# 95-869 Big Data and Large Scale Computing

## Spring 2018

### CLASS MEETS:
- **Time:** Tue & Thu 9:00AM - 10:20AM  
- **Place:** HBH 1002

### PEOPLE:
- **Instructor:** Leman Akoglu, (lakoglu @ andrew)  
- **Office:** HBH 2118C, office ph 412-268-30 four three  
- **Office hours:** Thu 3-4 PM; also, by appointment
- **Teaching Assistant:** Abhinav Maurya, (amaurya @ andrew)  
- **Office:** HBH 3034  
- **Office hours:** TBD
- **Graders:**  
  - **Yuan Qin**  
    - **Email:** invert (andrew.cmu.edu @ yuanq)  
    - **Office hours:** by appointment
  - **TBD**  
    - **Email:** invert (andrew.cmu.edu @ TBD)  
    - **Office hours:** by appointment

### COURSE DESCRIPTION:

The rate and amount of data being generated in today's world by both humans and machines are unprecedented. Being able to store, manage, and analyze large-scale data has critical impact on business intelligence, scientific discovery, social and environmental challenges.

The goal of this course is to equip students with the understanding, knowledge, and practical skills to develop big data / machine learning solutions with the state-of-the-art tools, particularly those in the Spark environment, with a focus on programming models in MLlib, GraphX, and SparkSQL. See the syllabus for more details. Students will also gain hands-on experience with MapReduce and Apache Spark using real-world datasets.

This course is designed to give a graduate-level student a thorough grounding in the technologies and best practices used in big data machine learning. The course assumes that the students have the understanding of basic data analysis and machine learning concepts as well as basic knowledge of programming (preferably in Python or Java). Previous experience with Hadoop, Spark or distributed computing is NOT required.

### Learning Objectives

By the end of this class, students will
- gain understanding of the MapReduce paradigm and Hadoop ecosystem
- understand scalability challenges for common ML tasks
- study distributed machine learning algorithms
- understand details of SparkSQL, GraphX, and MLlib (Spark’s ML library)
- implement distributed pipelines in Apache Spark using real datasets

### RECOMMENDED TEXTBOOKS:

There is no official textbook for the course. I will post all the lecture notes and several readings on course website. Below you can find a list of recommended reading.

- **Scaling up Machine Learning: Parallel and Distributed Approaches**, Cambridge University Press  
  Ron Bekkerman, Mikhail Bilenko, John Langford
- **Hadoop in Practice**, Manning Publications Co.  
  Alex Holmes
- **Learning Spark**, O'Reilly  
  Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia
• Advanced Analytics with Spark, O'Reilly
  Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills

BULLETIN BOARD and other info
• We will use the Canvas for course materials, homework deposits, announcements, and grades.
• We will use Piazza for questions and discussions.
• Carnegie Mellon 2017-2018 Official academic calendar

