INF 626: *Big Data and Stream Analytics* (3 cr)
Fall 2018

I tell my students, 'When you get these jobs that you have been so brilliantly trained for, just remember that your real job is that if you are free, you need to free somebody else. If you have some power, then your job is to empower somebody else. This is not just a grab bag candy game.' – Toni Morrison

Course Instructor

Instructor: George Berg
Email: gberg@albany.edu

- Office Hours:
  Tuesdays and Thursdays: 2:50 – 3:50 in the Campus Center. Ground floor near the rear grand staircase.
- Wednesdays: 2:50 – 3:50 in Draper 105.

Other Contact Info:

- Office: UAB 413
- Phone: 1-518-437-4937
- Twitter: @GBerg_UAlbany
- FB: @GeorgeBergUAlbanyCS

Course Description

INF 626 *Big Data and Stream Analytics* (3)
In data science, the analysis of large amounts of data is frequently expressed as the 4 V's: volume, velocity, variety, and veracity. This course examines the underlying concepts and practical implications of each of these dimensions at the frontier of data analytics. The size and amount of time available to process data both affect the types of analysis that are possible, as does the variety of data. In addition, issues of data source, distribution, and how much it can be trusted as the basis for analysis are increasingly important.
Prerequisite(s): INF 624.

Expected Student Outcomes

This is a comprehensive graduate course in the analysis of big data and streaming data. Specifically big data refers to amounts of data that preclude analysis by normal software methods. Streaming data introduces time challenges as well. The volume and pace of data introduce their own challenges in analyzing the data, especially in time critical situations.
By the end of this course, students will

- Examine data with statistical, machine learning, and data mining concepts to discern meaningful patterns, and to create predictive models.
- Connect how those techniques are affected by the size and pace of the incoming data.

- Use various computer packages to implement the above concepts and to analyze data.
- Recognize the challenges of variety in type, distribution and other relevant properties of data to analyze.
- Be aware of problems with the source and provenance of data. This can range from statistical properties of data used through potentially malevolent attempts to affect analyses.

Class Meetings

Lecture

The lecture meets twice week: Tuesdays and Thursdays, 1:15 – 2:35 PM in Husted 225.

Required Texts

1. Russell Jurney, *Agile Data Science*, O'Reilly, 2017. ISBN-13 978-0133892026.
2. Sandy Ryza, Uri Laserson, Sean Owen & Josh Wills, *Advanced Analytics with Spark*, O'Reilly. 2015. ISBN-13

Recommended Text

There is no recommended text for this class.

Additional Readings

There will be readings that will be available to the students online or via Blackboard. When these readings are assigned, the class will be told where they can be found.

TEAM-BASED LEARNING (TBL)

This course uses Team-based Learning (TBL). This section describes how we will be using TBL in this class.

AN ABSOLUTELY CRUCIAL POINT: The course is divided into learning modules. You ***must*** do the readings for each module ***before*** the unit's start. This is because each unit starts with a Readiness Assessment Test (RAT). Readings must be done before the RAT tests for the module (dates given in the syllabus below). The RAT tests are based solely upon the readings, and not on lecture or other in-class preparation beforehand.

Teams

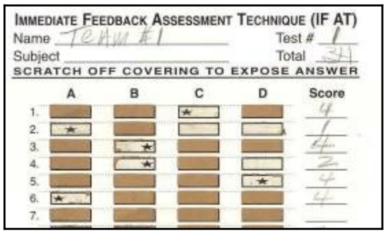
This course will be using a Team-Based-Learning (TBL) format (http://www.teambasedlearning.org). This instructional method aims to help develop your learning skills and will be done in a way that will hold teams accountable for using course content to make decisions that will be reported publically and subject to cross-team discussion/critique. You will be assigned to a team with approximately 6 members. Teams will be formed during the first week of the term. Teams will work together for most in-class activities throughout the semester.

Your grade will be influenced by team performance on team-based assignments. While in many courses, group work can be structured unfairly, such that some students end up doing all the work while everyone shares in the credit, two factors will prevent that from happening in this class. First, nearly all graded team work will be preceded by one or more preparatory assignments, for which each individual will be accountable (*e.g.* the RATs), thus ensuring that individual team members are each prepared to contribute to the team effort. Second, each individual's contribution to team work will be assessed by his or her teammates several times during the semester.

Phase 1 – Preparation: You will complete **specified readings** to begin each module

Phase 2 – Readiness Assurance Test: At the first class meeting of each module, you will be given a **Readiness Assurance Test** (RAT). The RAT test (10 multiple-choice questions) measures your comprehension of the assigned readings, and helps you learn the material needed to begin problem solving in phase 3. The purpose of phase 2 is to ensure that you and your teammates have sufficient foundational knowledge to begin learning how to apply and use the course concepts in phase 3. **RATs are closed book and based on the assigned readings.**

- **Individual RAT (iRAT)** – You individually complete a 10 question multiple-choice test based on the readings.
- **Group/Team RAT (gRAT)** - Following the iRAT, the same multiple-choice test is re-taken with your team. These tests use a "scratch and win" type answer cards known as an IF-AT. You negotiate with your teammates, and then scratch off the opaque coating hoping to reveal a star that indicates a correct answer. *Your team is awarded **10** points if you uncover the correct answer on the **first** scratch, **6** points for **second** scratch, and **2** point for **third** scratch.* No points are awarded for fourth or fifth.



