MA4635/DS4635
DATA ANALYTICS AND STATISTICAL LEARNING

The focus of this class will be on statistical learning: the intersection of applied statistics and modeling techniques used to analyze and to make predictions and inferences from complex real-world data. Topics covered include regression; classification/clustering; sampling methods (bootstrap and cross validation); and decision tree learning. Recommended background: Linear Algebra (MA2071 or equivalent), Applied Statistics II (MA2612 or equivalent), Probability (MA2631 or MA2621 or equivalent). The ability to write computer programs in a scientific language is assumed.

Where and When:
MTRF 9am-9:50am
Fuller Labs 320

Instructor information

Prof. Randy Paffenroth
Office location: UH 364 (Unity Hall)
Office hours: Mondays and Fridays 10am-11am (right after class). Also, I attempt to arrive at every class a half-hour early. **Other times are available by appointment, and I am always happy to meet.**

Best ways to contact me:

- WPI email: rcpaffenroth@wpi.edu

I should be able to turn around email questions relatively quickly 9am-5pm, Monday-Friday. My availability at night and on weekends is more limited and I certainly check my email far more infrequently, but you may feel free to try and contact me.

Teaching Assistant/Grader

Dashiell Lipsey
Office hours:
UH341
Mondays 2-3pm
Wednesdays 2-3pm

High level course goals and learning objectives.

By the end of the class, you should be able to:

- *Use tools* such as regression, classification, clustering, resampling, decision trees, etc. for making predictions from data.
- *Avoid* common pitfalls such as overfitting and data snooping.
- *Be able to assess* the validity of the prediction using cross validation and ensemble learning.
- *Diagnose* what can go wrong with a prediction.

Recommended background for course

The recommended background for the course are: Linear Algebra (MA2071 or equivalent), Applied Statistics II (MA2612 or equivalent), Probability (MA2631 or MA2621 or equivalent). The ability to write computer programs in a scientific language is assumed.

In particular, you will need to know some linear algebra:

- Vectors (that they can represent points in space, column vs. row, etc.)
- Matrices (transposes, that they don't commute, etc.)
- Inner products
- Least squares
- How to solve linear systems
- etc.

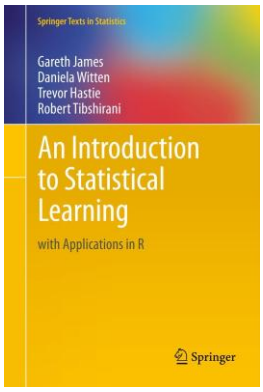
You will also need to know some probability and statistics

- Random variables (what they represent, etc.)
- Descriptive statistics (mean, variance, etc.)
- Hypothesis testing
- Estimation and prediction
- etc.

You will need to be able get your hands dirty playing with, processing, and plotting data using the **Python** computer language! The textbook uses **Python,** the *homework* uses **Python,** and that will be the officially supported language for the course and all lecture examples will be in **Python**. Now, this is not intended to be a programming course (i.e., your code will not be graded), but actually working with data will be extremely important (i.e., the *results* of the code will be graded)!

Textbook

An Introduction to Statistical Learning
Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
Downloadable at: [An Introduction to Statistical Learning (statlearning.com)]
(Be sure to get the Python version)



Recommended texts

Other texts that would be useful for the course are:

Secondary text for class:

- **Learning From Data**, by Yaser S. Abu Mostafa, Malik Magdon Ismail, and Hsuan Tien Lin. This book is used in the Caltech “Learning from Data” course and does a great job covering things like cross validation and VC dimension.

More advanced text:

- **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. This is the “big brother” of our textbook, and a great resource that covers a lot of interesting material.

Background texts that would be useful for the course are:

- **Linear Algebra and Its Applications**, by David Lay. This has been used as the textbookfor MA2071 (one of the requirements for the course).
- **Applied Statistics for Engineers and Scientists**, by Joseph Petruccelli, Balgobin Nandram, and Minghui Chen. This has been the textbook for MA2611 and MA2612 (the other requirement for the course).
- **Python Data Science Handbook**, Jake VanderPlas

Evaluation/Grades

Final grades will be determined based upon the following breakdown:

Homeworks (3 assignments, 1-2 person teams)	15%
Midterm exam	20%
Final project (3-5 person teams)	35%
Final exam	30%

The midterm exam and final exam will be in class, cumulative, and open note, but **no collaboration will be allowed** and the exams be graded based upon demonstrated understanding of key concepts. For each exam, you are allowed to bring in up to two (2) 8 ½ by 11 sheets of paper (either printed or handwritten) with whatever notes you want for the exam. The homework problems will be performed in **groups of at most two** and will be graded for demonstrated understanding of key concepts and quality of presentation. You can choose your own teammate, but team changes will need to be approved by Prof. Paffenroth. The final project will be performed in **groups of 3-5** and will be graded based upon the quality and completeness of a final presentation and final report.

Schedule

As this is an experimental class, I reserve the right to change the order and content of lectures to improve the learning experience for the course. I will ensure that the homework's and exams match the material actually covered.