MISC - FUN:
Joke-1  Joke-2  Joke-3
### Tentative Syllabus

<table>
<thead>
<tr>
<th>Date</th>
<th>Lectures and Readings</th>
<th>Out / Due</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review</td>
<td>Please take this <a href="http://www.andrew.cmu.edu/user/lakoglu/courses/95869/syllabus.htm">Python mini-quiz</a> before the course and take this <a href="http://www.andrew.cmu.edu/user/lakoglu/courses/95869/syllabus.htm">Python mini-course</a> if you need to learn Python or refresh your Python knowledge.</td>
<td></td>
</tr>
<tr>
<td><strong>Lecture 1: Introduction</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/20</td>
<td>* Big Data applications&lt;br&gt; * Technologies for handling big data&lt;br&gt; * Apache Hadoop and Spark overview</td>
<td></td>
</tr>
<tr>
<td><strong>Lecture 2: Hadoop Fundamentals</strong></td>
<td></td>
<td>HW0 out</td>
</tr>
<tr>
<td>3/22</td>
<td>* Hadoop architecture&lt;br&gt; * HDFS and the MapReduce paradigm&lt;br&gt; * Hadoop ecosystem: Mahout, Pig, Hive, HBase, Spark</td>
<td></td>
</tr>
<tr>
<td>3/27</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Lecture 3: Introduction to Apache Spark</strong></td>
<td></td>
<td>HW1 out</td>
</tr>
<tr>
<td>3/27</td>
<td>* Big data and hardware trends&lt;br&gt; * History of Apache Spark&lt;br&gt; * Spark's Resilient Distributed Datasets (RDDs)&lt;br&gt; * Transformations and actions</td>
<td></td>
</tr>
<tr>
<td>3/29</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Lecture 4: Machine Learning Overview</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4/3</td>
<td>* Basic machine learning concepts&lt;br&gt; * Steps of typical supervised learning pipelines&lt;br&gt; * Linear algebra review&lt;br&gt; * Computational complexity / Big O notation review</td>
<td></td>
</tr>
<tr>
<td><strong>Lecture 5: Linear Regression and Distributed ML Principles</strong></td>
<td></td>
<td>HW2 out</td>
</tr>
<tr>
<td>4/5</td>
<td>* Linear regression&lt;br&gt;   o formulation and closed-form solution&lt;br&gt;   o gradient descent&lt;br&gt;   o grid search&lt;br&gt; * Distributed machine learning principles&lt;br&gt;   o computation, storage, and communication</td>
<td></td>
</tr>
<tr>
<td>4/10</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Lecture 6: Logistic Regression and Click-through Rate Prediction</strong></td>
<td></td>
<td>HW3 out</td>
</tr>
<tr>
<td>4/12</td>
<td>* Online advertising&lt;br&gt; * Linear classification&lt;br&gt; * Logistic regression&lt;br&gt;   o working with probabilistic predictions</td>
<td></td>
</tr>
<tr>
<td>4/17</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
- categorical data and one-hot-encoding
- feature hashing for dimensionality reduction

Lecture 7: Principal Component Analysis and Neuroimaging
4/19
- Exploratory data analysis
- Principal Component Analysis (PCA)
- Formulations and solution
- Distributed PCA
4/24

Lecture 8: Big Data ML with MLlib
4/26
- k-means Clustering
- Decision Trees and Random Forests
- Recommenders

Lecture 9: Introduction to SparkSQL
5/1
- Working with tables in Spark
- Higher-level declarative programming

Lecture 10: Analyzing Networks with GraphX
5/3
- Understanding network structure
- Computing graph statistics

TBD
Final Exam
Assignments

COURSEWORK:
Coursework consist of 4 homework assignments, 1 take-home course project on big data analytics, and 1 final exam (grading in parentheses):

- **Homework** (55%)
- **Final Exam** (35%)
- **Class Participation** (10%)

IMPORTANT DATES:

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Note</th>
<th>Out</th>
<th>Due</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homework 0</td>
<td>Installation, Set up</td>
<td>3/27</td>
<td>--</td>
<td>0%</td>
</tr>
<tr>
<td>Homework 1</td>
<td>pySpark and RDDs</td>
<td>3/29</td>
<td>4/5</td>
<td>7%</td>
</tr>
<tr>
<td>Homework 2</td>
<td>Regression in Spark</td>
<td>4/5</td>
<td>4/12</td>
<td>12%</td>
</tr>
<tr>
<td>Homework 3</td>
<td>Classification in Spark</td>
<td>4/12</td>
<td>4/19</td>
<td>12%</td>
</tr>
<tr>
<td>Homework 4</td>
<td>Data Analysis with PCA in Spark</td>
<td>4/19</td>
<td>4/26</td>
<td>12%</td>
</tr>
<tr>
<td>Homework 5</td>
<td>Hands-on with ML-lib and SparkSQL</td>
<td>4/26</td>
<td>5/3</td>
<td>12%</td>
</tr>
<tr>
<td>Final Exam</td>
<td>You are allowed to bring a single, A4-size ‘cheat’ sheet with your notes</td>
<td>TBD</td>
<td>TBD</td>
<td>35%</td>
</tr>
<tr>
<td>Class participation</td>
<td>Pop-quizzes (any time during class)</td>
<td>--</td>
<td>--</td>
<td>10%</td>
</tr>
</tbody>
</table>

HOMEWORK:
The goal of the homework is to enable the students to practice the concepts learned in class using real-world datasets.

- ASSIGNMENTS ARE DUE AT THE BEGINNING OF LECTURE ON THE DUE DATE.
- All assignments are to be done individually. Please see the collaboration policy.
- To submit:
  - Submit your soft-copy (all source files) on Canvas.
  - Return your hard-copy (print out) submission in class.
  - Make sure that your answers are legible and coding is clear.
  - See course policies for assignment questions, late submissions, graded homework pick-up.

