IFT Advanced Big Data/AI Syllabus, Fall 2020
version 1.0 (08-08)

Instructor:Rob Rucker **Credits:** **3**

Telephone:480 332 9798(cell)

email: robert.rucker@asu.edu

zoom: https://asu.zoom.us/my/robrucker

Office Hours :TBD

Office: Sutton 301 E

Course Description: IFT **Adv Big Data/AI** expands on the earlier content of IFT 598 Big Data Analytics. The student will learn advanced data analysis and statistical techniques that arise in data analytic and monitoring applications. In this class, you will learn and work hands-on within a cloud/cluster environment using algorithms from: Regression, Classification, Clustering, Graph , and Text analysis with an introduction to Neural Nets. Familiarity with databases and statistics is essential. NoSQL databases will be employed for persistent storage.

Objectives for this course

1.**Set the Context:** This course will continue to immerse you in the lore, principles, and algorithms of: BigData, Data Mining, Data Engineering, and Data Science (that is, all the ubiquitous languages of Big Data)

2. **Master the 'Glue' Code:** Guide you in learning the '*scripting*' languages that form the substrate and interweaveshe algorithms of Big Data, namely Scala (Python & SQL will be available as well).

3. **Configure Working Environment:** Provide you with a notebook cloud & cluster environment to implement the algorithms introduced and/or the IntelliJ IDE (or VS code)

4. **Run Real Tasks:** Introduce and exercise algorithms typically used for Big Data tasks. (see weekly content below)

Prerequisites: IFT 598 Big Datq Analytics or industry equivalent ; STP226 (intermediate stats)

Textbooks (Required):

Provost, F. and T. Fawcett (2013) *Data Science for Business*, O'Reilley

This has 14 chapters that explain what data science consists of, vocabulary and tasks. Even if the student has taken a data mining/AI course, this text will expand their understanding as to what concepts to be aware of, and , is invaluable in providing bridge concepts linking the other areas: Stats, Data Engineering, Data Science, Big Data.

ISBN: 978 144 936 1327

Chambers & Zaharia (2018) *Spark: The Definitive Guide*, O'Reilly

Scala, Python, and SQL 2003(ANSI subset) are covered co-equally. The Spark text uses all three languages interchangeably) .The Functional Reactive concepts, illustrated by Spark/Scala, are used throughout the course.

ISBN: 978-1-4842-0965

(this text provides information about the nitty-gritty of actually *doing* Big Data/A.I within a distributed computing engine environment, Spark. Written by the main creator of Spark, M. Zaharia)

Swartz, Jason(2015) *Learning Scala,* O'Reilley

(The premier scripting language for Big Data/AI (Spark is written in Scala)).

Damji, Wenig, et. al (2020) *Learning Spark 2nd edition* OReillly

ISBN: 9781492050049

Optional (not required but valuable reads):

(We will also draw on the online specs for Spark , especially the Machine Learning apis that include Neural Net (Perceptrons).

Chamberlin,D. (2018) *SQL++ A Tutorial for SQL Users ,* Couichbase Press (free download)

This short 128 pg. tutorial will get the student up to speed for using the Couchbase analytic and text search features.

Bugnion, P. (2016) *Scala for Data Science,*Packt Publishing

ISBN:978-1-78528-137-2

Wickens,T.(1995)*The Geometry of Multivariate Statistics*, Erlbaum

ISBN: 0-8058-1656-9

Reading Materials: Electronic version of additional required research papers are accessible through ASU library links.

**Valuable data sets for projects, practice, and learning from: *Kaggle.com and Universities*

**Scala tutorials are on the net and available from various MOOCs

Twitter has a Scala school for their employees that they have made generally available and is a popular way to learn Scala.

I prefer to learn from the definitive Scala text by Odersky, Spoon and Venners.

Instructor supplied videos/ files, and pdfs on an as-needed basis

Software Requirements:

Databricks (free) Community Edition: This is a cloud environment, offered by the developers of Spark, and has built in capabilities to run Hadoop and Spark programs on a *cluster*. This will be introduced in the course sequence. IntelliJ is also used for specific tasks.

