# CS 5683: Algorithms and Methods for Big Data Analytics
# Fall 2020

## Course Information:

Instructor: Arunkumar Bagavathi
Email: abagava@okstate.edu
Office: MSCS 215
Fall 2020 Office hours: Tuesdays 9:00am-12:00pm – *Online meetings only*

Lectures: Monday, Wednesday 2:30pm – 3:45am
Classroom: Engineering South 201B
Credits: 3

Teaching Assistant: Reza Marzban
TA email: reza.marzban@okstate.edu

Prerequisites:

- *CS 5513 Numerical Computation (with 'C' or better grade) OR permission of instructor* (mandatory)
- Fundamentals of algorithms and data structures
- Fundamentals of Python programming, and Jupyter notebook
- Familiarity with basic probability theory and linear algebra
- Familiarity with rigorous algorithm analysis and writing proofs

Other information: The class will follow Canvas. All course announcements, assignments, instructions, grades, and course materials will be posted in Canvas

**Course Description:** This course will discuss several data mining and machine learning algorithms for big data analytics: clustering, recommendation systems, social media analysis, big graph mining, data stream analysis, large-scale machine learning, and online algorithms. The course includes hands-on experience for big data frameworks and algorithms with tutorials and assignments. All assignments will be based on Apache Spark and Python programming. But prior knowledge on Apache Spark is *not required*.

May not be used for degree credit with **MSIS 5683**.

## Topics

Following is the tentative syllabus for the course. It is intended to change based on availability of time.

1. Introduction to Spark
2. Frequent Itemset Mining: Apriori algorithm
3. Similar items: Shingling, Locality-Sensitive Hashing
4. Clustering
    a. Similarity measures
    b. Hierarchical clustering
    c. K-means clustering
    d. BFR algorithm
    e. CURE algorithm

5. Recommender systems
    a. Content-based model
    b. Collaborative filtering
    c. Dimensionality reduction
        i. Eigen values and eigenvectors
        ii. SVD
        iii. CUR decomposition
6. Text mining
    a. POS tagging
    b. NER models
    c. Simple topic modelling
7. Graph Mining
    a. Introduction to (social) networks
    b. PageRank
    c. Link spam
    d. Community detection
        i. Personalized PageRank
        ii. Motif-based clustering
        iii. Louvain algorithm
8. Data Stream Mining
    a. Sampling methods
    b. Querying – standard and ad-hoc querying
    c. Filtering methods – Bloom filters
    d. Counting distinct elements – Flajolet-Martin algorithm
9. Large-scale Machine Learning
    a. Parallel ML
    b. Decision Trees
    c. SVM
    d. Neural networks
        i. CBOW and skip-gram models
        ii. Intro to text and graph representation learning
10. Online Algorithms (OR) Intro. to Visual Analytics

## Grading:

There are only 3 activities in the course throughout the semester: Assignments, Projects, and Final exam. Ph.D. students however have to make an extra online paper presentation, related to their research, to the class. Assignments are more like tutorials to understand the basics of algorithms and frameworks, and Projects will involve both Python or PySpark programming along with problem solving tasks. All submissions will be on Canvas. The final exam will be on the Finals week of the semester. The exam will be conducted online. The exam questions format and syllabus will be decided during the pre-finals week. Class participation (both in-class and Canvas discussion boards) will be utilized to assign extra grades.

## M.S. students grading:

Assignments: 25%
Projects: 50%
Final exam: 25%

## Ph.D. students grading:

Assignments: 25%
Projects: 40%
Paper presentation: 10%
Final exam: 25%

## Assignments Policy:

**Assignments and Projects:** The course has two activities: 1) *Assignments:* These are short (more like tutorial) activities and it does not require much time to complete, and 2) *Projects:* These are much broader activities and it takes substantial amount of your time to complete it. Both activities involve both programming and problems.