EXAM:
There will be a final exam. It will be closed everything -- no books, slides, computers, etc. You will only be allowed to bring along with you 1 A4-size paper of your own notes ('cheat sheet'), you can use both sides. The final date will be announced during the semester.

CLASS PARTICIPATION:
Attendance will be quantified via pop-up quizzes in class. Each quiz will be a short-answer question, to be answered given 2-3 minutes. The quiz can be any time during the class, at the beginning, at the end, or anywhere in between.
There will be 10 such quizzes, 1 point each, for a total of 10% of the final grade. Students who attend class will get 0.5 point even if they return a blank answer. Partial answers will earn 0.75, and correct answers will get 1 point. Students who miss class and hence the quiz will receive 0 point.
Course Policies

LECTURES

- All devices such as laptops, cell phones, noisy PDAs, etc. should be turned off for the duration of the lectures and the recitations, because they may distract other fellow students.
- Please come to all lectures on time and leave on time, again so that there are no distractions to the classmates.

PREREQUISITES

Students are expected to have the following background:

- Basic knowledge of data analysis and machine learning concepts; having taken:
  - 95-791 Data Mining, OR
  - 95-828 Machine Learning for Problem Solving
- Basic programming skills at a level sufficient to write a reasonably non-trivial computer program in a language of preference (preferably Python)

ASSIGNMENTS

- Assignments are due at the * beginning of lecture * on the due date.
- The due date of assignments are posted at the assignments page.
- Assignments will be posted on Canvas.
- Students should submit the assignments electronically via Canvas, and return hard copy print outs in class.

Important Note: As we reuse problem set questions, covered by papers and webpages, we expect the students not to copy, refer to, or look at the solutions in preparing their answers. Since this is a graduate-level class, we expect students to want to learn and not google for answers. The purpose of problem sets in this class is to help you think about the material, not just give us the right answers. Therefore, please restrict attention to the books mentioned on the front page when solving problems on the problem set. If you do happen to use other material, it must be acknowledged clearly with a citation on the submitted solution.

Academic integrity

All students are expected to comply with CMU's policy on academic integrity. Please read the policy and make sure you have a complete understanding of it.

Collaboration

You are encouraged to discuss homework problems with your fellow students. However, the work you submit must be your own. You must acknowledge in your submission any help received on your assignments. That is, you must include a comment in your homework submission that clearly states the name of the student, book, or online reference from which you received assistance.

Submissions that fail to properly acknowledge any help from other students or non-class sources will receive NO credit. Copied work will receive NO credit. Any and all violations will be reported to the Heinz College administration and may appear in the student's transcript.

Questions and requests

- You should use Piazza for all your questions about the assignments and the course material. Instructor and TA(s) will do their best to answer your questions timely.
- Regrade requests should be done in writing/email,
  - within 2 days after graded assignments are distributed
  - to the TA that graded the question, and specifying
    - the question under dispute (e.g., 'HW1-Q.2.b')
    - the extra points requested (e.g., '2 points out of 5')
    - and the justification (e.g., 'I forgot to divide by variance, but the rest of my answer was correct')
  - In the remote case there is no satisfactory resolution, please contact the instructor.
Homework pick-up information

- You may pick up graded homeworks etc., from the course admin
  - Mrs. Adrienne McCorkle, HBH 2250
  - 9:00-11:30am, 1:30-4:30pm every weekday
  - with photo-id (for your privacy protection)

Late policy

- No delay penalties, for medical/family/etc. emergencies (bring written documentation, like doctor’s note).
- Each student is granted ‘3 slip days’ total for the whole course duration, to accommodate for coinciding deadlines/interviews/etc. That is, no questions asked, if the total delay is 3 days or less.
  - You can use the extension on any assignment during the course. For instance, you can hand in one assignment 3 days late, or 3 different assignments 1 day late each.
  - Late days are rounded up to the nearest integer. For example, a submission that is 4 hours late will count as 1 day late.
  - After you have used up your slip days, any assignment handed in late will be marked off 25% per day of delay.
- To use slip days:
  - upload your homework on Canvas to mark the time of your submission.
  - No emails to TA are necessary -- we will use the latest upload time as the submission time.
  - Make sure to return your printed hard copy next time you are in class.
  - Note down on front page of hard copy submission: count of slip days you used, as well as the count of slip days left.