Couchbase is a NoSQL database that we will use to store and query various datasets. Extensive documentation and tutorials will be available to students.(It runs on Linux, Windows, and Mac).

The agenda:

Week 1	Big Data Overview Readings from Data Science text Set up and run a Databricks notebook/ IntelliJ is ok too...
Week 2	Scala Part I, Spark Part I, as basis for Big Data, and its ecology Data Science text readings
Week 3	Spark Part II Core concepts and architecture Scala Part II Functional principles applicable to Big Data
Week 4	Scala Part III, Spark Part III, Spark Data structures
Week 5	Spark Modules and examples Spark SQL/DataFrames/Datasets (the required data formats for Big Data) Deep dive into the data types required for Data Engineering/ Data Science / Big Data and Data MIning

Week 6	Spark DataFrames, Datasets(continued) (note: Datasets are simply Scala *case class* collections, and DataFrames are simply Datasets of type ‘Row” ,i.e. Dataset[Row])
Week 7	Spark ETL with pipelnes Project proposals due (choose from: logistic regression, linear regression(multivariate clustering, similarity, naive Bayes, Text analysis,. Neural Nets)
Week 8	Linear Regression – theory and implementation (the geometry of multivariate statistics and machine learning vocabulary)
Week 9	Linear regression multivariate ratio variables and nominal variables (dummy variables such as gender, country. . . .
Week 10	Logistic Regression – theory and implementation (Multinomial as well as Bivariate) Feature extraction, transformations
Week 11	Logistic regression Part II Logistic regression as the Littlest Neural Net
Week 12	Data Lake and MLFlow investigations (New(2020) Spark modules (see the Learning Spark book)
Week 13	Delta Lake and MLFlow (cont)
Week 14	Text analysis TF-IDF analyses (and entropy again)!
Week 15	Neural net building on Logistic modues
Week 16	Standard Neural Net building (cont)
Finals week	**Final Exam** (take home) / (Zoom) Project presentations Take home final due and in-class(as necessary) presentations

Tests: There will be a quiz approximately every 4 weeks. The dates for these quizzes will be announced in advance and the final exam will be a take-home, details to follow. Unless prior arrangements are made in advance, or there is an excused absence, makeups for exams will not be given. The grade for the course is the student cumulative total divided by the cumulative possible total. There is no weighting. Refer to the table below to equate your fraction with a letter grade.

Homework: Homework will be assigned on a regular basis Homework must be **neat** and organized. If the person (usually me) grading your homework cannot evaluate your answer, you will lose points. Homework is to be submitted on **Canvas**. NOTE: submit your text, code, and graphic within a ***word*** .doc file. That is, cut and paste *within* a word document *.doc.* (Canvas will open such files and I can then review them. Don’t send me jpegs, or pngs files. Also, **File names should NOT have spaces in them.**

Grading: Your final grade for this course is your cumulative total

score / cumulative possible score and referenced to the table below.

98% - 100%	A+
93% - 97.9%	A
90% - 92.9%	A-
88% - 89.9%	B+
83% - 87.9%	B
80% - 82.9%	B-
78% - 79.9%	C+
70% - 77.9%	C
60% - 69.9%	D
0% – 59.9%	E

Classroom Policies:

1. Academic Integrity:

Each student has an obligation to act with honesty and integrity, and to respect the rights of others in carrying out all academic assignments. In IFT443, **any student who is found to have violated the academic integrity policy will, as a minimum, receive an E in the course**. The provided URL defines the policy as well as the process to be used if a student wishes to appeal this action.

Arizona State University maintains the highest standard for academic honesty and trusts that each student will perform ethically and professionally when preparing required work for this course. Each assignment must represent the student's collective original work, even for work designated as group work. Although ASU encourages collaboration between students, and faculty, in the sharing of ideas and experiences, individual work needs to represent the student's original thought and be distinguishably different from other students work. While discussions between students are encouraged, cheating will not be tolerated. Any student found cheating on an exam, a quiz, or assignment may be given a failing grade for the course and flagrant violations can result in additional consequences. You are cheating if you represent someone else's work as your own or if someone else represents your work as theirs. All graded work (exams, homework assignments, as well as any written exercises or quizzes) in this class must represent your own individual work only. Students may discuss the conceptual aspects of an assignment, but students must turn in their own, independently developed solutions. Grading will include comparing the structure and content of your solution with that of other students. By registration in this class, you are assumed to have read, understand and agreed to this policy, as well as to the procedures conveyed at the web sites below.

