COLLEGE OF ENGINEERING, UIC
CME594 APPLIED DATA SCIENCE FOR NON-COMPUTER SCIENTISTS
(4 CREDIT HOURS)

I. INSTRUCTOR & COURSE DETAILS

Instructor Name: Professor Sybil Derrible
Email address: derrible@uic.edu
Website: https://sybilderrible.com/
Drop-In Office Hours: open door policy / on demand
Drop-In Hours location: 2071 ERF or on Zoom

Blackboard Course Site: https://uic.blackboard.com/ultra/courses/_266168_1/cl/outline
Students are expected to log into the course site regularly to learn about any developments related to the course, upload assignments, and communicate with classmates. For all technical questions about Blackboard, email the Learning Technology Solutions team at LTS@uic.edu.

Course Modality and Schedule
This course is taught **ON CAMPUS**.

- Hours: Mondays and Wednesdays from 3:00 PM to 4:15 PM
- Location: Room 311 in Adams Hall

II. COURSE INFORMATION

Catalog Course Description and Prerequisite/corequisite Statement
Introduction and application of data science and machine learning techniques useful for research for non-computer scientists. One new technique is covered every week. The main assessment is a final paper for which the students are asked to pick any data set (preferably from their own research) and apply one or multiple techniques from the course.

Growth Mindset:
Course materials and assignments can be complex and challenging, but they are crucial to your intellectual and personal growth and development. There are times you may need extra help. Students who attend class consistently, complete all assignments, thoughtfully engage with feedback on work, develop good study strategies, visit the tutoring center, and contact faculty when they are struggling can develop a thorough understanding of the course material and ultimately succeed in the course!

Course Goals and Learning Outcomes

Course Learning Outcomes:

This course aims to provide students with introductory knowledge of several data science techniques that can be used for data analysis. The material learned should then be useful in the student's own research. More specifically, at the end of this course, students should be able to:

1. explain the main concepts behind all the techniques covered
2. identify the type of technique preferable to use depending on the type of data to analyze
3. use the various Python libraries learned to be able to apply these techniques
4. apply rigorously one or multiple of these techniques learned in their own research

General Education Learning Outcomes:
TBD

Course type:
Elective

Brief list of topics to be covered:
This course exposes students to one new technique per week. Topics and techniques covered include: k-nearest neighbor, principal component analysis, K-means and hierarchical clustering, decision tree learning, random forest, gradient boosting, neural networks, deep learning, and text mining.

Required and Recommended Course Materials
No textbook is required, but the following books will be used and are free to download:

- Han, J., Kamber, M., Pei, J., 2011, Data Mining: Concepts and Techniques, *Elsevier Science*, Waltham, MA
- James, G., Witten, D., Hastie, T., Tibshirani, R., 2017, An Introduction to Statistical Learning with Applications in R, *Springer*, New York, NY
- Murphy, K., 2012, Machine Learning: A Probabilistic Perspective, *MIT Press*, Cambridge, MA (not used in the class but a comprehensive resource)

Required Technology (all freely accessibly)

- Python 3.7.x: https://www.python.org/downloads/
- Libraries: NumPy, SciPy, Pandas, SciKit learn, Keras (for Windows, see: http://www.lfd.uci.edu/~gohlke/pythonlibs/)

or preferably use/install

- Google CoLab: https://colab.google/
- Jupyter Notebook: https://jupyter.org/
- Spyder (Anaconda) – Python 3.x (recommended package that includes Python and most recommended libraries; sometimes the 32bit version works better even for 64bit computers): https://www.spyder-ide.org/.

Respect for Copyright
Please protect the copyright integrity of all course materials and content. Please do not upload course materials not created by you onto third-party websites or share content with anyone not enrolled in our course.

III. COURSE POLICIES & CLASSROOM EXPECTATIONS

Grading Policy and Point Breakdown
Attendance, participation, behavior (15%)
Homework (30%)
Teaching Assignment (15%)
Abstract (5%)
Poster Presentation (5%)
Final Paper (30%)

Policy for Missed or Late Work
Missed work will be given a 0 grade. Work submitted late will be deducted 5% per day that it is late. For example, if it is submitted two days late, 10% will be deducted. Work submitted more than 20 days late will be given a 0 grade.