Homework 1	3/28/2025
Homework 2	4/4/2025
Midterm	4/10/2025
Project proposal	4/11/2025
Homework 3	4/18/2025
Final project due	5/2/2025
Final exam	5/7/2025

I reserve the right to curve the final grades (either up or down) based upon the aggregate performance of the class. Also, given the circumstances of this term, I reserve the right to change the courses activities, grading, etc. to achieve the courses learning goals.

Make-up Exam Policy

The exam dates are listed on the syllabus, and you are responsible for avoiding conflicts with the exams. If you need to take an exam early because of a conflict, then please let me know at least two weeks before the exam date. Any late exams will be penalized 30% (i.e., the maximum possible grade for the late exam will be 70%). **I will make every effort to be flexible, while respecting the need to be fair to all students.**

Late Assignment Policy

The assignment due dates are given in the syllabus and will be reiterated when each assignment is assigned. **One assignment** may be submitted up to one week late for full credit. However, any additional late assignments, or any assignment more than one week late, will be penalized 30% (i.e., the maximum possible grade for the late assignment will be 70%). To receive any credit for an assignment it must be submitted before the last class of the term. All assignments will be submitted using canvas.wpi.edu.

Collaboration and Academic Honesty Policy

Each student is expected to familiarize him/herself with WPI's Academic Honesty policies which can be found at https://www.wpi.edu/about/policies/academic-integrity/dishonesty. All acts of fabrication, plagiarism, cheating, and facilitation will be prosecuted according to the university's policy. If you are ever unsure as to whether your intended actions are considered academically honest or not, please contact your instructor in advance. Further information is available via https://www.wpi.edu/about/policies/academic-integrity.

Collaboration is prohibited on exams. Collaboration is encouraged on case studies, and you will be allowed to select your own teams of **2-5** for the case studies. On case studies you **may** discuss problems across teams, but each team is responsible for generating solutions and writing up results on their own **from scratch**. All violations of the collaboration policy will be handled in accordance with the WPI Academic Honesty Policy.

As examples, each of the following would be a violation of the collaboration policy (this list is **not** exhaustive):

- Two different case study teams share a solution to any assigned problem.
- One case study team allows another case study team to copy any part of a solution to an assigned problem.
- Any code or plots are shared between case study teams.

As examples, each of the following would not be a violation of the collaboration policy:

- Students within a team share solutions and code for a problem.
- Students from different teams discuss an assignment at the level of goals, where ideas for solutions can be found in the book or notes, what parts are more challenging, or how one might approach the problem.
- Of course, you can ask Prof. Paffenroth any questions you like, show them code, etc.

If there is any doubt as to what is allowed and what is not allowed, please just ask!

Plagiarism is defined as using the words, ideas, data, code, or other original academic material of another without providing proper citation or attribution. Plagiarism can apply to any assignment, including final or drafted copies. Examples include, but are not limited to:

- Misrepresenting the work of another as one's own,
- Inaccurately or inadequately citing sources,
- Paraphrasing (using the ideas of others in your own words) without citation.

Note that this includes the use of generative learning AI models such as ChatGPT, Bard, and other Large Language Models. If you use such tools, you will have to provide both prompt and answers received as an appendix (as well as check the claims independently!)

(*Partially based on the syllabi of Prof. Stephan Sturm*)

Inclusivity in the classroom

I consider this classroom to be a place where you will be treated with respect, and I welcome individuals of all ages, backgrounds, beliefs, ethnicities, genders, gender identities, gender expressions, national origins, religious affiliations, sexual orientation, ability—and other visible and nonvisible differences. All members of this class are expected to contribute to a respectful, welcoming, and inclusive environment for every other member of the class.

You deserve to be addressed in the manner you prefer. To guarantee that I address you properly, you are welcome to tell me your pronoun(s) and/or preferred name at any time, either in person or via email.
(*Partially based on materials from the Morgan Teaching Center and Prof. Bernardi*)

Accommodation for Special Needs or Disabilities

Students with approved academic accommodations should plan to submit their accommodation letters through the Office of Accessibility Services Student Portal. Should you have any questions about how accommodations can be implemented in this particular course, please contact me as soon as possible. Students who are not currently registered with the Office of Accessibility Services (OAS) but who would like to find out more information about requesting accommodations, documentation guidelines, and what the accommodated interactive process entails should plan to contact OAS either by email at AccessibilityServices@wpi.edu, by phone (508) 831-4908, or by stopping by the office on the 5th floor of Unity Hall.
(*Partially based on materials from the Morgan Teaching Center)*

Religious Observances

The observance of religious holidays (activities observed by a religious group of which a student is a member) and cultural practices are an important reflection of diversity. As your instructor, I am committed to providing equivalent educational opportunities to students of all belief systems. At the beginning of the semester, you should review the course requirements to identify foreseeable conflicts with case-study due dates and exams. If at all possible, please contact me within the first two weeks of the first class meeting to allow time for us to discuss and make fair and reasonable adjustments to the schedule and/or tasks.
(*Partially based on materials from the Morgan Teaching Center)*

Personal Emergencies

In the event of a medical or family emergency, please contact Prof. Paffenroth to find appropriate accommodation.