**Syllabus**

**Instructor**

Dr. Artur Dubrawski  
awd@cs.cmu.edu, Newell-Simon Hall 3121  
Office hours: Mondays, 4:45pm-5:55pm (advance notice please).

**Head Teaching Assistants**

Benedikt Boecking  
boecking@andrew.cmu.edu, NSH 3122  
Office hours: TBA

Jessie Chen  
bieshic+@andrew.cmu.edu, NSH 3123  
Office hours: TBA

Dr. Mathieu Guillame-Bert  
mathieug@andrew.cmu.edu, NSH 3123  
Office hours: TBA

**Teaching Assistants**

Lujie (Karen) Chen  
lujiec@andrew.cmu.edu  
Office hours: TBA

Shuai Cao  
shuaic@andrew.cmu.edu  
Office hours: TBA

Mahesh Gajwani  
mgajwani@andrew.cmu.edu  
Office hours: TBA

Aleksandr Glumov  
aglumov@andrew.cmu.edu  
Office hours: TBA

**Meetings**

Mondays, 6:00pm–8:50pm, Hamburg Hall 1000.  
Fridays, 9:00am–10:20am, Hamburg Hall 1000.

**Prerequisites**

95-796 “Statistics for IT Managers” or instructor’s permission based on the student’s knowledge of fundamentals of probability and statistics. Previous experience with data analysis will be considered a plus, although it is not absolutely necessary.

**Introduction**

Data mining – intelligent analysis of information stored in data sets – has gained a substantial interest among practitioners in a variety of fields and industries. Nowadays, almost every organization collects data, which can be analyzed in order to support making better decisions, improving policies, discovering computer network intrusion patterns, designing new drugs, detecting credit fraud, making accurate medical diagnoses, predicting imminent occurrences of important events, monitoring and evaluation of reliability to preempt failures of complex systems, etc.

About the Instructor: Artur Dubrawski is a scientist and a practitioner. He has been researching machine intelligence and its applications for twenty years. In the past, he has been affiliated with an advanced data mining firm, Schenley Park Research, and served as Chief Technical Officer at Aethon, a local high-tech company making autonomous delivery robots. Currently Dr. Dubrawski is a faculty member at the CMU Robotics Institute, where he directs the Auton Lab: a data mining and machine learning research group.
Auton Lab’s work has yielded multiple deployments of analytic solutions and software in various government and industrial applications.

**Course Objectives**

This course will provide participants with an understanding of fundamental data mining methodologies and with the ability to formulate and solve problems with them. Particular attention will be paid to practical, efficient and statistically sound techniques, capable of providing not only the requested discoveries, but also estimates of their utility. The lectures will be complemented with hands-on experience with data mining software, primarily R, to allow development of basic execution skills.

The scope of the course will cover the following groups of topics.

- **Foundations. How to make data mining practical?** (approximately 40% of class time)
  
  ✓ Learning from data: why, what and how?
  ✓ Fundamental tasks, issues and paradigms of learning models from data.
  ✓ Real world data is noisy and uncertain. How much can we trust the results of our analyses?
  ✓ Model selection.
  ✓ Reduction of dimensionality and data engineering.
  ✓ Measures of association between data attributes: information theoretic, correlational.

- **Pragmatic methodologies for mining data** (approximately 60% of class time)
  
  ✓ Predictive analytics: classification and regression.
  ✓ Cost-sensitive model selection using ROC approach.
  ✓ Compression of data and models for improved reliability, understandability, and tractability of large sets of highly dimensional data.
  ✓ Association rule learning and decision list learning, decision trees.
  ✓ Introduction to density estimation, anomaly detection, and clustering.
  ✓ Overview of mining complex types of data.
  ✓ Illustrative examples of real-world applications.

**Registration**

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

**Academic Integrity and Classroom Habits**

Students are expected to strictly follow Carnegie Mellon University rules of academic integrity in this course. This means in particular that quizzes, examinations and homework 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 projects, 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 or telephones during lectures is very strongly discouraged, except for meetings specifically designated for hands-on software demonstrations, and the beginning of each lecture when a quiz is to be administered.
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.

**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. If the course blackboard site is not activated in time for the first class meeting, the notes will be emailed to all registered students. Email will also be used to deliver course material to those on the waiting list. Students are encouraged to bring their printed copies of the notes to class.

**Reading Material**

Unfortunately, the ideal textbook for this course does not exist. Instead, we will use a selection of readings excerpted from a variety of sources. These readings are intended to complement the material presented in class. Selected issues covered by the required readings will become topics of graded assignments and final examination.