IMMEDIATE FEEDBACK ASSESSMENT TECHNIQUE (IF AT)

Name Team #1 Test # 1

Subject Total 34

SCRATCH OFF COVERING TO EXPOSE ANSWER

	A	B	C	D	Score
1.			★		4
2.	★				1
3.		★			4
4.		★			2
5.				★	4
6.	★				4
7.					

- **Appeals Process** - Once your team has completed the team test, your team has the opportunity to complete an appeal. The purpose of the appeal process is to allow your team to identify questions where you disagree with the question key or question wording or ambiguous information in the readings. Instructors will review the appeals outside of class time and report the outcome of your team appeal at the next class meeting. Only teams are allowed to appeal questions (no individual appeals).
- **Feedback and Mini-lecture** - Following the RATs and Appeal Process, the instructor may provide a short clarifying lecture on any difficult or troublesome concepts.

Phase 3 - In-Class Activities: You and your team use the foundational knowledge, acquired in the first two phases, to make decisions that will be reported publically and subject to cross-team discussion/critique. We will use a variety of methods to have you report your team's decision at the end of each activity. The presentation of your team responses is critical to the team grade. You should expect each team member to present individually and for the entire team to present with smooth transitions.

Grading

Category	Assignment Type	Weight Within Category	Category Weight in the Course
Individual Grades			(45% – 70%)*
	iRAT Tests	25%	
	Individual Assignments	35%	
	Midterm Exam	15%	
	Final Exam	25%	

Team Grades			(20% – 45%)*
	gRAT Tests	50%	
	Team Exercises	50%	
Class Participation and Peer Evaluation			(10% – 25%)*
	Peer Evaluation	75%	
	Class Participation (Instructor Determined)	25%	
______________________________ _			______________________________
Total			**100%**

* The class will determine the grade weights on 08/25/2018. Student teams will negotiate the exact proportions of individual grades, team grades, and peer evaluation for the course, with in the ranges given above. For example, they may agree on individual grades at 50%, team grades at 30%, and peer evaluation at 20% of a students' course grade. The percentages *must* total to 100%, of course.

Grade Determination:

Although philosophically I would prefer not to "grade", grades for this course are based on the total number of points a student, completing all assignments successfully, would earn. Each assignment will carry a fixed number of points. At the end of the semester your final grade will be based upon the number of points you've attained divided by the maximum number of points that could possibly be attained. For example, if the maximum amount of possible points possible is 125 and you have accrued 100 points your final grade will be 100 divided by 125 or 80% (a B-); if you accrued 110 out of 125 it will be 88 (or a B+), etc. The University at Albany uses a letter-based grading system and utilizes pluses and minuses (+/-) to allow for variations of the assigned grades. Acceptable grades are **A, A-, B+, B, B-, C+, C, C-, D+, D, D-, E ("E" being the designation for failure). The University does not use grades of A+ or F.**

- 95-100=A
- 94-90 = A-
- 86-89 = B+
- 83-85 = B
- 82-80 = B-
- 76-79 = C+
- 73-75 = C
- 72-70 = C-
- 66-69 = D+
- 63-65 = D
- 62-60 = D-
- 59 and below = E (Designation for failure or E)

Policies

Attendance: Your in-class performance is key to your success in this course. Attendance, itself, is not explicitly graded (but it does factor into class participation). Instead, graded in-class activities and assignments constitute an important part of the course grade. Keeping a passing average on these is not possible without consistent attendance. Missing class means the student earns an automatic zero for all individual and team activities or assignments missed. No make-up opportunities will be available.

Tardiness: Missing an assignment or activity that happens before a student arrives or after a student leaves also earns a zero. No make-up opportunities will be available. Tardiness also factors into class participation.

If you know that it will be difficult for you to consistently get to class on time and stay for the entire period, you should take this course at a time that better fits your schedule. Missing or being late frequently will guarantee a low grade for the course.

Make-up Policy: There are generally no make-up opportunities for missed assignments except in extenuating circumstances. Instead of asking to make up missed work, please use the course ‘safety valves’ described below.

Since there will be situations in your life when missing a class meeting is simply unavoidable, this course has 2 no-fault safety valves.

Safety Valve 1: The lowest iRAT and gRAT is dropped (Peer Evaluations, individual Assignments, and Exams are *not* dropped). A missed assignment will count against this (*i.e.* a zero from a miss would be your low score; you don’t get a miss and a drop).

Safety Valve 2: If you become seriously ill during the semester, or become derailed by unforeseeable life problems, and have to miss so many assignments that it will ruin your grade, schedule a meeting with the instructor in order to make arrangements for you to drop the course to save your grade point average. Don’t wait until it’s too late to do this when you get in trouble.

