CHEM452/CHBE413
Data Science for Chemistry and Engineering
Fall 2025

Prerequisites: MATH 225, MATH 227, MATH257, or MATH 415

Restrictions: Junior or Senior standing for undergraduates. Knowledge of essential programming constructs (e.g. functions, loops, conditional statements) in the context of a programming language (e.g. C/C++, Fortran, Java) is required. Basic proficiency with the Python programming language is strongly recommended.

Instructors:

- Prof. Nick Jackson (NL 350E, jacksonn@illinois.edu)

Teaching Assistant: Mr. Jingdan Chen (jingdan2@illinois.edu)

Required Textbooks:

1. *An Introduction to Statistical Learning with Python* by James, Witten, Hastie, Tibshirani. ISBN 9781071614174. 1st or 2nd Edition (**freely available** https://hastie.su.domains/ISLP/ISLP_website.pdf.view-in-google.html). – We will call this book **ISL**.
2. *Understanding Deep Learning* by Prince (**freely available** https://udlbook.github.io/udlbook/).

Additional Resources:

- *Deep Learning with Python* 2nd Edition by Chollet
- *Dive into Deep Learning,* Zhang, Lipton, Li, Smola ***(freely available** https://d2l.ai/)*
- *Python Programming and Numerical Methods: A Guide for Engineers and Scientists* Kong, Siauw, Bayen (**freely available book** https://pythonnumericalmethods.berkeley.edu/notebooks/Index.html) We will call this book **PPNM**
- https://scikit-learn.org/stable/user_guide.html - This is the online user-guide for the main machine learning library we will be using known as Scikit-learn. It is an outstanding resource.
- https://machinelearningmastery.com/ - This is one of the best blogs out there for beginners and I strongly suggest consulting it. Everything is done in Python and Keras, like this course.
- *Deep Learning for Molecules & Materials* by White. https://dmol.pub/
- *The Elements of Statistical Learning* by Hastie, Tibshirani and Friedman. ISBN 9780387848570 2nd Edition. We will call this book "ESL".
- Keras code examples for deep learning: https://keras.io/examples/
- DeepChem tutorials: https://deepchem.io/tutorials/the-basic-tools-of-the-deep-life-sciences/

Website: http://canvas.illinois.edu. The syllabus, lecture notes, coding examples, and homework assignments will be posted on the class website.

Class Schedule: Tuesday & Thursday, 3:00 – 4:20 PM, 163 Noyes Laboratory
TA Office Hours: Tuesday 2-3 pm (3rd Floor Noyes Theory Center Common Area)
Prof. Jackson Office Hours: Wednesdays 1-2 pm, Noyes Lab 350E.

Homework: There will be 10 homework assignments throughout the semester that will be completed as executable Jupyter notebooks. Homework is to be submitted on Canvas by 5 pm on the due date.

There will be two 48-hour take-home exams listed in the schedule. These exams will take a similar form to the programming homework assignments done weekly. Make-ups will be scheduled to accommodate students who miss an exam for valid reasons. Please include appropriate evidence to support your request for a make-up exam.

Grading:

Homework – 40% **Exam 1 – 15%** **Exam 2 – 15%** **Final Project – 30% (20% Written Report and GitHub Code, 10% Presentation)**	**Graded on a curve*, with median grade in the B/B+ range. The curve can only help you, i.e. 90% is at least an A-, 80% is at least a B-, 70% is at least a C-, and 60% is at least a D-.**

Course Description:

This course serves as an introduction to data science and machine learning for advanced undergraduate students. Essential to the course is a practical programming component in which students will practice programming the machine learning methods introduced in lecture.

At the conclusion of this course, you will be able to:

1. Process and featurize datasets relevant to chemistry and chemical engineering.
2. Perform supervised and unsupervised machine learning using Python.
3. Quantify and assess the performance of machine learning models.
4. Critique machine learning models in the context of the bias-variance tradeoff.
5. Understand and apply common neural network architectures using Keras.
6. Learn how to apply machine learning in experimental optimization campaigns.
7. Abide by best practices in the machine learning community.
8. Understand how data science and machine learning is used in the chemical industry.
9. Function effectively on a team, provide leadership. Set goals, tasks, and objectives.

