95-885 Data Science and Big Data, Fall 2021

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Course Logistics

Class: Monday / Wednesday, 8:35 am – 9:55 pm (HBH 1206)

Instructor: Raja Sooriamurthi, PhD
Office: HBH 3017
Email: raja@cmu.edu

Teaching Assistants:
Sabrina Chua (sdchua) and Yunxiao (Kevin) Wang (yunxiaow)

Course Description: From Data to Value

This course is an introduction to the techniques and tools for analyzing and distilling actionable knowledge from data with the end goal of adding value.

The Age of Data: Currently we are in the midst of the next disruptive age of Information Technology. In 1945 electronic computers appeared ushering in what one could call the first disruptive age of *hardware*. Starting with the mainframes of the 60s to current cloud computing we have seen various hardware instances such as minicomputers, supercomputers, personal computers, handheld computers, and wearable computers. Paralleling advances in hardware, there have been many advances in *software*: programming paradigms (imperative, object-oriented, functional, concurrent), development methodologies (CMM, agile), and algorithms for solving a range of problems (e.g., systems, networking, graphics, AI, machine learning, analytics). Starting around the late 1960s to the explosion of the Web in the early 90s the third disruptive age was in *communication*—the ability for computer systems around the world to transmit, and share data. The combined advances in hardware, software, and communication forms the basis of our current disruptive age of *data*. Massive amounts of data (Tera* bytes and beyond) are available in a range of domains: science, commerce, finance, healthcare, social media, real-time sensors etc. At historically unprecedented levels we are able to collect, transmit, curate, and process huge amounts of data at enormous speeds resulting in our ability to do ongoing tasks better and to do tasks we couldn't do before.

Big Data: Since early 2000 the nature of data has morphed. Big Data is differentiated from traditional data in terms of the three 'V's: volume, velocity, and variety which raise interesting questions:

- *Volume*: When we process data at the Tera and Peta byte level what fundamental shift in our approach to solving problems occurs?
- *Velocity*: Given the fast transmission and computational speeds of current systems, what new capabilities are enabled by the processing of streams of data in real time?
- *Variety*: Estimates are that more than 90% of the world's data is not structured (i.e., not in classical relational databases amenable to SQL queries). What type of new actionable insights are facilitated by the processing of semi-structured (e.g., csv, JSON) and unstructured (e.g., text, images, audio) data?

* Various prefixes are used to denote data volumes: tera (10^{12}), peta (10^{15}), exa (10^{18}), zetta (10^{21}), yotta (10^{24}). It is estimated that we currently have around 12 zetta bytes of data in the world. To get an intuitive feel of these sizes consider the following analogy: if 1 byte is a grain of sand then a mega byte is a tablespoon of sand; a giga byte is a shoebox full of sand; a tera byte is a playground of sand; a peta byte is a mile long stretch of beach; an exa byte is a beach of sand from Maine to North Carolina; a zetta byte is a beach as big as all the coastlines in the world; a yotta byte is … (source EMC).

95-885 Data Science and Big Data, Fall 2021

Data Science and Machine Learning: Organizations and businesses need data driven actionable insights from data. For example, a casino may want to identify whether there is a certain group of customers from which more business occurs—a task known as customer segmentation. A cell phone company may want to know if there is a risk of customers leaving for another carrier—a business situation known as customer churn. Analytic tasks that facilitate such actionable insights include prediction, optimization, recommendation, classification, clustering etc.

This ongoing IT revolution driven by data is also viewed as the fourth paradigm of science. For more than 1000 years science has been driven by *empirical* methods. Starting a few hundred years ago a mathematics based *theoretical* science paradigm emerged. As human achievement progressed, it turned out that some phenomena cannot be approached empirically or they are not tractable to theoretical approaches (e.g., earthquakes, thermonuclear fission). Hence, few decades back, yet another paradigm of science fostered *computational* simulations to study these phenomena. Currently a new paradigm of doing science based on data has emerged—data science. In this course we will study the techniques and tools of these three intertwined themes (i) exploratory data analysis (ii) machine learning and (iii) big data.

Learning Objectives

Upon successful completion of this course, students will be able to:

1. Apply the principles of computational thinking (CT) to data science
2. Demonstrate development practices conforming to the Pythonic way
3. Express a business problem as a data problem
4. Perform exploratory data analysis from inception to the value proposition
5. Explain the core principles behind various analytics tasks such as classification, clustering, recommendation
6. Articulate the nature and potential of big data
7. Demonstrate the use of big data tools on real world case-studies

Assessment

Mastery of the course content will be demonstrated via a combination of quizzes, exercises, assignments, two projects and a take home exam:

Component	**Weight**
Quizzes + Exercises	15
Assignments	35
Project 1 (exploratory data analysis)	20
Project 2 (machine learning)	20
Take home case-study	10

This is an application-oriented course requiring skill in algorithmic problem solving. We will use Python based data science tools. Prior programming experience with Python at the level of the course 15-112 or 95-888 is required. As part of class preparation credit, periodically you will be required to setup infrastructure on your personal machine. We expect infrastructure concerns to be addressed before class so that we can focus on core

course content during class.