Attendance / Participation Policy
Attendance is mandatory. Please email me if you face an unexpected situation that may impede your attendance, participation in required class and exam sessions, or timely completion of assignments.

Academic Integrity
As a student and member of the UIC community, you are expected to adhere to the Community Standards of academic integrity, accountability, and respect. Please review the UIC Student Disciplinary Policy for additional information.

Email Expectations
Students are responsible for all information instructors send to your UIC email and Blackboard accounts. Faculty messages should be regularly monitored and read in a timely fashion.

IV. COURSE SCHEDULE & READINGS

Weekly Schedule of Class Topics, Assignments, Assessments, Due Dates, and Deadlines

---

Week 1: Python Installation, Tutorial, and Data Exploration

- Install Python and recommended libraries (see software section above)

Readings:

- James, G., Witten, D., Hastie, T., Tibshirani, R., 2017, "Chap. 1 Introduction" in *An Introduction to Statistical Learning with Applications in R*, Springer
- Han, J., Kamber, M., and Pei, J., 2011, "Chap. 2 Getting to Know Your Data" in *Data Mining: Concepts and Techniques*, Elsevier Science.
- Emerson, J., Green, W.A., Schloerke, B., Crowley, J., Cook, D., Hofmann, H., Wickham, H., 2012, "The Generalized Pairs Plot", *Journal of Computational and Graphical Statistics*, 22(1):79–91.

  And

- Learnpython.org: https://www.learnpython.org/ (Learn the Basics and Data Science Tutorials)
  or
- A Byte of Python: https://python.swaroopch.com/

Week 2: Statistical Learning, k-Nearest Neighbor, and Data Cleaning

Readings:

- Pythonprogramming (video), 2016, "Intro to Machine Learning with Scikit Learn and Python", pythonprogramming.net
- James, G., Witten, D., Hastie, T., Tibshirani, R., 2017, "Sec. 2.1 What Is Statistical Learning?" in Chap. 2 Statistical Learning in *An Introduction to Statistical Learning with Applications in R*, Springer
- Han, J., Kamber, M., and Pei, J., 2011, "Sec. 3.2 Data Cleaning" in Chap. 3 Data Preprocessing in *Data Mining: Concepts and Techniques*, Elsevier Science.
- Han, J., Kamber, M., and Pei, J., 2011, "Sec. 9.5.1 k-Nearest-Neighbor Classifiers" in Chap. 9 Classification: Advanced Methods in *Data Mining: Concepts and Techniques*, Elsevier Science.

Supplementary Readings:

- Han, J., Kamber, M., Pei, J., 2011, "Chap. 1 Introduction", in *Data Mining: Concepts and Techniques*, Elsevier Science.
- Han, J., Kamber, M., Pei, J., 2011, "Chap. 3 Data Preprocessing", in *Data Mining: Concepts and Techniques*, Elsevier Science.
- Sehn Korting, T. (video), 2014, "How kNN algorithm works", YouTube.com

---

Week 3: Linear Regression, Logistic Regression, and Model Validation

Readings:

- James, G., Witten, D., Hastie, T., Tibshirani, R., 2017, "Sec. 3.1 Simple Linear Regression" in Chap. 3 Linear Regression in *An Introduction to Statistical Learning with Applications in R*, Springer
- James, G., Witten, D., Hastie, T., Tibshirani, R., 2017, "Sec. 4.3 Logistic Regression" in Chap. 4 Classification in *An Introduction to Statistical Learning with Applications in R*, Springer
- James, G., Witten, D., Hastie, T., Tibshirani, R., 2017, "Sec. 2.2 Assessing Model Accuracy" in Chap. 2 Statistical Learning in *An Introduction to Statistical Learning with Applications in R*, Springer
- Han, J., Kamber, M., and Pei, J., 2011, "Sec. 8.5 Model Evaluation and Selection" in Chap. 8 Classification: Basic Concepts in *Data Mining: Concepts and Techniques*, Elsevier Science.
- Gupta, P., 2017, "Cross Validation in Machine Learning", Towards Data Science

