67-364 / 95-885 Data Science, Spring 2017

Course Information

Class: 67-364: Tuesday / Thursday, 3:00 - 4:20 pm, HBH 1005
95-885: Friday, 9:00 - 11:50 am, HBH 1002

Instructor: Raja Sooriamurthi
Office: Porter Hall 224D
Email: raja@cmu.edu

Course Description

This course is an introduction to the techniques and tools for analyzing and distilling actionable knowledge from data.

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, 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 huge amounts 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 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 8 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 mega byte is a table spoon 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).

Data Science and Machine Learning: Organizations and businesses need data driven actionable insights. 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.

Learning Objectives

Upon successful completion of this course, students will have achieved the following learning objectives:

- Appreciate the value of data as a strategic resource for organizations
- Understand core analytics tasks e.g., exploratory data analysis, classification, prediction, optimization, recommendation etc.
- Hands on experience with data science tools and real-world case studies
- Exposure to the nature, potential, and tools for processing Big Data

Course Requirements

To complete the course, a combination of topic readings, exercises, assignments and two projects will be required.

Component	**Weight**
Attendance + Participation	5
Quizzes + Exercises	10
Assignments	40
Project 1	17
Project 2	18
Exam	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 is needed.

As part of class preparation credit, periodically you will be required to setup infrastructure on your personal machine (e.g., before we meet Friday morning of the first week you need to have the Miniconda Python analytics ecosystem setup). 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 this site. The set of themes we plan to discuss include:

I. Data Science
   1. The Data Science pipeline
   2. Exploratory Data Analysis (scraping and visualization)
   3. Machine learning (supervised and unsupervised approaches)
   4. Analytics tasks: classification, prediction, recommendation, clustering

II. Big Data
   5. The Hadoop platform (HDFS, map-reduce)
   6. Cloud based platforms such as AWS
   7. Data processing on Hadoop (MrJob and Pig)
   8. Spark

The actual topics we discuss and their depth will depend on the classes' interest and pace.

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.

Laptops and Mobile Phones: This is a technology-oriented course but there is a time and place to use technology. As the need arises, we will have hands on class sessions where you will need to use your laptop. But in other instances, laptops and other devices (iPads, smart phones etc.) tend to hinder classroom participation and discussions. Hence, unless explicitly stated otherwise, please close or turn off all such devices when in class. In this context, amongst other articles discussing this issue, you may find the Washington Post op-ed piece of David Cole, *Laptops & Learning,* to be interesting (http://www.washingtonpost.com/wp-dyn/content/article/2007/04/06/AR2007040601544.html).

Note taking: I recommend taking notes on paper. We will be discussing several concepts and ideas behind modern data science algorithms and techniques. You will find it more productive to think and engage during class. Slides and other material will be available after class from the course web site for your review.

Classroom Etiquette: Arriving late to class or leaving early disrupts the instructor and the learning environment. Please plan ahead and do everything you can to avoid these situations. Exiting and re-entering the classroom during session is a significant distraction interrupting class flow. If there is an emergency exit and re-enter with a minimum of disruption to the class.

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 not be accepted. 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: Disability Resources, 102 Whitfield Hall 412.268.2013, or by email at: lpowell@andrew.cmu.edu.

Take care of yourself

Do your best to maintain a healthy lifestyle this semester 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

Let's have a fun and productive course!

Last updated: Jan 16, 2017