Tentative Course Schedule

Please check the detailed week by week course schedule for slides, handouts etc. All course content will be available from Canvas. The tentative set of themes we plan to discuss include:

I. Exploratory Data Analysis
   1. The Data Science pipeline
   2. Basic tools: Jupyter notebooks, Pandas
   3. Visualization techniques and tools
   4. Web scraping

II. Machine Learning
   1. Supervised, Unsupervised, Reinforcement approaches
   2. Analytics tasks: classification, regression, prediction, recommendation, clustering, optimization
   3. Tools: scikit-learn, Spark ML

III. Big Data
   1. The Hadoop platform (HDFS, map-reduce, MrJob)
   2. Cloud based platforms(e.g., Google Collaboratory, Azure, AWS)
   3. Spark (Spark SQL and Graph analytics)

The actual topics we discuss, and their depth will depend on the classes' interest and pace. Given the broad nature of the topics in the course, there is no single book that discusses the course content. The lectures will be self-contained for the needed background for the assignments. For additional background, all of the following books are excellent resources:

- *Python Data Science Handbook: Essential Tools for Working with Data* by Jake VanderPlas.
- *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd edition) by Wes McKinney.
- *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* by Aurélien Géron.
- *Spark: The Definitive Guide: Big Data Processing Made Simple* by Bill Chambers and Matei Zaharia.

All of these books are available online via our CMU library https://www.library.cmu.edu.

Class Policies

Attendance and Preparation for Class: Data Science is an exciting and rapidly evolving field. To fully engage in classroom discussions, you are expected to attend all class sessions and come prepared for each class. Class participation contributes towards the final grade assessment. There will be in-class assignments and occasionally unannounced short quizzes at the beginning of class. Students who have an unexcused absence or tardiness will not be able to make up these assignments and quizzes. Unexcused absences can reflect upon your grade. In the event of a situation requiring you to be absent (e.g., job interview) please contact the professor in advance.

Flex days: Part of professional behavior is submitting deliverables on time. Due dates of all deliverables (assignments, projects etc.) will be specified when issued and it is expected that assignments will be submitted on time. At the same time 'life happens' — you may have to travel for an interview, may fall sick, it may be an extremely busy week etc. To accommodate such situations, each student has a total of 4 flex days. Unless explicitly

specified otherwise, you may apply at the max 2 flex days (48 hours) for submitting an assignment beyond the due date. After that, submissions will be accepted with a 20% penalty per day late. Please email the professor ahead of time when you avail of a flex day.

Academic Integrity: Unless explicitly stated otherwise, *all work needs to be individually done*. While it is fine to discuss general ideas, all submitted work must be your own. Sharing of work with another student or using the work of another's when completing your own will result in a grade of zero. Any case of suspected cheating will be brought to the Dean's attention. If you referred to external sources or consulted with others be sure to clearly indicate so. Be sure to familiarize yourself with the University policies on academic integrity http://www.cmu.edu/policies/student-and-student-life/academic-integrity.html .

Reassessment: If you would like a component of the course (assignment, exam etc.) to be reevaluated, submit your request in writing (email will suffice) explaining in detail why you feel your response needs to be re-assessed. Any reassessment requests need to be submitted within two weeks of the assignment or exam bring returned.

For Students with Learning Disabilities: If you wish to request an accommodation due to a documented disability, please inform your instructor and contact the Office of Disability Resources http://www.cmu.edu/disability-resources

Take care of yourself

Do your best to maintain a healthy lifestyle by eating well, exercising, avoiding drugs and alcohol, getting enough sleep, and taking time to relax. Despite what you might hear, using your time to take care of yourself will actually help you achieve your academic goals more than spending too much time studying.

All of us benefit from support and guidance during times of struggle. There are many helpful resources available on campus. An important part of the college experience is learning how to ask for help. Take the time to learn about all that's available and take advantage of it. Ask for support sooner rather than later – this always helps.

If you or anyone you know experiences any academic stress, difficult life events, or difficult feelings like anxiety or depression, we strongly encourage you to seek support. Consider reaching out to a friend, faculty or family member you trust for assistance connecting to the support that can help. Counseling and Psychological Services (CaPS) is here for you: call 412-268-2922 and visit their website at http://www.cmu.edu/counseling. Over 25% of students reach out to CaPS some time during their time at CMU.
http://www.cmu.edu/teaching/designteach/design/syllabus/syllabussupport.html

Every individual must be treated with respect.

We are diverse in many ways, and this diversity is fundamental to building and maintaining an equitable and inclusive campus community. Diversity can refer to multiple ways that we identify ourselves, including but not limited to race, color, national origin, language, sex, disability, age, sexual orientation, gender identity, religion, creed, ancestry, belief, veteran status, or genetic information. Each of these diverse identities, along with many others not mentioned here, shape the perspectives our students, faculty, and staff bring to our campus. We, at CMU, will work to promote diversity, equity and inclusion not only because diversity fuels excellence and innovation, but because we want to pursue justice. We acknowledge our imperfections while we also fully commit to the work,

inside and outside of our classrooms, of building and sustaining a campus community that increasingly embraces these core values. Each of us is responsible for creating a safer, more inclusive environment.

Unfortunately, incidents of bias or discrimination do occur, whether intentional or unintentional. They contribute to creating an unwelcoming environment for individuals and groups at the university. Therefore, the university encourages anyone who experiences or observes unfair or hostile treatment on the basis of identity to speak out for justice and support, within the moment of the incident or after the incident has passed. Anyone can share these experiences using the following resources:

- Center for Student Diversity and Inclusion: csdi@andrew.cmu.edu, (412) 268-2150
- Report-It online anonymous reporting platform: www.reportit.net username: tartans password: plaid

All reports will be documented and deliberated to determine if there should be any following actions. Regardless of incident type, the university will use all shared experiences to transform our campus climate to be more equitable and just.

Let's have a fun and productive course!

Last updated: August 2021