---

Week 4: Principal Component Analysis

Readings

- Smith, L. I., 2002, "A tutorial on Principal Components Analysis", Notes for Course COSC453

Supplementary Readings:

- scikit-learn, n.d., "2.5. Decomposing signals in components (matrix factorization problems)", scikit-learn.org

---

Week 5: Clustering Analysis

Readings:

- Han, J., Kamber, M., and Pei, J., 2011, "Chap. 10 Cluster Analysis: Basic Concepts and Method" in *Data Mining: Concepts and Techniques*, Elsevier Science.
- scikit-learn, 2016, "2.3 Clustering", scikit-learn.org

Supplementary Readings:

- Tan, P-N., Steinbach, M., Kumar, V., 2006, "Chapter 8. Cluster Analysis: Basic Concepts and Algorithms", in *Introduction to Data Mining*, Pearson, pdf of chapter and slide accessible at http://www-users.cs.umn.edu/~kumar/dmbook/index.php

---

Week 6: Support Vector Machine

Readings

- Han, J., Kamber, M., and Pei, J., 2011, "Section 9.3 Support Vector Machine", in Chap. 9 Classification: Advanced Methods in *Data Mining: Concepts and Techniques*, Elsevier Science.

- Berwick, R., 2009, "An Idiot's guide to Support vector machines (SVMs)", Notes for Course CAP 6412 (Advanced Computer Vision)
- Udiprod, 2007, "SVM with polynomial kernel visualization", YouTube.com
- scikit-learn, n.d., "4.2. Permutation feature importance", scikit-learn.org
- scitkit-learn, n.d., "4.1. Partial Dependence and Individual Conditional Expectation plots", scikit-learn.org

Supplementary Readings:

- scikit-learn, n.d., "1.4. Support Vector Machines", scikit-learn.org
- Sehn Korting, T. (video), 2014, "How SVM (Support Vector Machine) algorithm works", YouTube.com
- Molnar, C., n.d., 8.5 Permutation Feature Importance", christophm.github.io

---

Week 7: Decision Tree Learning, Tree-Based Feature Importance, and Permutation Importance

Readings

- Han, J., Kamber, M., and Pei, J., 2011, "Chap. 8 Classification: Basic Concepts" in *Data Mining: Concepts and Techniques*, Elsevier Science.
- Wang, T., 2011, "Information & Entropy", Slides for Comp 595 DM
- Ronaghan, S., 2018, "The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark", Towards Data Science

Supplementary Readings:

- Khanacademy (video), 2016, "Information Theory", khanacademy.org
- scikit-learn, n.d., "1.10. Decision Trees", scikit-learn.org
- Catalano, M., Leise, T., and Pfaff, T., 2009, "Measuring Resource Inequality: The Gini Coefficient", *Numeracy*, 2(2), DOI: http://dx.doi.org/10.5038/1936-4660.2.2.4

Week 8: Random Forest, Gradient Boosting, and SHAP

- Install the shap library.

Readings

- Yiu, T., 2019, "Understanding Random Forest", Towards Data Science
- Bowers, M., 2020, "How to Build a Gradient Boosting Machine from Scratch", blog.mattbowers.dev
- Mazzanti, S., 2020, "SHAP Values Explained Exactly How You Wished Someone Explained to You", Towards Data Science
- Luvsandorj, Z., 2021, "Explaining Scikit-learn models with SHAP", Towards Data Science

Supplementary Readings:

- Wood, T., n.d., "What is Random Forest?", DeepAI

- Normalized Nerds, 2021, "Random Forest Algorithm Clearly Explained!", YouTube.com
- Brownlee, J., 2016, "A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning", machinelearningmastery.com
- StatQuest, 2019, "Gradient Boost Part 1 (of 4): Regression Main Ideas", YouTube.com
- Saini, A., 2021, "Gradient Boosting Algorithm: A Complete Guide for Beginners", Analytics Vidhya

---

Week 9: Backpropagation Neural Networks and Partial Dependence Plots

Readings

- Han, J., Kamber, M., and Pei, J., 2011, "Section 9.2 Classification by Backpropagation", in Chap. 9 Classification: Advanced Methods in *Data Mining: Concepts and Techniques*, Elsevier Science.
- Welch Labs (video), 2014, "Neural Networks Demystified", Part 1 to Part 7, YouTube.com

Supplementary Readings:

- scikit-learn, n.d., "1.17. Neural network models (supervised)", scikit-learn.org
- Shiffman, D., 2012, "Chapter 10. Neural Networks" in *The Nature of Code*, Creative Commons Attribution-NonCommercial 3.0 Unsupported License, ISBN: 0985930802

Week 10: Deep Learning (CNN, LSTM, GRU)

- Install the Keras library.

