# CSCI 4364/6364 Machine Learning

**Fall 2023**

Section 81

3 Credits

Syllabus v.20230824b

THE GEORGE
WASHINGTON
UNIVERSITY

WASHINGTON, DC

## Instructor

John Sipple
jsipple@gwu.edu

**Research Interests**: Deep Reinforcement Learning in real-world applications, Explainability, Anomaly Detection, IoT, Generative Language Models

Staff Software Engineer
Core Enterprise Machine Learning at Google

Adjunct Professor for Computer Science
The George Washington University

## Grader

Mike Rossetti
rossetti@gwu.edu

**Research Interests**: Applied Machine Learning, Unsupervised Learning , Natural Language Processing , Systems Development Methodologies, Software Testing and Verification, Information Security and Privacy, Social and Information Networks

## Time and Location

Tuesday, Thursday
9:35 am - 10:50 am,
8/24 - 12/11/2023,
Tompkins 208

## Office Hours

Wednesday
9:00 - 10:00 am via Zoom
(or by appointment)

## HW Office Hours

Wednesday
7:00 - 8:30 pm via Zoom

# General Information

## Description

This course presents a broad overview of Machine Learning, covering conceptual and theoretical foundations, and hands-on application. This course is organized into four phases. In Phase 1 (Fundamentals), we will lay the groundwork with evaluation methods, mathematics, probability, and information theory and familiarize ourselves with Python and machine learning software libraries used in industry and academics today. In Phase 2 (Supervised Methods), we will study ML techniques that learn from labeled data, covering linear, non-linear and deep learning methods. In Phase 3 (Unsupervised Methods), we'll explore unsupervised techniques like k-nearest neighbor search, clustering and anomaly detection that do not train from labeled data. In Phase 4 (Advanced Topics), we'll explore important, contemporary topics like convolutional and recurrent neural networks, autoencoders, explainability, and reinforcement learning, large language models, and social aspects of AI.

Students will have the opportunity to train and evaluate several different types of machine learning models and present their results in practical homework assignments and term projects.

No experience in Python or machine learning is required; however, since this is a hands-on course, students will submit individual homework assignments and projects in the Python programming language using **Google Colaboratory**, **Scikit-Learn**, **TensorFlow**, and **Keras**.

Students will be able to pursue advanced projects aligned with their current research, or choose from a list of projects provided by the instructor. **At the end of the course, students will be equipped with enough knowledge to follow and pursue their own research in machine learning, and apply machine learning solutions to real-world problems**.

## Objectives

1. Students will gain sufficient knowledge of the fundamentals of machine learning that will enable them to follow and pursue research in machine learning.
2. Students will have enough knowledge of the fundamentals to develop unique and effective solutions in research or industry.
3. Students will gain hand-on experience with core machine learning libraries, such as Scikit-Learn, TensorFlow and Keras.

## Prerequisites

While no experience in programming Python or machine learning is required, students should have had coursework in algorithms, linear algebra, calculus, and probability. Additionally, students should be able to program in at least one object-oriented programming language.

- Analysis of algorithms (CSCI 3212 or 6212),
- Probability theory or statistics (e.g. CSCI 3362/6362)
- Linear algebra (MATH 2184), or equivalent.
- Coursework, familiarity, and experience in at least one object-oriented programming language, such as Python, Java, C++, or C#.

Students with non-traditional preparation should email the instructor and/or get approval from the CS department.

## Lectures

Lectures are in-person, but Zoom recordings and slides will be uploaded to Blackboard within 48 hours of the lecture. In general, students should use the recordings to review lectures and for exceptional circumstances when they cannot attend, not as an alternative to regular in-person attendance. Students are strongly encouraged to attend the lectures in person, since it provides an opportunity to ask questions and obtain immediate feedback.

# Course textbooks and supplementary resources

Since no single textbook provides a complete picture on the topic, this course has four required textbooks, several papers, and a few optional resources. All required textbooks and papers are available for free online, or can be purchased as e-books or hardcopy on Amazon.