***Programming is an integral part of this course*.** We will use Python due to its straightforward interfacing with many powerful machine learning libraries (e.g. scikit-learn, Keras). To remove barriers to programming, we will be using Jupyter notebooks in Google Colab to write and run our machine learning algorithms. Google Colab can be run entirely through your browser and only requires a valid gmail account to use. (At least) the following Python libraries will be utilized in this course: numpy, scipy, scikit-learn, keras, tensorflow, rdkit, seaborn, matplotlib. Due to variable backgrounds in programming, supplementary programming reviews in Python will also be uploaded to the course website. While basic Python programming will be presented in the first two weeks of class, all homework will involve writing Python code in Jupyter notebooks, and **thus it is your responsibility to be sufficiently prepared.**

The first half of the course will cover introductions to Python programming via Google Colab and Jupyter notebooks in conjunction with a brief review of the necessary mathematics for machine learning. It will then be followed by fundamental techniques in machine learning: linear and nonlinear regression, classification, model assessment, regularization, unsupervised learning, Bayesian optimization, active learning, and chemical featurization. The second half of the course will introduce neural network architectures (feed-forward, convolutional, recurrent) and their many variations and applications in a chemical context. All datasets utilized in this class constitute **real** chemical or engineering datasets for problems that are being actively studied.

There is a critical computational laboratory component to the course. To solidify the concepts introduced in lecture, there will be a computational laboratory for the course. In this session, the professor and TA will lead a hands-on laboratory utilizing Google Colab and Jupyter notebooks to implement the machine learning techniques (using Python) discussed in class on real chemical datasets. The laboratory will cover practical programming aspects particularly related to efficiency and code debugging, which is essential for becoming proficient in the techniques. Later in the course, this computational lab will also serve as a point of collaboration for students when working on the final group projects.

The course will feature guest lectures from chemical data scientists in industry and academia. We tentatively anticipate guest lectures from researchers at Dow. These lectures will give you the opportunity to understand how the models you learn about in class are applied in practical settings, as well as the potential career opportunities that the skills in this course might (eventually!) lead to.

The course will culminate in a final project. 30% of the course grade will be derived from a team-based project on a cutting-edge machine learning topic selected by the students. This will result in a written project report and slide presentation before students and professors at UIUC. This will provide students in the course (i) the opportunity to study machine learning techniques of high interest that were not covered in the lecture schedule, (ii) practice working on data science projects in a team-based setting, and (iii) practice explaining and presenting data science concepts to non-experts.

I. Tentative Lecture Schedule

Week	Date		Outline	Reading/Notebooks	Comments
1	Aug	26	Course Overview and Software Ecosystem	• ISL Ch. 1 • Prince Ch. 1 • **CDS25_IntroPython.ipynb**	
	Aug	28	Crash Course in Python	• ISL Ch. 1 • PPNM Ch. 1-5, 11, 12 (Python Practice) • **CDS25_IntroPython.ipynb** • **CDS25_NumpyScipyMatrices.ipynb**	HW1 Assigned
2	Sep	2	Using Large Language Models for Science	• **CDS25_IntroPython.ipynb** • **CDS25_NumpyScipyMatrices.ipynb**	
	Sep	4	Exploratory Data Analysis **(Zoom lecture)**	• **CDS25_Pandas_EDA.ipynb**	HW 1 Due HW2 Assigned
3	Sep	9	Machine Learning Method Overview and Linear Regression	• ISL Ch. 3.1-3.3, 7.1 • Prince Ch. 2 • **CDS25_LinearRegression_GradientDescent.ipynb**	
	Sep	11	Regularization, Overfitting, and Partial Least Squares	• ISL Ch. 6.1-6.4 • **CDS25_LassoRidgePLS.ipynb**	HW2 Due HW3 Assigned
4	Sep	16	Model Assessment: Cross-Validation	• ISL Ch. 2.2, 5.1-5.3 • Prince Ch. 5 • **CDS25_ModelAssessment.ipynb**	**Projects Introduced**

