Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

95-869 Big Data and Large Scale Computing
Spring 2021

HOME
SYLLABUS
ASSIGNMENTS
COURSE POLICY

Announcements

- This course starts on Tuesday February 2, 2021.
- Please see the course prerequisites here.

CLASS MEETS:

Time: TUE/THU 4:50PM - 6:10PM
Place: ONLINE @Zoom (Calendar invitations with link sent individually)

PEOPLE:

Instructor: Leman Akoglu
- Email: *invert* (andrew.cmu.edu @ lakoglu)
- Online office hour: **THU @9PM-10PM EDT**; also, by appointment

Teaching Assistants:
Sameera Kodi
- Email: *invert* (andrew.cmu.edu @ skodi)
- Online office hour: **TUE @9-10AM EDT**

Lingxiao Zhao
- Email: *invert* (andrew.cmu.edu @ lingxia1)
- Online office hour: **THU @8-9PM EDT**

Please find the Zoom links to office hours on Canvas.

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
- Learning Spark, O'Reilly
  Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia
- Advanced Analytics with Spark, O'Reilly

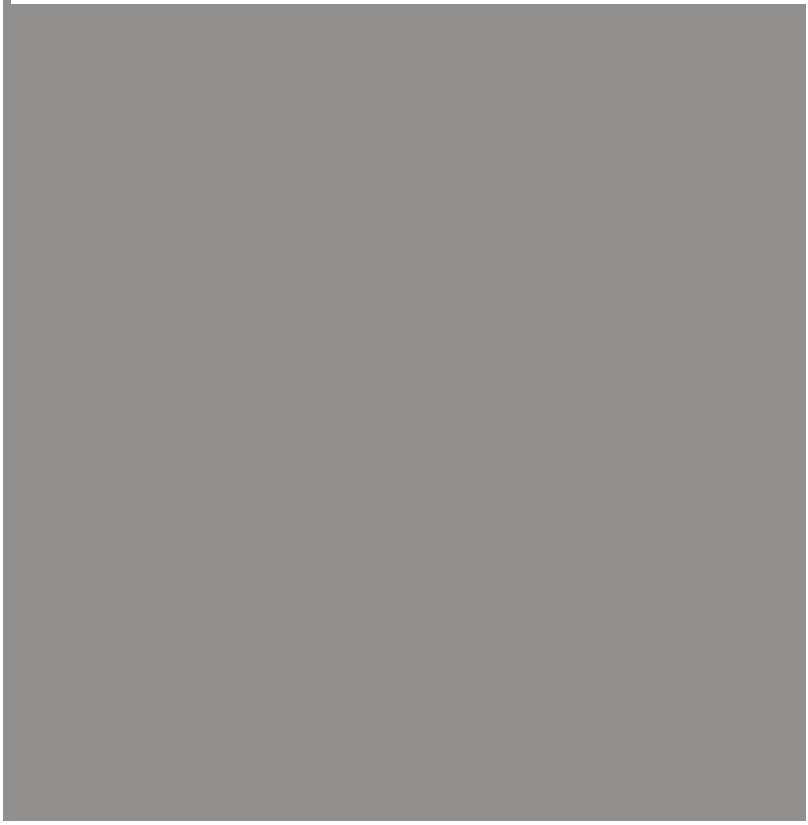
Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills

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 2020-2021 official academic calendar

MISC - FUN:

Joke-1 Joke-2 Joke-3



Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

95-869 Big Data and Large Scale Computing

Tentative Syllabus

Week	Lectures and Readings	Out / Due
	Review Please take this Python mini-quiz before the course and take this Python mini-course if you need to learn Python or refresh your Python knowledge.	
Week 1	**Lecture 1: Introduction** • Big Data applications • Technologies for handling big data • Apache Hadoop and Spark overview	
Week 2	**Lecture 2: Hadoop Fundamentals** • Hadoop architecture • HDFS and the MapReduce paradigm • Hadoop ecosystem: Mahout, Pig, Hive, HBase, Spark	**HW0 out**
	Lecture 3: Introduction to Apache Spark • Big data and hardware trends • History of Apache Spark • Spark's Resilient Distributed Datasets (RDDs) • Transformations and actions	**HW1 out**
Week 3	**Lecture 4: Machine Learning Overview** • Basic machine learning concepts • Steps of typical supervised learning pipelines • Linear algebra review • Computational complexity / Big O notation review	

