DS 501

STATISTICAL METHODS FOR DATA SCIENCE

This course provides an overview of Data Science, covering a broad selection of key challenges in and methodologies for working with big data. Topics to be covered include data collection, integration, management, modeling, analysis, visualization, prediction and informed decision making, as well as data security and data privacy. This introductory course is integrative across the core disciplines of Data Science, including databases, data warehousing, statistics, data mining, data visualization, high performance computing, cloud computing, and business intelligence. Professional skills, such as communication, presentation, and storytelling with data, will be fostered. Students will acquire a working knowledge of data science through hands-on projects and case studies in a variety of business, engineering, social sciences, or life sciences domains. Issues of ethics, leadership, and teamwork are highlighted. Prerequisites: None beyond meeting the Data Science admission criteria.

Where and When

Wednesdays from 6:00pm until 8:50pm - HL154

Instructor information

Prof. Randy Paffenroth
Office location: AK124
Office hours: 1-2pm on Tuesdays and 5-6pm on Wednesdays. **Other times are available by appointment, and walk-ins are always welcome if I am around and not otherwise indisposed.**
Best ways to contact me:

- WPI email: rcpaffenroth@wpi.edu
- Gmail and Google hangouts: randy.paffenroth@gmail.com
- Office phone: (508) 831-6562

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

TBD

High level course goals and learning objectives

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

- *Use tools* such as Linear Regression, Logistic Regression, Trees, etc. for making predictions from data.
- *Explain* the pros and cons of various approaches.
- *Avoid* common pitfalls such as overfitting and data snooping.
- Given a prediction generated from such a method, be able to *assess* the validity of the prediction.
- *Diagnose* what can go wrong with a prediction.

Recommended background for course

Basic knowledge of programming and statistics.

In particular, you will need to be able get your hands dirty playing with, processing, and plotting data using the **Python** computer language! The case studies will be done in **Python** and that will be the officially supported language for the course and all lecture examples will be in **Python**. Now, with that being said, 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

No official textbook but, a recommended texts will be discussed at the appropriate points in the class.

Evaluation/Grades

Final grades will be determined based upon the following breakdown:

Case studies and presentations (4 assignments)	50%
Midterm exam	20%
Final exam	30%

The midterm exam and final exam will be in class, non-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 4 8 ½ by 11 sheets of paper (either printed or handwritten) with whatever notes you want for the exam. The case studies will be performed in **groups of 2-4** and will be graded based upon the quality and completeness of presentations and the submitted reports.
I reserve the right to curve the final grades (either up or down) based upon the aggregate performance of the class.

Make-up Exam Policy

Make-up exams will only be allowed in the event of a documented emergency or religious observance. The exam dates are listed on the syllabus and you are responsible for avoiding conflicts with the exams.

Late Case Studies Policy

As the case studies are team based, and presentations will be expected on the day the case studies are due, **late case studies will not be accepted**!

Collaboration and Academic Honesty Policy

Collaboration is prohibited on the exams. Collaboration is encouraged on case studies and you will be allowed to select your own teams of 2-4 for the 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 sharing solutions and code for a problem.
- Students from different teams discussing 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 him code, etc.

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

Schedule

On this schedule the homework, exam, and final project dates are fixed. On the other hand, 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 homeworks and exams match the material actually covered.

	Tuesday	
Class 1	January 20	
Class 2	January 27	Case Study 1 assigned
Class 3	February 3	
Class 4	February 10	Case Study 1 due
Class 5	February 17	Case Study 2 assigned
Class 6	February 24	
Class 7	March 2	Case Study 2 due Midterm exam
	March 9	Term break
Class 8	March 16	
Class 9	March 23	Case Study 3 assigned
Class 10	March 30	
Class 11	April 6	Case Study 3 due
Class 12	April 13	Case Study 4 assigned
Class 13	April 20	
Class 14	April 27	Case Study 4 due Final exam

Accommodation for Special Needs or Disabilities

If you need course adaptations or accommodations because of a disability, or if you have medical information to share with me, please make an appointment with me as soon as possible. If you have not already done so, students with disabilities who believe that they may need accommodations in this class are encouraged to contact the Office of Disability Services as soon as possible to ensure that such accommodations are implemented in a timely fashion. This office is located in the West St. House (157 West St), (508) 831-4908.

Accommodation for Religious Observance

Students requiring accommodation for religious observance must make alternate arrangements with Prof. Paffenroth at least one week before the date in question.

Personal Emergencies

In the event of a medical or family emergency, please contact Prof. Paffenroth to work out appropriate accommodations.