	Sep	18	Molecular Featurization **(Zoom Lecture)**	• **CDS25_MolecularFeaturization.ipynb**	HW3 Due HW4 Assigned
5	Sep	23	Classification: Logistic Regression	• ISL Ch. 4.1-4.3 • **CDS25_LogisticRegression.ipynb**	
	Sep	25	Support Vector Machines	• ISL Ch. 9.1-9.5 • **CDS25_SVM.ipynb**	HW4 Due HW5 Assigned
6	Sep	30	Decision Trees and Random Forests	• ISL Ch. 8.1 • **CDS25_DecisionTrees.ipynb**	
	Oct	2	Kernel Ridge Regression and Gaussian Process Regression	• https://dmol.pub/ml/kernel.htm • **CDS25_GPRfromScratch.ipynb**	**Project Team Selection due** HW5 Due HW6 Assigned
7	Oct	7	Data Efficient Discovery with Active Learning	• **CDS25_ActiveLearning_1D.ipynb**	
	Oct	9	Data Efficient Discovery with Bayesian Optimization	• **CDS25_BayesianOptimization_1D.ipynb**	HW6 Due
8	Oct	14	**Exam 1** Unsupervised Learning: Clustering **(Zoom Lecture)**	• ISL Ch. 12.1,12.4 • Prince Ch. 14 • **CDS25_KMeans.ipynb**	**Exam 1 available online**
	Oct	16	Unsupervised Learning: Dimensionality Reduction	• ISL Ch. 12.3 • **CDS25_PCA_tSNE.ipynb**	**Exam 1 due** HW7 Assigned
9	Oct	21	Explainable AI	• https://dmol.pub/dl/xai.html • **CDS25_SHAP_ReactivityRatios.ipynb**	
	Oct	23	**Data Science in Industry Lecture:** Alix Schmidt, Dow	Senior data scientist in Dow Chemical's R&D Information Research Team	**Project Topic Selection due** HW7 Due HW8 Assigned
10	Oct	28	Introduction to Deep Learning	• ISL Ch. 10	
	Oct	30	Neural Networks I	• ISL Ch. 10 • Prince Ch. 3, 4, 5 • **CDS25_NeuralNetworkWithKeras_DeepDive.ipynb**	HW8 Due HW9 Assigned

11	Nov	4	Neural Networks II	• ISL Ch. 10 • Prince Ch. 6, 7 • **CDS25_Keras_NNRegression.ipynb**	
	Nov	6	Neural Networks III	• ISL Ch. 10 • Prince Ch. 8, 9 • **CDS25_Keras_NNClassification.ipynb**	HW9 Due HW10 Assigned
12	Nov	11	Convolutional Neural Networks	• ISL Ch. 10 • Prince Ch. 10 • **CDS25_Keras_ConvNet_Example.ipynb**	
	Nov	13	Graph Neural Networks	• Prince Ch. 13 • **CDS25_DeepChem_GraphConv.ipynb** • https://distill.pub/2021/gnn-intro/ • https://distill.pub/2021/understanding-gnns/	HW10 Due
13	Nov	18	**Exam 2** Neural Networks for Time Series and Language	• ISL Ch. 11 • **CDS25_RNN_weatherforecasting.ipynb**	**Exam 2 available online**
	Nov	20	Attention and Transformers	• Prince Ch. 12	**Exam 2 due**
	Nov	25	**Fall Break**		
	Nov	27	**Fall Break**		
14	Dec	2	Deep Learning and Ethics **(Zoom Lecture)**	• Prince Ch. 20	
	Dec	4	**Final Presentations of Projects I**		**Project Reports Due**
15	Dec	9	**Final Presentations of Projects II**		

II. Homework Policies:

The following homework guidelines will assist the graders, and will also help you to receive appropriate credit for your effort:

1. **Your homework and code must be legible and turned in as both a .pdf and an executable Jupyter notebook file**. Your codes must be properly commented. https://stackoverflow.blog/2021/12/23/best-practices-for-writing-code-comments/. Submit your homework online on the course website. Appropriate folders will be created every week for homework submission.

2. Please provide explanations that will help the grader follow your solution. For example, mathematical solutions and algorithms used for solving a problem should always be accompanied by explanations. Any assumptions, simplifications, and use of physical or mathematical reasoning or intuition should be clearly stated and justified.

