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.

For more information and for education specific resources, visit <https://library.northeastern.edu>
Network Campus Library Services: [Northeastern University Library Global Campus Portals](#)

24/7 Canvas Technical Help

For immediate technical support for Canvas, call 617-373-4357 or email help@northeastern.edu

Canvas Student Resources: <https://canvas.northeastern.edu/student-resources/>

For assistance with my Northeastern e-mail, 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.