**The Elements of Statistical Learning: Data Mining, Inference, and Prediction**, Second Edition (Springer Series in Statistics) 2nd Edition by Hastie, Tibshirani, Friedman

- Foundational supervised and unsupervised learning.
- Hard copy available on Amazon (hardcover or eTextbook)
- Online (free) at http://www-stat.stanford.edu/ElemStatLearn

**Deep Learning (Adaptive Computation and Machine Learning series)** by Goodfellow, Bengio, Courville (Goodfellow)

- Deep Learning topics and foundational math and statistics.
- Hard copy available on Amazon or at the GW bookstore
- Online (free) at https://www.deeplearningbook.org/

**Dive into Deep Learning** (D2L)

- An interactive deep learning book with code, math, and discussions.
- Contains latest developments in Deep Learning
- Available on-line (free) at: https://d2l.ai/

**Reinforcement Learning: An Introduction**, 2nd Ed, by Sutton and Barto (Sutton)

- Introduction and topics on Reinforcement Learning

- Available on-line (free) at: https://www.andrew.cmu.edu/course/10-703/textbook/BartoSutton.pdf

Additionally, this course has an optional textbook that provides hands-on programming along with a solid background of machine learning:

**Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems** 3rd Edition, 2023 (O'Rielly) by Géron

- Available on Amazon
- (2nd Ed) https://www.knowledgeisle.com/wp-content/uploads/2019/12/2-Aur%C3%A9lien-G%C3%A9ron-Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-Tensorflow_-Concepts-Tools-and-Techniques-to-Build-Intelligent-Systems-O%E2%80%99Reilly-Media-2019.pdf

Alternatively, students may reference some on-line references of the machine learning libraries we will be using throughout the course, including:

- Scikit-Learn (https://scikit-learn.org) Provides detailed tutorials and hands-on guides.
- Keras (https://keras.io/) has excellent documentation on APIs, and provides detailed guides and examples.
- TensorFlow (https://www.tensorflow.org/) Provides guides and tutorials of the TensorFlow library.

Students that have had little or no experience in the Python programming language are encouraged to review and refer to the on-line Python documentation, guides and tutorials available at https://www.python.org/doc/ .

There will be several papers published in research conferences and journals for some of the advanced topics (see schedule below).

## Supplementary Resources

| Abbreviation | Paper |
|---|---|
| Alammar2020 | The Illustrated Transformer http://jalammar.github.io/illustrated-transformer/ |
| Géron | Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 2nd Edition (O'Rielly) by Géron https://www.knowledgeisle.com/wp-content/uploads/2019/12/2-Aur%C3%A9lien-G%C3%A9ron-Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-Tensorflow_-Concepts-Tools-and-Techniques-to-Build-Intelligent-Systems-O%E2%80%99Reilly-Media-2019.pdf |
| Ruff2020 | A Unifying Review of Deep and Shallow Anomaly Detection (2020), Ruff, L., Kauffmann, J., Vandermeulen, R., Montavon, G., Kloft, M., Dietterich, T., Müller, K.R. https://arxiv.org/abs/2009.11732 |
| Samek2020 | Toward Interpretable Machine Learning: Transparent Deep Neural Networks and Beyond (2020), Samek, |

| | |
|---|---|
| | W., Montavon, G., Lapuschkin, S., Anders, C., Müller, K.R. https://arxiv.org/abs/2003.07631 |
| Schubert2017 | DBSCANRevisited, Revisited: WhyandHowYouShould(Still) Use DBSCAN (2017), ACMTransactions on Database Systems, Vol. 42, No. 3, Article 19. Publication date: July 2017 https://www.ccs.neu.edu/home/vip/teach/DMcourse/2_cluster_EM_mixt/notes_slides/revisitofrevisitDBSCAN.pdf |
| Sipple2020 | Interpretable, Multidimensional, Multimodal Anomaly Detection with Negative Sampling for Detection of Device Failure https://proceedings.mlr.press/v119/sipple20a.html |
| Vaswani2017 | Attention is all you need (2017) https://arxiv.org/abs/1706.03762 |

# Assignments and Grading

**Coding Standards**
Throughout this course, we will use Google's Python Style Guide (https://google.github.io/styleguide/pyguide.html) as the coding standard for homework and project submission. A big part of machine learning in industry is applying good programming practices.