**Code of conduct:** We follow OSU students code of conduct *extremely seriously*. Standard penalty for students who jeopardize the code of conduct is suspension from the university. We recommend students to form online groups and discuss about assignments and projects. However, each student should write their answers independently along with list of people they discussed with in the submission.

**Questions:** Post your questions on discussion board or email instructor/TA. We will try hard to reply as soon as possible. You can expect our reply anytime within 12 hours of your question. *Please do not post questions in last few hours of the deadline.*

**Late submissions:** Students have a total of **2 grace periods** for both assignments and projects. However, we will not accept submissions after 1 week of the due date.

**Regrading:** We try our best to make fair and consistent grading for all activities. So the grades you receive are very unlikely to change significantly. However, if you want to regrade your work, email us your requests within a week after receiving your grades.

## Books:

Books are not mandatory for this course. Following are good books to read for this course:

1. "Mining Massive Datasets", Second Edition by Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman
2. "Spark: The Definitive Guide", 2018 Edition by Bill Chambers and Matei Zaharia
3. "Introduction to Data Mining", Second Edition by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar

**Student Guidance on Wearing Facial Coverings**

**All OSU students, employees, and visitors must wear a facial covering (mask) upon entering any campus building and when near or encountering others. This includes during class and in laboratory settings. Please be aware that additional personal protective equipment (PPE), such as a face shield along with a mask, may be required in certain classroom and laboratory settings. Students who fail to wear their facial covering will be asked to leave the room and return after retrieving their facial covering.**

**Students who continuously fail to comply with this university expectation will be referred to the Office of Student Conduct Education and Administration for the Student Code of Conduct's Failure to Comply policy.**

**COVID-19 can be spread when people are asymptomatic, which means they do not know they are sick yet. Wearing facial coverings has been shown to reduce the spread of COVID-19 to others. It is important that OSU is a safe place to work and study, and taking this step creates a safe environment for all of us as advised by the CDC. More on facial covering guidelines. (https://go.okstate.edu/coronavirus/campus-reopening-plan/plan-at-a-glance/face-coverings.html)**

**COVID-19 UPDATES**
**https://go.okstate.edu/coronavirus/**

Please visit this webpage for information regarding the university's response to the COVID-19 pandemic, answers to frequently asked questions, and other important updates.

**ACADEMIC INTEGRITY**
101 Whitehurst/405-744-5627/http://academicintegrity.okstate.edu
OSU is committed to maintaining the highest standards of integrity and ethical conduct. This level of ethical behavior and integrity will be maintained in this course. Participating in a

behavior that violates academic integrity (e.g., unauthorized collaboration, plagiarism, multiple submissions, cheating on examinations, fabricating information, helping another person cheat, unauthorized advance access to examinations, altering or destroying the work of others, and altering academic records) will result in an official academic sanction. Violations may subject you to disciplinary action including the following: receiving a failing grade on an assignment, examination or course, receiving a notation of a violation of academic integrity on your transcript, and being suspended from the University. Students have the right to appeal the charge.

## COPYRIGHT & FAIR USE POLICY OF COURSE MATERIALS

Course materials may not be published, leased, sold to others, or used for any purpose other than appropriate OSU-related individual or group study without the written permission of the faculty member in charge of the course and other copyright holders. This paragraph grants you a limited license giving you access to materials for this course, including PowerPoint slides, audio/video recordings, written, or other materials, for appropriate OSU-related educational use only. Lectures should not be recorded without permission from the faculty member and must not be further disseminated or shared. Assignments, quizzes, and exams (individual questions or in its entirety) should not be uploaded to websites offering note-sharing, tutoring, or other academic help (free or by paid subscription).

## CLASS PARTICIPATION

Class participation is a critical component of learning; therefore, you are expected to participate fully in all scheduled class meetings. While no penalty may be assessed for class absences during Fall 2020, you may not be permitted to make up certain class activities if absent. If you are ill, you should stay home. If you are required to participate in official university-sponsored activities or military training, you should receive an excused absence unless the written course attendance policy indicates otherwise. If you will be absent from class for sponsored activities, you must provide prior notification of the planned absence to the instructor. You may be required to submit assignments or take examinations before the planned absence.