Some of the required readings might be handed out in hard copies, but most of the required material will be distributed electronically, either via email or through the course blackboard site, or as pointers to the resources available on the internet for free download. Note that many of the readings are protected under copyright law. In order to use them in this course it was necessary to purchase official permissions from the copyright holders. Each enrolled student could have their HUB account charged with an equal share of the copyright fees. Although the exact amount of the individual share is not known at the moment of writing this document, it is estimated to not exceed $30.00. Please note that it is illegal to distribute copies of the copyrighted materials without obtaining permissions from their legal owners.

Interested students are welcome to go beyond the scope of the required readings. In particular, the following books are recommended - but not required - listed in no particular order:


**Software and Hands-on Exercises**

We will primarily rely on R free software to demonstrate and operationalize concepts presented during lectures. Students are expected to download and install the software, as well as learn basic usage skills on their own using tutorials available online. Appropriate resources will be recommended during the first lecture.

Recitations will review concepts taught in lectures and connect them to homework problems through examples. Recitation sessions when software tools are introduced (R and Tableau) will provide hands-on-experience opportunity: the students will be asked to follow the presenter using their laptops and they will work on assigned exercises while in session.

**Distribution of Assignments and Submission of Reports**

All assignments will be distributed electronically either through the course blackboard or via email. All reports (including homework) must be submitted electronically through the course blackboard. During the first week of the course, assignments will be emailed to students who are on the wait list. They will be required to email their reports in the electronic form to the TAs and the instructor by or before designated deadlines, unless they receive access to the course blackboard after being moved to the regular roster.
**Grading**

Grades will be based upon the results of four homework assignments, one analytical project, six in-classroom quizzes, and a written, in-class final examination.

Homework assignments are due at the beginning of the class when the homework is due. Late homework will be accepted until 24 hours past the deadline, but it will be subject to an automatic 50% grade reduction.

Open book quizzes will be administered at the beginning of every Monday class starting from the second week of classes. They will measure the level of comprehension of the key concepts covered in the lectures and obligatory readings.

The analytical projects will be conducted in small groups of students. Each team will analyze specific real-world data. The project will be graded based on preparedness to review meetings, two written progress reports and one final report, and a recording of an oral presentation of the results.

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 recitations and office hours tend to grasp the taught concepts in a more effective manner than those who prefer to assume a passive approach. The activity component of the grade will consider records of attendance in lectures and recitations, and the quality (not just the quantity) of the student’s contributions to classroom discussions.

The final grade for this course will be the result of a combination of elementary evaluations of a spectrum of skills specifically looked for by top employers. They include: the ability to solve problems individually given an abundance of resources (homework); the ability to perform individually under stress and time constraints with (quizzes) and without access to external sources of information (final exam); teamwork (group projects); and the capability of focused verbal interaction (active participation).

<table>
<thead>
<tr>
<th>Component</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homework (4 times 6%)</td>
<td>24%</td>
</tr>
<tr>
<td>Quizzes (6 times 2%)</td>
<td>12%</td>
</tr>
<tr>
<td>Analytical project (in teams)</td>
<td>30%</td>
</tr>
<tr>
<td>Active participation</td>
<td>6%</td>
</tr>
<tr>
<td>Final exam</td>
<td>28%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

**Timetable and Schedule of Lecture Topics (subject to possible revisions)**

*Note: Details of the recitation sessions will be provided separately*

**Monday, October 26th**

Lecture 1. Course introduction. Learning from data: Why, What and How?

Lecture 2. How to identify reliable models in Data Mining?

*Homework #1 handed out; reports due at 6pm on Monday November 02nd.*

**Friday, October 24th**

Lecture 3. Predictive analytics: Classification.

**Monday, November 02nd**

*Homework #1 due.*

Recitations 1 and 2.
Homework #2 handed out; reports due at 6pm on Monday November 09th.

Project teams formed and project assignments distributed electronically.

Friday, November 06th


Monday, November 09th

Homework #2 due.

Lecture 5. Preprocessing of data. Reduction of dimensionality.

Friday, November 13th


Monday, November 16th

Analytical project first milestone reports due.

Recitations 3 and 4.

Homework #3 handed out; reports due at 6pm on Monday November 23rd.

Friday, November 20th


Monday, November 23rd

Homework #3 due.

Lecture 7. Descriptive analytics: Density estimation, anomaly detection, and clustering.

Friday, November 27th

No class. Thanksgiving break.

Monday, November 30th

Analytical project second milestone reports due.


Homework #4 handed out; reports due at 6pm on Monday December 07th.

Friday, December 04th

Recitation 5.

Monday, December 07th
Homework #4 due.

Lecture 10. Overview of mining complex types of data.

Lecture 11. Final Review.

Friday, December 11th

Recitation 6.

Monday, December 14th (tentatively at the regular class time and place)

Project final reports and presentations due.

Final examination.