- § Academic integrity policy: https://provost.asu.edu/academicintegrity/
-
- § Student code of conduct: https://eoss.asu.edu/dos/srr/codeofconduct
-
- Studentlife's Student Academic Integrity Policy:
- http://www.asu.edu/studentlife/judicial/integrity.html
- Fulton School of Engineering's Academic Integrity Information Page: http://engineering.asu.edu/integrity/honor-code/
-

1. **ADA – Students with Disabilities:**

The Americans with Disabilities Act (ADA) is a federal antidiscrimination statute that provides comprehensive civil rights protection for persons with disabilities. One element of this legislation requires that all qualified students with documented disabilities be guaranteed a learning environment that provides for reasonable accommodation of their disabilities. If you believe you have a disability requiring an accommodation please contact the Disability Resource Center at ASU Polytechnic located in Sutton Hall, Suite 240, or call 480-7271039 / TTY: 480-727-1009. Eligibility and documentation policies are online at: https://eoss.asu.edu/drc. Students must request DRC involvement.

Title IX

Title IX is a federal law that provides that no person be excluded on the basis of sex from participation in, be denied benefits of, or be subjected to discrimination under any education program or activity. Both Title IX and university policy make clear that sexual violence and harassment based on sex is prohibited. An individual who believes they have been subjected to sexual violence or harassed on the basis of sex can seek support, including counseling and academic support,

from the university. If you or someone you know has been harassed on the basis of sex or sexually assaulted, you can find information and resources at https://sexualviolenceprevention.asu.edu/faqs.

As a mandated reporter, I am obligated to report any information I become aware of regarding alleged acts of sexual discrimination, including sexual violence and dating violence. ASU Counseling Services, https://eoss.asu.edu/counseling, is available if you wish discuss any concerns confidentially and privately.

1. **Other On-Campus Resources:**

There are lots of valuable resources on campus to help you achieve success both personally and academically. Don't hesitate to use them! A few of these are listed here (note that the services are available at all ASU campuses – from the links below, you can find the location(s) of the resource you're interested in at the Poly campus (or another campus):

a. Writing centers – https://tutoring.asu.edu/writing-centers

a. Tutoring, student success centers: https://tutoring.asu.edu/tutoring
b.
c. Counseling / consultation: https://eoss.asu.edu/counseling
d. Career services: https://eoss.asu.edu/cs

1. **Behavioral Policies:**

Any violent or threatening conduct by an ASU student in this class will be reported to the ASU Police Department and the Office of the Dean of Students.

1. **Absence Policies:**

Absences related to religious observances/practices that are in accord with ACD 304–04, "Accommodation for Religious Practices" (http://www.asu.edu/aad/manuals/acd/acd304-04.html) or related to university sanctioned events/activities that are in accord with ACD 304–02, "Missed Classes Due to University-Sanctioned Activities" (http://www.asu.edu/aad/manuals/acd/acd304-02.html) will not be accommodated **unless the instructor is alerted in advance such absences (advance notice = at least 72 hours)**.

Be present for class. **Do not** expect to receive a passing grade for this course if you do not attend class on a regular basis. You **are** responsible for material covered during class or
included in homework assignments regardless of whether you are in class or not.

Students who attend classes regularly have been consistently shown to perform better than students who do not. The truth of that statement may be obvious, but students do not always adhere to it. The concepts in this course are difficult to comprehend when you are first introduced to the subject. Class attendance may help provide better understanding.

1. **Keep** all homework and unit exams in case there is a problem with your grade.

2. **You are** encouraged to see me before or after class (time permitting) or make an appointment with me if you need help or have any problem with the class.

3. **Please remember,** I will do everything within reason to see that you learn as much as possible in this class, and I expect you will do the same.

4. **Cellular or Smart** phones *are* allowed in the classroom . If it rings just take it outside.