Readings:

- Khan, S., Rahmani, H., Ali Shah, S.A. and Bennamoun, M., 2018, "Chap. 4 Convulationanl Neural Networks" in *A Guide to Convolutional Neural Networks for Computer Vision*, *Synthesis Lectures on Computer Vision*, 8(1):43-68
- dshahid380, 2019, "Convolutional Neural Network: Learn Convolutional Neural Network from basic and its implementation in Keras", Towards Data Science
- Phi, M., 2018, "Illustrated Guide to Recurrent Neural Networks", Towards Data Science
- Phi, M., 2018, "Phi, M., 2018, "Illustrated Guide to Recurrent Neural Networks", Towards Data Science", Towards Data Science

---

Week 11: Spring Break

- No class

Week 12: Abstract Presentation

- No readings / Abstract Presentation

Week 13: Text Mining

- Install the nltk library, nltk data, the TextBlob library, and the gensim library.

Readings:

- Bird, S., Klein, E., Loper, E., 2010, "Chap. 1 Language Processing and Python" in *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, O'Reilly Media, Inc, Sebastopol, CA
- Lettier, 2018, "Your Easy Guide to Latent Dirichlet Allocation", medium.com
- Keitakurita, 2017, "A Practical Introduction to NMF (non-negative matrix factorization)", mlexplained.com
- Napitupulu, J., 2015, "Text Learning with scikit-learn", available at https://napitupulu-jon.appspot.com/posts/text-learning-ud120.html

Supplementary Readings:

- Doig, C., 2015, "Introduction to Topic Modeling in Python", PyTexas 2015, available at https://chdoig.github.io/pytexas2015-topic-modeling/#/
- Zhai, C., 2017, "Text Mining and Analytics", coursera course, available at https://www.coursera.org/learn/text-mining

Week 14: Generalized Regression and Probabilistic Neural Networks

- Install the Python library NeuPy

Readings

- McCormick, C., 2013, "Radial Basis Function Network (RBFN) Tutorial", mccormickml.com
- macheads101 (video), 2017, "RBF Networks", YouTube.com
- Hong, X., n.d., "Probabilistic neural network", reading.ac.uk
- Xoaxdotnet (video), 2009, "Neural Networks Lesson 2: Probabilisitic Neural Networks", YouTube.com

Supplementary Readings:

- Touretzky, D.S., 2006, "Radial Basis Functions", cs.cmu.edu
- Specht, D., 1990, "Probabilistic neural networks", Neural Networks, 3(1):109-118

Week 15: Fisher Information

- Download Fisher Information library at http://csun.uic.edu/codes/fisher.html

Readings

- Ahmad, N., et al., 2016, "Using Fisher information to track stability in multivariate systems", *Royal Society Open Science*, 3:160582

Supplementary Readings:

- Karunanithi, A.T., Cabezas, H., Frieden, B.R., Pawlowski, C.W., 2008, "Detection and assessment of ecosystem regime shifts from fisher information", Ecology and Society, 13(1):22

Week 16: Final Presentation

- No readings / Final Presentation

Disclaimer
This syllabus is intended to give the student guidance on what may be covered during the semester and will be followed as closely as possible. However, as the instructor, I reserve the right to modify, supplement, and make changes as course needs arise. I will communicate such changes in advance through in-class announcements and in writing via Blackboard Announcements.

