UCLA Extension Data Science Certificate Updated 9/24

Transforming Data into Knowledge and Vision
Understanding the Power and Beauty of Data

Instructor: **Dr. William Yu** *Quarter:* Fall 2018

UCLA Extension Public Contact: **Damar Douglas,** ddouglas@unex.ucla.edu, (310) 825-7609

Meeting times: MWTh: 6:00-9:00pm at Gayley 121B

Students can expect to do additional 15-20 hours of work outside of the classroom.

Course Website: Canvas will be the main site for communications, discussion, materials, etc.

Email: william.yu@anderson.ucla.edu *Office phone:* (310) 825-7805

Office: Anderson School C-506 *Office hours:* MWTh: 9:00-9:30pm or appointment

Description

This certificate is designed to fully immerse students in the fundamentals of the practice of data science. The goal of this course is to prepare students for work as a data analyst or data scientist. Instruction strikes a balance between the rich and beautiful theory behind this exciting field and putting principles and open-source software tools into practice on real datasets. Students learn the why of data analysis but also the how using tools such as R, RStudio, and Python.

Courses and Sequence

The UCLA Extension Data Science Certificate is a four-course program that includes the following courses:

- UCLA CSX 450-1 // Introduction to Data Science
- UCLA CSX 450-4 // Machine Learning with R
- UCLA CSX 450-2 // Exploratory Data Analysis and Visualization
- UCLA CSX 450-3 // Hadoop and Managing Big Data

To assist in the fast-paced nature of the program, the program uses a unique sequencing. Courses CSX 450-1, CSX 450-4, and CSX 450-3 are taught in sequence, each course building upon the previous courses knowledge. CSX 450-32, *Exploratory Data Analysis and*

Visualization, supplements the week's lessons in modeling and analysis with lessons in engineering and infrastructure.

Course Outline (subject to change)

Week	Date	Course Content	Data Project
1	9/24, 26,27	Introduction: Data science, data cleaning and management, R, RStudio, statistics, and basic infrastructure	Project 1
2	10/1, 3, 4	Introduction to Machine Learning (ML) models, Linear regression	Project 2
3	10/8, 10, 11	ML: Logistic regression, classification models: KNN, LDA, Caret	Project 3
4	10/15, 17,18	Classification models: random forest, naïve Bayes, SVM, ANN	Project 4
5	10/22,24,25	Unsupervised learning: PCA, K-means clustering, dimensionality reduction and model selection	Project 5
6	10/29,31, 11/1	Introduction to Python, Anaconda, Spyder, Jupyter, its applications	Project 6
7	11/5,7,8	Predictive analytics and time series modeling and forecasting	Project 7
8	11/**13**,14,15	Visualization: exploratory and interactive data analysis	Project 8
9	11/19,**20**,21	Big data analytics and applications	Capstone project
10	11/26,28,29	Deep learning and capstone project presentations and discussions	Capstone project

Note: 11/12 and 11/22 are holidays so we meet on **11/13 and 11/20** instead. The classroom will be at Gayley **121E** for these two special Tuesday meetings.

All the data projects will use real-world dataset, big or small. You will be assigned and guided to clean, manage, analyze these data and apply the tools and models you learn to turn these data into interesting stories and valuable knowledge and insights.

Outcomes

Upon completion students can expect to have a solid grasp of applied data analysis and machine learning principles. Students will be able to work with an unseen dataset from the retrieval and cleaning phase through the modeling and presentation phase. Students will have

produced at least one portfolio-ready capstone project and prepared them for presentation using GitHub.

Textbooks

- "An Introduction to Statistical Learning with Applications in R," by James, Witten, Hastie, and Tibshirani, Springer. (http://www-bcf.usc.edu/~gareth/ISL/). Free download: http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf.
- "Machine Learning with R," 2nd Ed., Packt, 2015, by Brett Lantz.
- "Using R for Introductory Statistics," by John Verzani. Simple to understand. Free download: https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf.
- "Introduction to Statistical Thought," by Michael Lavine. Free download: http://people.math.umass.edu/~lavine/Book/book.pdf

Description for Each Course

Introduction to Data Science

This course introduces students to the evolving domain of data science and to the food-chain of knowledge domains involved in its application. Students learn a wide range of challenges, questions, and problems that data science helps address in different domains, including social sciences, finance, health and fitness, and entertainment. The course addresses the key knowledge domains in data science, including data development and management, machine learning and natural language processing, statistical analysis, data visualization, and inference. The course also provides an exposure to some of the technologies involved in application of data science. The course includes case studies that require students to work on real-life data science problems.

Machine Learning Using R

This course focuses on machine learning, which is concerned with algorithms that transform information into actionable intelligence. This field is made possible due to the rapid and simultaneous evolution of available data, statistical methods and computing power. The machine learning language, R, is a cross-platform, zero-cost statistical programming environment, which offers a powerful but easy-to-learn set of tools that can assist students with finding data insights. Students learn the origins and practical applications of machine learning, how knowledge is defined and represented by computers, and the basic concepts that differentiate machine learning approaches. Machine learning algorithms can be divided into two main groups: supervised learners that are used to construct predictive models, and unsupervised learners that are used to build descriptive models. Students learn the

classification, numeric predictor, pattern detection and clustering algorithms. Students learn to train a model, evaluate its performance, and improve its performance. Algorithm uses are illustrated with real-world cases, such as breast cancer diagnosis, spam filtering, identifying bank loan risk, predicting medical expenses, estimating wine quality, identifying groceries frequently purchased together, and finding teen market segments.