Late Assignments: Out of class assignments are due on the due date, by the assigned time. Late individual assignments will be accepted, but at the cost of a full letter grade for missing the deadline, and an additional letter grade for each additional 24 hours late. In-class assignments may be done only on the days they are scheduled.

Withdrawal from the Course: The **drop date** for the Fall 2018 semester is **Monday, November 9, 2018 for graduate students in full semester courses.** That is the last date you can drop a course and receive a 'W'. It is your responsibility to take action by this date if you wish to drop the course. In particular, grades of "incomplete" will not be awarded to students because they missed the drop deadline. Given that dropping a course can have financial aid implications, please see your advisor or the Financial Aid office before dropping a course so you understand the implications that action can have on your aid.

Electronic Devices: For some team activities, you will need to use a phone/tablet/laptop. Other than that, make sure your devices are put away during class unless we are using them in a team exercise. *Non-class device use will count negatively against the entire class’s participation grade.*

Students with Disabilities: Students who feel that they have disabilities that require special arrangements for them to take the course *must* register with the Disability Resource Center. Students are eligible for special services to which both the Center and the professor agree. In general, *it is the student's responsibility* to contact the professors at least one week before the relevant assignment to make arrangements. You can contact the Disability Resource Center in Campus Center 137, or at 442-5490, if needed.

Incompletes: As per both the Graduate and Undergraduate Bulletins, the grade of Incomplete (I) will be given ”only when the student has nearly completed the course requirements but because of circumstances beyond the student's control the work is not completed.” A student granted an incomplete will make an agreement specifying what material must be made up, and a date for its completion. The incomplete will be converted to a normal grade on the agreed upon completion date based upon whatever material is submitted by that time.

Important: Incompletes will *not* be given to students who have not fulfilled their classwork obligations, and who, at the end of the semester, are looking to avoid failing the course. This is asking for special treatment.

Responsible Use of Information Technology: Students are required to read the University at Albany Policy for the Responsible Use of Information Technology available at the ITS website: https://wiki.albany.edu/display/public/askit/Responsible+Use+of+Information+Technology+Policy

Academic Integrity

In this class, some course work and examinations are *individual* exercises. The individual work that you do must be *yours* – not that of other students, friends, tutors, *etc*. While it may seem like the easy way out of doing the assignments to copy them from others, this strategy will backfire on the tests, when you will not know the material you would have learned from doing the assignments. You may of course form study groups, discuss assignments and techniques in general terms, *etc.*, but the assignments themselves *must* be your own work. In particular, two or more people may not create an individual assignment together and submit it for credit. *Please ask if you have any questions about academic integrity.*

I am also personally offended by cheating, in part because it hurts the honest students in the class. We will try our hardest to catch cheaters. If we catch a student cheating, we will not go easy on him or her. Given that, is it really worth it?

The Graduate and Undergraduate Bulletins state the university's policies on academic integrity. You will be held to these policies. You are expected to be familiar with them.

A (non-exhaustive) list of unacceptable activities is:

- Allowing other students to see or copy your assignments.
- Examining or copying another student's assignments.
- Allowing other students to see or copy your work during an exam.
- Examining or copying another student's work during an exam.
- Getting answers or help from people, or other sources (*e.g.* research papers, web sites) without acknowledging them.

- Defacing or deleting class shared documents.
- Lying to the Professor about issues of academic integrity.

Any incident of academic dishonesty in this course, no matter how "minor" will result in

- No credit for the affected assignment.
- A written report will be sent to the appropriate University authorities.
- One of -
  - A final mark reduction by *at least* one-half letter grade (*e.g.* B → B-, C- → D+),
  - A Failing mark (E) in the course, and referral of the matter to the University Judicial System for disposition.

Policies from Graduate Bulletin: http://www.albany.edu/graduate_bulletin/regulations.html

Timeline

Week	Topics	Readings
1	Big Data	Jurney, Ch. 1.
2	Big Data	Ryza, Ch. 1.
3	Agile Data Analytics/Hadoop	Jurney, Ch. 2.
4	Agile Data Analytics/Hadoop	
5	Data Issues	Jurney, Ch. 3.
6	Data Issues	Jurney, Ch. 4.
7	Spark	Ryza, Ch. 2.
8	Spark	
9	Visualization	Jurney, Ch. 5, Ryza, Ch. 7.
10	Decision Trees	Ryza, Ch. 4.
11	Anomaly Detection	Ryza, Ch. 5.
12	Prediction	Jurney, Chs. 7&8
13	Comprehensive Case Studies I	
14	Comprehensive Case Studies II	

Miscellaneous

Extra credit opportunities

During the semester the university and others hold events that may be of interest to students in this course. If you attend an event and write a summary and reflection piece on the event (specified in individual assignments) you may receive extra credit worth up to 1% of the course value. A maximum of 5% of extra credit can be accrued this way. There are no other extra credit mechanisms available in this course.