3. **Policy on Homework Collaboration:** You are encouraged to discuss the underlying concepts and approaches with your classmates. However, such discussions should stop short of jointly prepared solutions. That is, the work you turn in should reflect your own efforts. Strict penalties are associated with any forms of cheating. Be aware of the University Student Code found at: https://studentcode.illinois.edu/ respectively.

4. **Policy on late submissions:** Please try to submit the homework on time. If you need an extension of the homework submission deadline, please contact the instructor or the TA. For late submission, 50% of the points will be deducted. No points will be awarded for any homework submissions after the solution is posted online.

III. Office Hours

1. Professor and TA office hours are listed at the beginning of the syllabus. If there are any major conflicts with these days/times, please let us know.

2. Please feel free to ask any questions during office hours. There is never a bad question, and we are here to help you learn the material.

IV. Non-Office Hours

Please try to come to office hours if you have questions. It is easier for us, and you will benefit from interacting with other students who come. However, we do recognize that there may be times when you will want to discuss a topic or problem outside of office hours. The best way to reach us is by Illinois email (jacksonn@illinois.edu). We will try to answer your questions via email, or we can schedule a time to meet if needed. We will do our best to respond to your emails as quickly as possible, but please do not expect an immediate reply. When you send emails to us regarding the course, please include course number in the header of your message; this will allow us to flag your messages, and thus respond to them more quickly.

V. Academic Misconduct

The University of Illinois at Urbana-Champaign expects its faculty, staff, students and guests to conduct themselves in accordance with the community values of civility, respect, and honesty; to maintain the highest level of integrity and exercise critical judgement in all dealings, decisions and encounters; and to maintain and strengthen the public's trust and confidence in our institution. It is the responsibility of each student to refrain from infractions of academic integrity, from conduct that may lead to suspicion of such infractions, and from conduct that aids others in such infractions. Students have been given notice of this Part by virtue of its publication. Regardless of whether a student has actually read this Part, a student is charged with knowledge of it. Ignorance is not a defense.

Academic misconduct (cheating, plagiarism, ….), as a form of fraud, undermines the public trust, both in the institution and in the degree. When you sign your name to work, you are stating that the work is yours, you created it or contributed to it, and you comprehend everything in it. **Academic misconduct of any sort will not be tolerated.** Instructors are required to report all suspected infractions of academic integrity in the online FAIR system that guides both the instructor and the student through the different phases of the process exactly as stated in the Student Code, including any appeals regarding the finding and/or the sanction. Additional text regarding academic integrity expectations can be found in the University Student Code, Article 1, Part 4: https://studentcode.illinois.edu/article1/part4/1-401/ and https://studentcode.illinois.edu/article1/part4/1-402/

VI. Disability Services

Students with disabilities will be appropriately accommodated and should inform the instructor as soon as possible of their needs. Disability Resources and Educational Services (DRES) is located at 1207 South Oak Street; telephone 333-1970; disability@illinois.edu; http://www.disability.illinois.edu. Please inform the instructor of any special accommodation needs at the start of the semester.

VII. Religious Observances

Illinois law requires the University to reasonably accommodate its students' religious beliefs, observances, and practices in regard to admissions, class attendance, and the scheduling of examinations and work requirements. You should examine this syllabus at the beginning of the semester for potential conflicts between the course deadlines and any of your religious observances. If a conflict exists, you should notify your instructor of the conflict and follow the procedure at https://odos.illinois.edu/community-of-care/resources/students/religious-observances/ to request appropriate accommodations. This should be done in the first two weeks of classes.

VIII. Dealing with Stress, Personal Issues

Counseling services are available to all our students here on campus. College can be stressful for a variety of reasons. The departments of chemistry and of chemical and biomolecular engineering believe your mental health is as important as your physical health and intellectual growth. *If you are feeling overwhelmed, depressed, or anxious, there are many resources on campus to assist you*. The Counseling Center offers same-day first time appointments, time-limited counseling, long-term group therapy, and several skill development workshops. Please call them at 217-333-3704 to make an appointment, or visit their website, www.counselingcenter.illinois.edu, for more information. If you are experiencing a mental health crisis and feel you are in immediate danger, please call 911. The Champaign County Crisis Line (217-359-4141) is also available 24 hours a day, 7 days a week, 365 days a year.

