MSBA 423: Big Data Analytics

Professor: Pantelis Loupos

Office: Gallagher Hall, Room 3217

Phone: 530-752-7662

Email: ploupos@ucdavis.edu

Office Hours: By appointment

Course Description

We live in the information era, where quintillions of data are created every day. This explosion of data has led many firms to develop analytic capabilities that can deliver insights and a scientific decision making. Big data have tremendous potential to transform companies; yet, in practice many companies find real measurable value-gain to be elusive. It is all too easy to suffer from "analysis paralysis" in the face of a sea of metrics, to make misinformed recommendations based on flawed data or analytics, or to invest in an analytics tool that makes strong promises but doesn't deliver actionable insights. The ultimate goal of this class is to equip you with the state of the art big data skills to become an effective data scientist in this evolving data landscape. Emphasis is put not only on technical skills, but also on how you are going to be an effective communicator of your analytics work to your team and upper management.

Learning Objectives

By the end of the course, you will be able to:

1. Acquire a working knowledge of the current big data analytics methodologies and ecosystem, such as Distributed Computing, Streaming Processing, Text and Social Network Analytics, and Deep Learning.
2. Develop a technical understanding of the data analytics processes, and design an analytics plan, in order to evaluate questions such as:
    - Did a marketing promotion for an e-commerce retailer succeed?
    - How to develop a deep learning algorithm for cancer detection?
3. Develop a general framework for taking charge in new situations, as well as a personal philosophy and style of analytics management when collaborating and communicating with the data analytics team of your company.

Evaluation of Work	Weight
1. Class Participation	25%
2. Homework Assignments	35%
3. Final Group Project	40%

Each of these is discussed below:

1. Class Participation

- Class participation is 25% of your grade. Positive contributions that deepen our collective understanding of a topic and build class discussion increase your score. I am particularly sensitive to comments that demonstrate that you have been carefully listening to the previous speaker. You can also "earn" contribution points by sending me articles of general interest that are relevant to our class topics and discussions. A third way to enhance your participation score and for me to get to know you better is to sign-up for an office visit with me about your career plans.

2. Homework

- There will be weekly or biweekly homework assignments. Instructions for these assignments will be posted on the course web site (Canvas). In fairness to everyone in the class, **late homework assignments are not accepted.** We will discuss homework assignments in-class the day that they are submitted. All assignments must be submitted electronically through Canvas.

3. Final Group Project

- The final project requires you to apply the course concepts to a current business problem that you face or find interesting. You must develop an Analytics Plan that articulates how you will use data and analytic methods to assess your business problem. Detailed instructions will be provided in class. I invite student groups to set up a meeting with me and discuss potential final projects and methods.

Grading and "Rules of the Game"

Grading

Your course grade is a weighted average of class participation, homework assignments, and the final project. You must submit any re-grading requests via email within 10 days from when the assignment is returned. In your email, you should clearly explain why you are requesting a re-grade. While I will consider the specific concerns cited in your email, I will hold the right to re-grade the entire assignment. Please remember that small changes in your grade on a single assignment typically do not affect your overall course grade.

Classroom Etiquette

You are expected to fully follow the UC Davis honor code. When you are in class, these are my three important etiquette aspects:

1. **Electronics**: All cell phones must be muted before the start of class. The computer should only be used to take class notes. All other programs should be shut down before the start of class. Any IG/FB chatting, web surfing, e-mail etc. disturbs the class and is a breach of classroom etiquette.
2. **Seating**: All computer users should make sure to be seated in the last row of the class.
3. **Attendance and Punctuality**: Class will begin on time. Any unexplained absences, late arrivals, and/or early exits will count against students' class participation score. If you miss two sessions, this will not affect your course grade. Each additional absence will decrease your participation grade by 4%, i.e., a third absence will reduce your participation grade from 25% to 21%, a fourth absence to 17%, etc. **If you miss six or more classes, you will fail the course**. I realize that students face many issues during the quarter that impact class attendance. I will handle requests for excused absences and exceptions to the attendance policy on a case-by-case basis. Explanations must be provided to me via email no less than 24 hours prior to the start of the class. Late explanations will not be considered. When a guest speaker is scheduled for class, late arrivals and/or early exits are strictly prohibited regardless of the explanation. There are exceptions in accordance with UC Davis policy for religious holidays, funerals,

and student/dependent hospitalizations. If a student has a medical condition that may occasionally necessitate their leaving class mid-class, this should be disclosed to me via email during Week 1 of the course.

Feedback

Feedback about your course experience is super important to me. If at any point during the quarter you would like to tell me something anonymously, I have set up an online Suggestion Box. We will also have a class liaison to coordinate feedback and class activities.

Course Textbook and Material

I will provide all readings of each class in electronic format (try to save some trees by not printing everything!). We will make extensive use of the textbook "Spark: The Definitive Guide" by Bill Chambers and Matei Zaharia (O'Reilly Media, Inc.
ISBN: 9781491912201). This is a book you want to keep even after the end of this course.

Detailed Class Description and Readings

Note: A "*" will indicate an "Advance Reading". You don't have to read those, if you don't want to.

Class 1: Recap of Data Analytics

Objectives:

1. Recap of the three pillars of analytics – descriptive, predictive and prescriptive.
2. Recap of Prescriptive Analytics Methods:
    a. A/B Experiments
    b. Natural Experiments
    c. Quasi Experiments
        i. Matching
        ii. Diff n Diff
        iii. Phased Roll Outs

Readings:

1. Eric T. Anderson and Duncan Simester (2011), "A Step by Step Guide to Smart

Business Experiments"

Homework 1 (Individual): Due at the beginning of Class 3.