V. ACCOMMODATIONS

Disability Accommodation Procedures
UIC is committed to full inclusion and participation of people with disabilities in all aspects of university life. If you face or anticipate disability-related barriers while at UIC, please connect with the Disability Resource Center (DRC) at drc.uic.edu, via email at drc@uic.edu, or call (312) 413-2183 to create a plan for reasonable accommodations. To receive accommodations, you will need to disclose the disability to the DRC, complete an interactive registration process with the DRC, and provide me with a Letter of Accommodation (LOA). Upon receipt of an LOA, I will gladly work with you and the DRC to implement approved accommodations.

Religious Accommodations
Following campus policy, if you wish to observe religious holidays, you must notify me by the tenth day of the semester. If the religious holiday is observed on or before the tenth day of the semester, you must notify me at least five days before you will be absent. Please submit this form by email with the subject heading: **"YOUR NAME: Requesting Religious Accommodation."**

VI. CLASSROOM ENVIRONMENT

Inclusive Community
UIC values diversity and inclusion. Regardless of age, disability, ethnicity, race, gender, gender identity, sexual orientation, socioeconomic status, geographic background, religion, political ideology, language, or culture, we expect all members of this class to contribute to a respectful, welcoming, and inclusive environment for every other member of our class. If aspects of this

course result in barriers to your inclusion, engagement, accurate assessment, or achievement, please notify me as soon as possible.

Name and Pronoun Use
If your name does not match the name on my class roster, please let me know as soon as possible. My pronouns are *[she/her; he/him; they/them]*. I welcome your pronouns if you would like to share them with me. For more information about pronouns, see this page: https://www.mypronouns.org/what-and-why.

Community Agreement/Classroom Conduct Policy
- Be present by turning off cell phones and removing yourself from other distractions.
- Be respectful of the learning space and community. For example, no side conversations or unnecessary disruptions.
- Use preferred names and gender pronouns.
- Assume goodwill in all interactions, even in disagreement.
- Facilitate dialogue and value the free and safe exchange of ideas.
- **Try not to make assumptions, have an open mind, seek to understand, and not judge.**
- Approach discussion, challenges, and different perspectives as an opportunity to "think out loud," learn something new, and understand the concepts or experiences that guide other people's thinking.
- Debate the concepts, not the person.
- Be gracious and open to change when your ideas, arguments, or positions do not work or are proven wrong.
- Be willing to work together and share helpful study strategies.
- Be mindful of one another's privacy, and do not invite outsiders into our classroom.

Content Notices and Trigger Warnings
Our classroom provides an open space for a critical and civil exchange of ideas, inclusive of a variety of perspectives and positions. Some readings and other content may expose you to ideas, subjects, or views that may challenge you, cause you discomfort, or recall past negative experiences or traumas. I intend to discuss all subjects with dignity and humanity, as well as with rigor and respect for scholarly inquiry. If you would like me to be aware of a specific topic of concern, please email or visit my Student Drop-In Hours.

VII. RESOURCES: Academic Success, Wellness, and Safety

We all need the help and the support of our UIC community. Please visit my **drop-in hours** for course consultation and other academic or research topics. For additional assistance, please contact your assigned college advisor and visit the support services available to all UIC students.

Academic Success
- UIC Tutoring Resources
- College of Engineering tutoring program

- Equity and Inclusion in Engineering Program
- UIC Library and UIC Library Research Guides.
- Offices supporting the UIC Undergraduate Experience and Academic Programs.
- Student Guide for Information Technology
- First-at-LAS Academic Success Program, focusing on LAS first-generation students.

Wellness

- **Counseling Services**: You may seek free and confidential services from the Counseling Center at https://counseling.uic.edu/.
- Access U&I Care Program for assistance with personal hardships.
- **Campus Advocacy Network**: Under Title IX, you have the right to an education that is free from any form of gender-based violence or discrimination. To make a report, email TitleIX@uic.edu. For more information or confidential victim services and advocacy, visit UIC's Campus Advocacy Network at http://can.uic.edu/.

Safety

- UIC Safe App—PLEASE DOWNLOAD FOR YOUR SAFETY!
- UIC Safety Tips and Resources
- Night Ride
- Emergency Communications: By dialing 5-5555 from a campus phone, you can summon the Police or Fire for any on-campus emergency. You may also set up the complete number, (312) 355-5555, on speed dial on your cell phone.