IX. Use of Electronic Devices and Class Courtesy

Cellphones and other electronics must be turned off during lecture. Laptops are allowed on only during in-class coding exercises that will be provided. In addition, talking and other behavior that distracts students will not be tolerated. The instructors respect the time of the student, and the same level of maturity is expected in return. Thank you!

X. Emergency

Please take a few minutes this week and learn the different ways to leave this building. If there's ever a fire alarm or something like that, you'll know how to get out, and you'll be able to help others get out too. If you have not already signed up for emergency text messages at emergency.illinois.edu, please do it this week. You'll receive information from the police and administration during emergency situations. If you have any questions, go to https://police.illinois.edu/, or call 217-333-1216. Please see police.illinois.edu/safe for more information on how to prepare for emergencies, including how to run, hide or fight and building floor plans that can show you safe areas.

Good luck with the course!

Prof. Jackson

Course Project: Advanced Topics in Chemical Data Science and Engineering

The frontier of machine learning changes at breakneck speed, so the lecture content of this course serves as a survey of the "greatest hits" of machine learning and builds the foundation to understand the methods emerging every day. To provide students a chance to dig deep into exciting new topics, we will have a final course project on an advanced topic not covered (or only minimally covered) in the course. I provide a (non-definitive) list of potential topics below. **Student groups are required to schedule a meeting with Prof. Jackson to discuss project topics.**

https://deepchem.io/tutorials/the-basic-tools-of-the-deep-life-sciences/

https://udlbook.github.io/udlbook/

Projects from Previous Years:

- Integrating LLM embeddings and LSTM's to predict B-factors of proteins from sequence.
- Isolation forests for anomaly detection of degradation in Li-ion batteries.
- Predicting the yields of Suzuki-Miyaura coupling reactions.
- Bayesian Optimization of molecular species.
- Active learning of cyclic voltammogram data.
- Graph Neural Networks for the enantiomeric systems
- Deep reinforcement learning for de novo small molecule design to maximize solubility

Other for Ideas for ML Techniques to Explore: Interpretable AI, symbolic regression, advanced nonlinear dimensionality reduction, diffusion models, variational autoencoders.

Critical Elements of the Project:

1. Identification of a potential methodological topic (above).
2. Identification of a domain-specific application (broadly defined across Chemistry, ChBE, Materials, etc) which you will apply this method to.
3. A simple example of a code utilizing the method applied to your problem of interest (with a real chemical dataset), **with this code uploaded to the course GitHub.**
4. A written report summarizing your topic, application, code, and findings.
5. A brief (15 minute) oral presentation of your results to the class.

Important Dates:

1. Group team members due: October 2nd
2. Group topics due: October 23rd
3. Written reports due: (both hardcopy reports and electronic code submissions are needed): December 4th
4. Class presentations: December 4th and December 9th

Team Size and Composition: You should organize yourselves into teams of 4 people. As upperclassmen and graduate students, you are responsible for organizing your own teams.

Written Reports: A written report submitted by the group will be required. The report should be ~10-15 pages, single-spaced, Times New Roman font, including figures (no more than five) and references (no

more than 20). Please begin with an abstract, not to exceed 250 words, which presents the rationale of the research, its scientific objective, and an estimate of the significance to the field of research if the objective is reached. Include pertinent literature citations, with titles.

Oral Presentations: Each team will present brief (15 minutes) slide set on their project. The overall idea is to share the story behind this recent breakthrough to your class. The goal of the presentation is to explain the complex science in an accessible manner using the concepts and ideas covered during the course. Every member of the team is required to speak during the presentation.

Grading Rubric for Final Project:

Oral Presentation (20 points)

- Understanding of topic (5 points).
- Application to a chemical/engineering system (5 points).
- Organization of presentation (5 points).
- Quality of oral communication (5 points).

Written Report (40 points)

- Background and motivation for the method and application (5 points).
- Technical description of the method being used (15 points).
- Functioning Jupyter notebook application to a chemical or engineering system that obeys all data science and machine learning protocols (e.g. cross-validation) learned in the course (15 points).
- Formatting, grammar, references, quality of writing (5 points).