Class 2: Recap Continued and Introduction to Big Data

Objectives:

1. Recap of Predictive Analytics Methods
2. Introduction to Big Data
3. When do we need Big Data?
4. Business Analytics Framework: How to communicate with the C-suite.

Readings:

1. Florian Zettelmeyer and Matthias Bolling, "Big Data Doesn't Make Decisions, Leaders Do"
2. Foster Provost and Tom Fawcett (2013), "Data Science and its Relationship to Big Data and Data-driven Decision Making"
3. Nicholas Henke, Jacques Bughin, Michael Chui, James Manyika, Tamim Saleh, Bill Wiseman and Guru Sethupathy (2016), "The Age of Analytics: Competing in A Data Driven World"
4. * Xiao-Li Meng (2018), "Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election"

Plan ahead:

- **For class 3, you will need to install PySpark in your laptop. Instructions will be given in class.**
- **Try to run a simple task on AWS using an Amazon EC2 Spot Instance with Amazon EMR (link).**

Class 3: Introduction to Hadoop and PySpark

Objectives:

1. Learn about distributed computing.
2. Familiarize yourself with our main programming tool, Pyspark.

Readings:

1. Chapter 2 and 3 of "Spark: The Definitive Guide"
2. AWS Cloud Computing. Read pages 1-17 (From Introduction to AWS Cost Management). You can skim through the rest.

Plan ahead:

- **For class 4, you will need to install Kafka in your laptop. Instructions will be given in class.**

Class 4: Real Time Analytics - Streaming

Objectives:

1. Learn about Stream Processing.
2. Familiarize yourself with Kafka. This is an extremely useful tool for industrial applications. Note that we will stick to Python.

Readings:

1. Read "Getting started with Apache Kafka in Python – Towards Data Science"
2. Read chapters 20-23 of "Spark: The Definitive Guide". I know it's a lot of reading, but you are going to thank me later!

Homework 2: Kafka Application. Due at the beginning of Class 6.

Plan ahead:

- **For class 5, you will need to submit your project proposal.**

Class 5: Text Analytics

Objectives:

1. Learn about Text Analytics.

Readings:

1. Dami Lee (2019), "Emoji are showing up in court cases exponentially, and courts aren't prepared"
2. Matthew Gentzkow, Bryan T. Kelly, and Matt Taddy (2017), "Text as Data"
3. Oded Netzer, Alain Lemaire, and Michal Herzenstein (2019), "When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications"
4. Skim through pages 445-454 of "Spark: The Definitive Guide"

Class 6: Social Network Analytics

Objectives:

1. Learn about Social Network Analytics.
2. Familiarize yourself with GraphX library.

Readings:

1. Pantelis Loupos, Alexandros Nathan, and Moran Cerf (2019), "Starting Cold: The Power of Social Networks in Predicting Non-Contractual Customer Behavior"
2. Chapter 30 of "Spark: The Definitive Guide"

Homework 3: Venmo Application. Due at the beginning of Class 8.

Class 7: Deep Learning: Part 1

Objectives:

1. Learn the fundamentals about Deep Learning.
2. Familiarize yourself with Tensor Flow.

Readings:

1. Chapter 1 and 2 of "Neural Networks and Deep Learning" by Michael Nielsen. It's in your class readings folder.
2. Chapter 31 of "Spark: The Definitive Guide"
3. * Natalie Wolchover, "New Theory Cracks Open the Black Box of Deep Learning"

Class 8: Deep Learning: Part 2

Objectives:

1. Learn about Generative Adversarial Networks (GANs).

Readings:

1. Martin Giles, "The GANfather: The man who's given machines the gift of imagination"
2. Dan Falk, "How Artificial Intelligence Is Changing Science"
3. * Goodfellow et al., "Generative Adversarial Networks"
4. * Hristina Uzunova, Jan Ehrhardt, Fabian Jacob, Alex Frydrychowicz, and Heinz

Handels, "Multi-scale GANs for Memory-efficient Generation of High Resolution Medical Images"

Homework 4: DL Application. Due at the beginning of Class 10.

Class 9: Advanced Topics

Objectives:

1. Learn about Reinforcement Learning
2. Ethics of Big Data
3. Prepare for what's next - Causal AI and Quantum Computing.

Readings:

1. Will Knight, "Google's AI Masters the Game of Go a Decade Earlier Than Expected"
2. Read section 2.6 of "Artificial Intelligence and Games" by Georgios N. Yannakakis and Julian Togelius -- provided in your class folder
3. Read pages 6-9 from "21 lessons for the 21st century" by Yuval Noah Harari -- provided in your class folder.
4. Geoffrey Fowler, " The spy in your wallet: Credit cards have a privacy problem" (link)
5. Kevin Hartnett, "To Build Truly Intelligent Machines, Teach Them Cause and Effect". I highly recommend buying and reading the "Book of Why: The New Science of Cause and Effect".
6. "Machine learning, meet quantum computing"
7. * Silver et al., "Mastering the game of Go with deep neural networks and tree search"
8. * Bard et al., "The Hanabi Challenge: A New Frontier for AI Research"

Class 10: Projects Presentation - Your time to shine!