Northeastern
College of Engineering

INFO6105

DATA SCIENCE ENGINEERING METHODS AND TOOLS

Course Information

Course Title: INFO 6105 Data Science Engineering Methods and Tools
Course Number: INFO6105
Term and Year: Fall 2025
Credit Hours: 4
CRN: 17642
Course Format: On-ground (Traditional)
Location: Ryder Hall 155
Lecture Hours: Saturday 13:00 - 16:30 ET
Student Hours (in person): Tuesday 12:30-13:00 & 16:30-17:00 ET | By Appointment
TA Student Hours: 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 Pan, a dedicated professional in data science and statistics, 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, Hong first joined Cornell University Medical College as a faculty member, where he conducted and oversaw 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.

For over 25 years, Hong has been a leader in data science efforts, serving as the subject matter expert on over 20 federal and institutional projects. His influence and impact in the field, particularly his expertise in AI/ML algorithms and advanced statistics, have been instrumental in developing statistical and data-driven diagnostic tools to guide the treatment of brain disorders. He has created best practice approaches for optimized data acquisition, data science solutions for biomarker discovery, and automated analytics and informatics pipelines based on functional neuroimaging methodology. His work has resulted in four granted patents, a successful spin-out startup, and earned him the Mass General Brigham Excellence in Innovation Award twice, as well as Brigham and Women's Hospital's Pillar Award in Research & Innovation. He has also published over 60 journal articles, solidifying his professional standing in the field.

In 2023, Hong joined Simmons University, a women-centered liberal arts college, as a faculty member, focusing on full-time teaching in data science and statistics.

Northeastern
College of Engineering

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.

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, and recommender systems) and supervised learning (parametric and non-parametric algorithms, support vector machines, kernels, neural networks, and 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.

Course Schedule

	September								October								November								December						
	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		1	2	3	4	5	6	W05				1	2	3	4	W09							1	W14		1	2	3	4	5	6
W02	7	8	9	10	11	12	13	W06	5	6	7	8	9	10	11	W10	2	3	4	5	6	7	8	W15	7	8	9	10	11	12	13
W03	14	15	16	17	18	19	20	W07	12	13	14	15	16	17	18	W11	9	10	11	12	13	14	15	W16	14	15	16	17	18	19	20
W04	21	22	23	24	25	26	27	W08	19	20	21	22	23	24	25	W12	16	17	18	19	20	21	22		21	22	23	24	25	26	27
W05	28	29	30					W09	26	27	28	29	30	31		W13	23	24	25	26	27	28	29		28	29	30	31			
																W14	30														

Class Meeting Time | Midterm Exam | Fall Break | 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, CLT
    d. Inferential Statistics
2. Multivariate Data Analysis:
    a. Interdependence: Principal Component Analysis (PCA), Clustering Analysis (k-Means)
    b. Dependence: Discriminant Analysis (LDA, QDA), Logistic Regression, Multiple Regression
3. Statistical Machine Learning:
    a. Conditional Independence and Classification (Naive Bayes)
    b. Resampling Methods (Cross-Validation and Bootstrap)
    c. Tree-Based Methods, Ensembles of Models & Evaluating Performance
    d. Support Vector Machine and Neural Networks

Northeastern
College of Engineering

Required Tools and Course Textbooks

Textbooks:

- **Univariate and Bivariate Statistics:**
  - (a) **Introductory Statistics 2e 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 K**
  - (c) **Learning Statistics with R** https://learningstatisticswithr.com/ **Referred to as LSR**
- **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 **Referred to as MDAR**
- **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, and contains a lab at the end of each chapter, demonstrating the chapter's concepts in R. It is available online at https://statlearning.com. You can also purchase a hard copy from Springer. **Referred to as ISLR**

Required Tools:

- R and RStudio (or Jupyter Notebook)

Additional Materials:

- Peter Dalgaard. **Introductory Statistics with R (2e).** Springer Verlag, 2008. **[Refer to as ISwR]**
- Roger Peng. **R Programming for Data Science.** LeanPub, 2014-2022. **[Refer to as RPDS]**
- Utilizing the interactive learning tool **swirl** at https://swirlstats.com/ and progressing with at least one submodule (~15 minutes) daily is recommended.
- Roger D. Peng and Elizabeth Matsui. **The Art of Data Science: A Guide for Anyone Who Works with Data.** LeanPub, 2015-2018.
- Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund. **R for Data Science (2e)**. https://r4ds.hadley.nz/
- Hadley Wickham, Danielle Navarro, and Thomas Lin Pedersen. **ggplot2: Elegant Graphics for Data Analysis (3e)**. https://ggplot2-book.org/
- Max Kuhn and Julia Silge. **Tidy Modeling with R**. https://www.tmwr.org
- Benjamin S. Baumer, Daniel T. Kaplan, and Nicholas J. Horton. **Modern Data Science with R (3e)**. https://mdsr-book.github.io/mdsr3e/
- Rafael A. Irizarry. **Introduction to Data Science I: Data Wrangling and Visualization with R.** https://rafalab.dfci.harvard.edu/dsbook-part-1/
- Rafael A. Irizarry. **Introduction to Data Science II: Statistics and Prediction Algorithms Through Case Studies.** https://rafalab.dfci.harvard.edu/dsbook-part-2/
- James Long and Peter Teetor. (2019). **R cookbook (2e)**. https://rc2e.com
- Winston Chang. **R graphics cookbook (2e)**. https://r-graphics.org
- Garrett Grolemund. **Hands-On Programming with R**. https://rstudio-education.github.io/hopr/
- Yihui Xie, Christophe Dervieux, Emily Riederer. **R Markdown Cookbook**. https://bookdown.org/yihui/rmarkdown-cookbook/