Week 4	**Lecture 5: Linear Regression and Distributed ML Principles**
	• Linear regression
	  ◦ formulation and closed-form solution
	  ◦ gradient descent
	  ◦ grid search
	• Distributed machine learning principles
	  ◦ computation, storage, and communication
		HW1 due
		HW2 out

Week 5	**Lecture 6: Logistic Regression and Click-through Rate Prediction**
	• Online advertising
	• Linear classification
	• Logistic regression
	  ◦ working with probabilistic predictions
	  ◦ categorical data and one-hot-encoding
	  ◦ feature hashing for dimensionality reduction
		HW2 due
		HW3 out

Week 6	**Lecture 7: Principal Component Analysis and Neuroimaging** • Exploratory data analysis • Principal Component Analysis (PCA) • Formulations and solution • Distributed PCA	**HW3 due** **HW4 out**
Week 7	**Lecture 8: Big Data ML with MLlib** • k-means Clustering • Decision Trees and Random Forests • Recommenders	**HW4 due** **HW5 out**
	Lecture 9: Introduction to SparkSQL • Working with tables in Spark • Higher-level declarative programming	
Bonus Lecture	**Lecture 10: Analyzing Networks with GraphX** • Understanding network structure • Computing graph statistics	**HW5 due**
See here	**Final Exam**	

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

95-869 Big Data and Large Scale Computing

HOME
SYLLABUS
ASSIGNMENTS
COURSE POLICY

Assignments

COURSEWORK:

Coursework consist of Coursework consist of 5 homework assignments, 1 final exam, and after-class quizzes that will determine your class participation (grading in parentheses):

- Homework (60%)
- Progress (15%)
- Final Exam (25%)

IMPORTANT DATES:

Assignment	Note	Out	Due	Weight
Homework 0	Installation, Set up	2/2	--	0%
Homework 1	pySpark and RDDs	2/11	2/18	11%
Homework 2	Regression in Spark	2/18	2/25	12%
Homework 3	Classification in Spark	2/25	3/4	12%
Homework 4	Data Analysis with PCA in Spark	3/4	3/11	12%
Homework 5	Hands-on with ML-lib and SparkSQL	3/11	3/18	13%
Progress	Quizzes (posted on Canvas)	after each class	wihtin 2 days	15%
Final Exam	on Canvas	TBD	--	25%

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 **11:59PM EDT** OF THE DUE DATE.
- All assignments are to be done **in groups**. Please see the collaboration policy.
- Submission (only electronically):
  - Submit all of your source files on Canvas.
  - Also submit your print out/pdf with answers on Canvas.
  - Make sure that your answers are legible and coding is clear.
  - See course policies regarding questions about the assignments, late submissions, etc.

EXAM:

There will be a final exam. It will be posted on Canvas. Decision regarding students taking the exam synchronously or asynchrounously will be made later in the semester.

PROGRESS:

Progress and participation will be quantified via quizzes. Each quiz will be a list of multiple-choice questions, to be posted on Canvas after each class -- with a due date and time.
There will be a total of 10-12 such quizzes during the semester, 1 point each. We will select the highest-scoring 10 out of those for each student, for a total of 15% of the final grade. Students who attempt all the questions in the quiz will get 0.5 point even if they answer all these questions wrong. Students who do not attempt to answer any questions in a quiz will receive 0 points.

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

95-869 Big Data and Large Scale Computing

HOME
SYLLABUS
ASSIGNMENTS
COURSE POLICY

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:
  - (Heinz) 95-791 Data Mining, or
  - (Heinz) 95-828 Machine Learning for Problem Solving
  - (SCS) or 10601 or 10701 or 10715 or 15388 or 11663 or 16791
- Working knowledge of linear algebra (e.g. matrix-matrix multiplication, eigenvectors, matrix rank, etc.)
- Programming skills at a level sufficient to write a reasonably non-trivial computer program in **Python**

ASSIGNMENTS

- Assignments are due at *** 11:59PM EDT *** 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 only** via Canvas.

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 expected to work on each HW with the students in your **'study group'**. We will announce the study groups beforehand, which we will create considering the time zones of the students as well as paying attention to diversity. Each group will upload a **single** submission on Canvas.

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 any book or online reference from which you received assistance. Submissions that fail to properly acknowledge any help from 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.

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.