## COURSE SCHEDULE ADJUSTMENTS FOR FALL

Courses that usually meet for 50 minute class periods (e.g. Monday, Wednesday, and Friday classes) have been shortened by 5 minutes per class meeting to allow more travel time for students and faculty between classes. Instructors will be expected to provide additional assignments or instruction to make up the contact time for these courses to ensure classes meet the required semester credit hour standards.

## INSTRUCTOR OFFICE HOURS

During Fall 2020 instructors and teaching assistants shall offer office hours online. Students are asked to respect the posted virtual office hours of all instructors and teachings assistants.

## PRE-FINALS WEEK POLICY

Final examinations are scheduled at the end of each semester and are preceded by pre-finals week, which begins seven days prior to the first day of finals. During Fall 2020 pre-finals week, all normal class activities will continue in an online format; however, no assignment, test, or examination accounting for more than 5% of the course grade may be given; and no activity or field trip may be scheduled that conflicts with another class. This excludes makeup and laboratory examinations, out-of-class assignments (or projects) made prior to pre-finals week and independent study courses. No student or campus organization may hold meetings, banquets,

receptions, or may sponsor or participate in any activity, program, or related function that requires student participation. For additional information, contact the Office of Academic Affairs, 405-744-5627, 101 Whitehurst.

## FINAL EXAM OVERLOAD POLICY

Final exams for Fall 2020 will be held in an online format. In the event you have three or more final exams scheduled for a single day, you are entitled to arrange with the instructor of the highest numbered course (4 digit course number) or two highest, if you have four finals on one day, to re-schedule that examination(s) at a time and place of mutual convenience during final exam week. You should submit this request in writing, with a copy of your class schedule, at least two weeks prior to the beginning of final exam week. The instructor has one week prior to the beginning of final exam week to arrange a mutually convenient time and place for administration of the final exam. After one week, if an agreement cannot be reached, take the request to the department head.

## EQUAL OPPORTUNITY

409 General Academic Building/405-744-7607
https://1is2many.okstate.edu/
OSU is committed to maintaining a learning environment that is free from discriminatory conduct based on race, color, religion, sex, sexual orientation, gender identity, national origin, disability, age or protected veteran status. OSU does not discriminate on the basis of sex in its educational programs and activities. Examples of sexual misconduct and/or sex discrimination include: sexual violence, sexual harassment, sexual assault, domestic and intimate partner violence, stalking, or gender-based discrimination. OSU encourages any student who thinks that he or she may have been a victim of sexual misconduct or sexual discrimination to immediately report the incident to the Title IX Coordinator (405-744-9153 or Deputy Title IX Coordinator (405-744-5470). If a reporting student would like to keep the details confidential, the student may speak with staff in the Student Counseling Center (405-744-5472) or one of the University's Sexual Assault Victim Advocates (Mon-Fri 8 AM-5 PM, 405-564-2129 or 24 Hour Help Line 405-624-3020).

## STUDENT ACCESSIBILITY SERVICES

1202 W. Farm Rd #155/405-744-7116/http://sds.okstate.edu/

According to the Americans with Disabilities Act, each student with a disability is responsible for notifying the University of the disability and requesting accommodations. If you think you have a qualifying disability and need accommodations, contact the Office of Student Accessibility Services to start the registration process and to ensure timely implementation of appropriate accommodations. To receive services, you must submit appropriate documentation and complete an intake process to verify the existence of a qualified disability and identify reasonable accommodations. Faculty have an obligation to respond when they receive official notice of accommodations but are under no obligation to provide retroactive accommodations.

**View more university syllabus details at:**
https://academicaffairs.okstate.edu/sites/default/files/Fall%202020%20Syllabus%20Attachment%20as%20of%20Aug%206%202020.pdf