**25% Homework**: Assignments will be mostly hands-on solving a machine learning problem in the ML programming environment, and students will be expected to submit functional and well-documented **Google Colaboratory notebooks**. While collaboration is permitted between students, each student is expected to submit their own solution. Each assignment is graded on execution (50%), correctness (30%), and readability (20%). **There will be six homework assignments provided, of which the top five grades will be used for grading.** Each student will submit their colab notebooks the deadline in Blackboard. Note that for Homework 5 and 6, students will need additional compute resources available with Colab Pro or Colab Pro+ (https://colab.research.google.com/signup).

**20% Midterm Exam**: The exam is focused on the student's ability to apply and understand the fundamentals of machine learning, covering lectures through supervised learning.The exam contains scenario-based essay (long answer) questions. The exam will be limited to topics discussed in class (i.e., no curve balls or trick questions), so to prepare properly,you should review the slides/recordings, required (and optional) readings, and homework assignments. **The exam is open-book, and you may bring electronic devices like laptop, ipad, etc.** However, no remote assistance via the internet (texting, chat, etc.) is permitted.

**30% Group Project**: Students will work as a team on a machine learning project based on a real-world machine learning task. Students must submit a project proposal and receive concurrence from the instructor. Graduate students are encouraged to work on an advanced topic with the goal of publishing their work at a peer-reviewed conference. Group and individual projects will be presented to the course participants, and invited guests from industry and academia. At the end of the course, students will submit a final report, which will account for the full project grade.

Students may choose machine learning term projects from three options:

(1) One to five students may participate in a project that augments or extends current research work of at least one student, if:
    - The proposed work has not been completed at the start of the semester, and
    - includes one or more of the topics covered in the schedule, and
    - is aligned with their graduate research work.

(2) Three to five students project proposed by the instructor on a real-world project:
    - Energy optimization using RL

   - Intelligent Diagnostics using Anomaly Detection, Explainable AI, and Large Language Models
   - Additional topics to be announced in the first weeks of the course.

(3) Three to five students may enter an official on-line ML competition, such as Kaggle. Students will have to submit their model's results and have a ranking on the competition leaderboard.

**25% Final Exam**: The final exam will cover the fundamentals of unsupervised machine learning and the advanced topics presented in the second half of the course (Lectures 10 - 26). The format will be the same as the mid-term exam.