Northeastern
College of Engineering

Course Activities

1. Homework Assignments

There will be **10 Homework Assignments**, assigned one week before the due date, focused on applying theory learned in the class to analyze a data set in R. Assignment submissions should be in a single **PDF** file. The R code used to generate your results should be appended to the end of your assignment.

2. Quizzes

There will be **two midterm exams (60 minutes) and nine weekly quizzes (15 minutes) at the beginning of class time to assess students' understanding of the concepts presented in** 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).

3. Attendance and Participation

There will be **in-class practices and exit tickets,** which will be submitted at the end of class.

4. Final Project

The project is open-ended, allowing students to choose their topics. In this project, students will frame and solve problems using the quantitative capabilities of Statistical Machine Learning with R. 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 3-minute presentation in the classroom in the final week of the course (95%).

5. Final Examination

- The final exam will be comprehensive, covering the entire course material.
- The final exam will be closed-notes and closed-books in pen-and-paper format.
- All the exams and quizzes allow a 2-sided Letter-sized cheat sheet you prepared for yourself.

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 experiences during this process, to reflect on personal development or changes in learning, including lessons learned about oneself, the learning process, 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 two absences per course; three or more absences result 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 for an extended period or experiences a medical issue that prevents them from attending class, it is strongly recommended that they speak with their Academic Advisor (coe-mgen-gradadvising@northeastern.edu) to learn more about the Medical Leave of Absence policy. If a student anticipates 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 requirements. 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 and graded for completion. The 4-bin grading scale (EMRN: Excellent/Exemplary, Meets Expectations, Revision Needed, Not Assessable) will provide feedback for a chance of revision within 2 weeks. **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 submit your assignments by the due dates and are provided with an additional week for revision if needed. The answer keys will be released 1 week after the due date, after which no work will be accepted. You are allowed **3 Late Work Tokens** for the course by **completing this Google Form**.
- **Grading Policies: All assignments will be graded on the EMRN 4-bin grading scale to provide timely feedback for revisions. A maximum of <u>one</u> revision and <u>one extra</u> week are allowed per assignment.**
- **Laptop Requirement:** Students are required to have a personal laptop. We will use laptops in the classroom to write R programs. Please ensure your computer is fully charged before coming to school every week.

Grading Scale

A system of multiple grading schemes is adopted to promote equitable learning and assessment experiences, providing space for growth and development. You will be provided with a Google Sheet "WHAT IF Grade Calculator" to track your progress.

- **Student Work:** [35~70%]
  - **Class Activities & Participation**
  - **10 Homework Assignments**
  - **Final Project and Presentation**
- **Assessments**: [30~65%]
  - **9 Weekly Quizzes**
  - **2 Midterm Exams**
  - **Final Exam: Cumulative**

Extra Credit: No individual extra credit will be given for this class. The instructor will determine whether extra credit is needed by examining the distribution of grades for the whole class.

Final grades will be assigned according to the following ranges:

A 95-100%	B+ 87-89.9%	C+ 77-79.9%	F ≤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 Northeastern University, faculty and students collaborate to create a respectful and inclusive learning environment. We aim to create and maintain a positive and supportive classroom atmosphere where the diversity, backgrounds, and perspectives of all members are valued and respected. The following guidelines will help us work toward this goal and clarify expectations for engagement with one another 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. Laptop computers must be put away during class time, except during designated class activity periods. Tablets (e.g., iPads) may be used for note-taking only if they are flat on the desk, similar to 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 haven't 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," such as abbreviations, acronyms, or slang, in the email text. Emails that do not meet these standards may not receive a response. Email questions must be sent a reasonable amount of time before the 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 the 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 by doing**:
    a. **Conceptual understanding** over memorizing
    b. **Experimenting** over being perfect
    c. **Process** over product
3. **Learning data science 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). The easiest way to manage your weekly calendar is to set a fixed starting time every day for this course (say, 2 pm) with varied lengths (say, from a minimum of 15 minutes to a maximum of 2 hours), resulting in at least 10.5 hours.
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 manage stress effectively.

We all 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 **entirely anonymous and confidential**. For courses 6 weeks in length or shorter, surveys will be open one week before the end of the course; for courses longer than 6 weeks, 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, students are responsible for knowing the requirements and restrictions regarding research and writing, examinations of any kind, collaborative work, the use of study aids, the appropriateness of assistance, and other relevant issues. Students are responsible for learning the conventions of documentation and acknowledging sources in their respective 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_cTIAbH7ZRaaw0Ki

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 the classroom. Over 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 university's activities. 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 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, and 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 modify course requirements and/or assignments in response to new materials, class discussions, or other legitimate pedagogical objectives.