95-852 Applied Data Science
Carnegie Mellon University, Spring 2016

Syllabus

Instructor
Dr. Artur Dubrawski
The Robotics Institute & Heinz College, CMU
Office: Newell-Simon Hall 3121, Phone: 412-268-6233, E-mail: awd@cs.cmu.edu
Office hours: Mondays, 8:30 – 10:00am (appointments recommended)

Lectures
Mondays & Wednesdays, 3:00 – 4:20pm, Hamburg Hall 1502

Recitations
Fridays 4:30 – 5:50pm, Hamburg Hall 1502

TAs
Jieshi Chen, Kyle Miller +TBA
Office, Phone, E-mail, Office hours: please refer to data on the course blackboard site.

Prerequisites
95-791 or permission of the instructor based on demonstrated equivalent experience.

Course Description

This course explores the rapidly developing field of Data Science in the context of its pragmatic applications. Modern enterprise is a complex system spanning a variety of functions in pursuit of a range of convoluted objectives. Its environment is exposed to the effects of globalization and “era of information”, producing the influx of large amounts of complex multisource data that may contain useful evidence. As a result, present-day decision makers face a truly formidable task of internalizing huge amounts of time-critical information while being expected to always make the right decisions at the right times. Conveniently, Data Science comes to their rescue.

Applied Data Science aims to achieve two main goals. The first is to optimize the efficiency of decision making by human managers. The second is to maximize the utilization of available data, so that no important clue is ever missed. This course aims at building expertise required to achieve those goals in practice. Students will have the opportunity to gain and solidify knowledge of the most important contemporary methods of Data Science, and to develop understanding of practical applicability of the studied topics in business scenarios. They will be able learn how to formulate analytic tasks in support of business objectives, how to define successful analytic projects, and how to evaluate utility of existing and potential applications of the discussed technologies in practice.

The instructor is a scientist and a practitioner. He has been involved in research towards various topics of machine intelligence and its applications for over two decades. He has been a technical lead and executive in the new technology industry. Currently Dr. Dubrawski is faculty at the CMU Robotics Institute where he directs the Auton Lab and leads multiple Applied Data Science projects in support of industry, government, and non-governmental organizations.

Required Name Tags

The students are expected to bring their letter-sheet-sized name tags to each of the meetings starting from the second lecture. Class attendance will be recorded at the beginning of each lecture based on the visual account of the displayed tags. Attendance in the first lecture will be taken by each participant signing the attendance sheet.
**Course Registration**

Pass-fail registration is not allowed in this course. Audit requests will be denied except for extraordinary circumstances.

**Grading**

Final grades will be determined by scoring the individual performance of each student at:
- Six in-class quizzes (each worth 4% of final grade points for the total of 24%),
- Active and constructive participation in classroom activities (6%),
- Written, in-class final examination (24%),

and by evaluating team activity:
- Blogs written weekly on specific assigned topics (16%),
- Data Science projects (30%).

The five-minute quizzes will be administered at the beginning of Wednesday meetings, starting from the second week of the course. Blogs authored by teams of 3 students will be due for review at midnight on each Friday starting from the second week of the course. Each team will be allowed to bid for their favorite blog topics selected from the extensive list of available topics. Class projects, also conducted in teams of 3 (these teams can but do not have to be the same as blog teams), will have specific schedule of deliverables including project proposals, milestone reports and progress presentations, as well as final reports and final presentation videos. Project progress presentations will be conducted during weekly recitations starting from the second week of the course. Their purpose is to showcase progress made by each team, enable constructive in-class discussion, and receive guidance regarding the next steps. All teams will need to be ready to present each week, however only an unannounced a priori subset of teams will be called to actually present during each particular recitation session. Class participation is considered to be a valuable criterion in grading performance. Empirically, the students who take an active part in lectures and those who frequent office hours tend to grasp the taught concepts in more effectively than those who resort to passive approach. The final grade for this course will reflect the sum of the above specified component grades.

**Academic Integrity and Classroom Habits**

The students are expected to strictly follow Carnegie Mellon University rules of academic integrity in this course. This means in particular that examinations and quizzes are to be the work of the individual student using only permitted material and without any cooperation of other students or third parties. It also means that usage of work by others is only permitted in the form of quotations and any such quotation must be distinctively marked to enable identification of the student’s own work and own ideas. All external sources used must be properly cited, including author name(s), publication title, year of publication, and a complete reference needed for retrieval. Regarding the group work (projects and blogs), the work should be the work of only the group members. In all their work students should not in any way rely on solutions to problems distributed in prior years or on the work of prior students or other current students. Violations will be penalized to the full extent mandated by the CMU policies. There will be no exceptions.

Usage of electronic equipment such as portable computers during classes is very strongly discouraged, except for meetings specifically designated for hands-on software demonstrations and exercises.

No student may record or tape any classroom activity without the express written consent of the instructor. If a student believes that he/she is disabled and needs to record or tape classroom activities, he/she should contact the CMU Office of Disability Resources to request an appropriate accommodation.
**Distribution of Lecture Notes**

Hard copies of the lecture notes will not be distributed. The notes will be available for download from the course blackboard site at least 12 hours before each lecture. The students are encouraged to bring their printed copies of the notes to class.

**Lecture Topics**

Note: Many of the topics listed below will take longer than one class meeting, and many of the class meetings will begin with a thorough in-class discussion of the due readings and cases. Preparation for these discussions and active, contributory participation in them will be the key factor used to determine the in-class activity grades.

- Introduction to Applied Data Science. Definitions and application environment. What challenges and needs could be addressed with Data Science? What can be achieved by applying Data Science in practice and what are the dimensions of attainable impact in business scenarios? How to estimate and validate added value?
- Reconciling prior knowledge against observations: Bayesian analytics.
- Building efficient models from complex data. Regularization. Graphical models. Predictive analytics with simultaneous use of multiple models.
- Revealing of the unexpected: Detection of anomalies and emerging patterns.
- Utility of link-entity data. Network analysis.
- Modeling customer preferences. Recommendation systems.
- What is the most profitable question to ask? Value of information vs. cost of acquiring more data. Active and semi-supervised learning. Selected aspects of causality and design of experiments.
- Representing data efficiently. How to make analysis possible when facing large amounts of data? How to enable almost-real-time ad-hoc querying of large data sets? Caching sufficient statistics. Parallel computing.

One or two class meetings will be devoted to lectures by invited speakers selected among seasoned practitioners of Data Science. Exact dates, scopes and topics of these lectures will be provided later.

**Readings**

The readings are intended to complement the material presented in class and to foster discussions of important topics related to Applied Data Science, and their real-world applications. Selected issues covered by the required readings will become topics of graded assignments. Some of the required readings will be provided electronically via course blackboard or as pointers to the resources available on the internet for free download. Note that many of the readings are protected under copyright law. Please note that it is illegal to distribute copies of the copyrighted materials without obtaining permissions from their legal owners. The list of readings will be provided separately.