# Course Schedule

| Wk | L | Date | Topic | Reading | Optional Reading | Due |
|---|---|---|---|---|---|---|
| **1** | **1** | 8/24 | Overview and history, paradigms, applications, ML programming environment | | | |
| **2** | **2** | 8/29 | Linear Algebra and Principal Components Analysis<br>**Overview of the final project.** | Goodfellow 1, 2 | D2L: 2.3 | |
| **2** | **3** | 8/31 | Uncertainty, Random Variables, Probability and Information Theory | Goodfellow 3 | D2L:2.6 | |
| **3** | **4** | 9/5 | Principal Components Analysis | Goodfellow 3 | Géron 8 | |
| **3** | **5** | 9/7 | Multivariate Calculus, Chain Rule, Gradient, and Optimization | Goodfellow 4 | D2L: 2.4, 2.5 | |
| **4** | **6** | 9/12 | Linear Regression: Least Squares and Regularization | Hastie: 2.1-2.3, 2.9, 3.1 - 3.4 | D2L: 3 | Homework 1: Model Training and Evaluation |
| **4** | **7** | 9/14 | Linear Classification | Hastie: 4.1, 4.2, 4.4 | D2L: 4<br>Géron 3 | |
| **5** | **8** | 9/19 | Additive Models, Trees and Related Methods, Random Forests | Hastie: 9.1, 9.2, 15.1-15.3 | Géron 6, 7 | |
| **5** | **9** | 9/21 | Support Vector Machines and Kernels | Hastie: 12.1-12.3 | Géron 5 | Project teams and proposals. |
| **6** | **10** | 9/26 | Deep Feedforward Networks | Goodfellow 6 | Géron 10 | Homework 2: Regression |
| **6** | **11** | 9/28 | Backpropagation for Deep Learning | Goodfellow 6 | Géron 10,<br>D2L: 22.4 | |
| **7** | **12** | 10/3 | Regularization for Deep Learning | Goodfellow 7 | Géron 11 | |
| **7** | **13** | 10/5 | Optimization for Deep Learning | Goodfellow 8 | Géron 11<br>D2L:12 | |
| **8** | | 10/10 | Mid-Term Examination (Lectures 1 - 9) | | | |
| **8** | | 10/12 | No Class - Fall Break | | | |
| **9** | **14** | 10/17 | Metric Properties, Measures of Association, k-Nearest Neighbors | Hastie 14.3.1-14.3.11 | Géron 9 | Homework 3: Multiclass Classification |
| **9** | **15** | 10/19 | Cluster Analysis: K-Means, Hierarchical, DBScan | Hastie 14.3.12, Schubert2017 | Géron 9 | |
| **10** | **16** | 10/24 | Anomaly Detection | Ruff2020 | Sipple2020 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| **10** | **17** | 10/26 | Fundamentals of Convolutional Neural Networks | Goodfellow 9 | Géron 14<br>D2L: 7 | Homework 4: Cluster Analysis |
| **11** | **18** | 10/31 | Computer Vision with Convolutional Neural Networks | Goodfellow 9 | Géron 14<br>D2L:8 | |
| **11** | **19** | 11/2 | Sequence Models with Recurrent and Recursive Networks | Goodfellow 10 | Géron 15<br>D2L: 9, 10 | |
| **12** | **20** | 11/7 | Attention and the Transformer Architecture | Vaswani2017<br>D2L: 11 | Géron 17<br>Alammar2020 | |
| **12** | **21** | 11/9 | Autoencoders and Generative Adversarial Networks | Goodfellow 14 | D2L: 20 | Homework 5: CNN |
| **13** | **22** | 11/14 | Explainable AI: SHAP, LIME, Integrated Gradients | | | |
| **13** | **23** | 11/16 | Large Language Models: Design and Application | TBD | TBD | |
| **14** | | 11/21 | No Class - Thanksgiving Break | | | |
| **14** | | 11/23 | No Class - Thanksgiving Break | | | |
| **15** | **24** | 11/28 | Reinforcement Learning: Elements, Bandits, Markov Decision Process, Q-Learning | Sutton: 1, 3 | Géron 18<br>D2L: 17 | Homework 6 Sequence-to-sequence models with Transformer |
| **15** | **25** | 11/30 | Reinforcement Learning: Policy and Value Iteration, Policy Gradient, Actor-Critic | Sutton: 6, 13 | Géron 18 | |
| **16** | **26** | 12/5 | AI and Society: Ethics, Fairness, Privacy and Intellectual Property | TBD | TBD | |
| **16** | | 12/7 | Group Project Presentation and Demos | | | |
| **17** | | 12/12 | Make-up day | | | |
| **17** | | 12/13-12/19 | Final Exam (Lectures 10 - 26) | | | |

# Academic Honesty

If you feel pressured about an assignment, please email the instructor instead of cheating. All work that you submit in this course for a grade should be your own (or the work of your group, whose names you have documented). If we detect cheating, we reserve the right to assign the student a 0 on the assignment, or an F in the course for more egregious violations. We will also be using automated software to check for cheating with code that is submitted to us.

Group assignments require collaboration within each group, but no collaboration between groups is permitted. Please refer to the academic integrity policy linked from the course web page. This policy will be strictly enforced. If you're having significant trouble with an assignment, please contact the instructors. Please see: Academic Integrity Policy