Exploratory Data Analysis and Visualization

The key goal of Data Science is to obtain insights from data. The insights could be about what happened in the past by analyzing historical data or about predicting what may happen in the future using predictive analytics. Data Scientists go through an iterative process to come up with means that lead to insights. This process is called Exploratory Data Analysis (EDA). In addition to a curious mind, data exploration and data visualization are key requisites for EDA. This course will teach you these skills with a specific focus on visualization. You will learn the iterative process of EDA, data analysis techniques, data exploration, and visualization. The course uses tools such as R Programming for data analysis, and several packages for data visualization.

Hadoop and Managing Big Data

The extent of data being produced and stored by organizations is increasing. In fact, IDC has projected to reach 40 zetta bytes by 2020. Organizations understand that being able to extract and leverage value and gain actionable insights from this big data can give them a tremendous competitive advantage. In this course, you learn all about Hadoop—its evolution a framework consisting of tools for distributed storage and data processing, to an open-source framework. This course addresses distributed storage and large data set processing focusing on architectures and technologies, specifically Hadoop. Additionally, students learn about other elements in the Hadoop ecosystem, NoSQL databases, and competing technologies. Students also install, setup, and use Hadoop on a single node.

Academic Honesty Policy

Academic dishonesty covers behavior in cheating, plagiarism, and fabrication of information. These behaviors are not tolerated. Students are encouraged to familiarize themselves with the UCLA Extension Student Conduct Code and the official statements regarding cheating and plagiarism at: https://www.uclaextension.edu/Pages/str/StudentConduct.aspx

Services for Students with Disabilities

In accordance with Section 504 of the Rehabilitation Act of 1973 and the Americans with Disabilities Act of 1990, UCLA Extension provides appropriate accommodations and support services to qualified applicants and students with disabilities. These include, but are not limited to, auxiliary aids/services, such as note takers, audiotaping of courses, sign language interpreters, and assistive-listening devices for hearing-impaired individuals, extended time for and proctoring of exams, and registration assistance. Accommodations and types of support services vary and are specifically designed to meet the disability-related needs of each stude3nt based on current, verifiable medical documentation. Arrangements for auxiliary aids/services are available only through UCLA Extension Disabled Student Services at (310) 825-4581 (voice/TTY) or by email at access@uclaextension.edu. Please request such arrangements with at least five working days' advance notice. All assistance is handled in confidence. Accommodations must be pre-approved. Requests for retroactive accommodation will not be accepted.

Grading

- **Component**
  - Class attendance and participation 30%
  - Projects 40%
  - Capstone project and presentation 30%
- **Scheme**

Grade Percentage	Letter Grade	Pass/Fail
90-100%	A	Pass
80-89%	B	Pass
70-79%	C	Pass
60-69%	D	Fail
>59%	F	Fail

Learning and Using R

- Download R
  - https://cran.r-project.org/bin/windows/base for Windows
  - https://cran.r-project.org/bin/macosx for Mac
- Download R Studio, a more productive platform for R
  - https://www.rstudio.com/products/rstudio/download
- Introduction to R
  - A short introduction to R

    - A introduction to R
    - Elementary Statistics with R
    - UCLA IDRE Resources on R
- Predictive Analytics
    - Forecasting: principles and practice (R based) by Rob Hyndman and George Athanasopoulos. https://www.otexts.org/fpp
    - Statistical forecasting: note on regression and time series analysis, by Robert Nau. http://people.duke.edu/~rnau/411home.htm

Public Datasets

- https://github.com/awesomedata/awesome-public-datasets
- Federal Reserve Economic Data (FRED, https://research.stlouisfed.org/fred2/)
- Bureau of Economic Analysis (BEA, http://www.bea.gov/)
- Bureau of Labor Statistics (BLS, http://www.bls.gov/)
- Census' American Fact Finder, including American Community Survey (ACS), providing abundant demographic, economic, social, housing data by state, county, city, and zipcode!
- Google Public Data Explorer with dynamic graphing.
- Wall Street Journal (WSJ)'s Economic Indicators Archive, Economic Forecasting Survey, Market Data, including all kinds of stock, bond, currency, commodities data
- Yahoo Finance (http://finance.yahoo.com/), providing downloadable historical finance data
- Fed Regional Banks' data, indicators, and research: New York, San Francisco, Chicago, Kansas City, Dallas, Philadelphia, Cleveland, Atlanta, Boston, Minneapolis, Richmond, DC Board
- Gapminder, a fact-based worldview, with a very cool chart!
- Visualizing Economics, providing a lot of vivid charts and maps in the world!
- World Bank's World Development Indicators and Worldwide Governance Indicators
- International Monetary Fund (IMF)'s World Economic Outlook Databases
- National Bureau of Economic Research (NBER)'s data, collecting a bunch of great research data!
- International regional and country's statistical sites by BLS and by EDIRC
- Index of Economic Freedom and CIA the World Factbook
- Robert Shiller's stock and housing markets data, providing useful financial market data
- S&P/Case-Shiller home price indices
- Federal Housing Finance Agency (FHFA)'s House Price Index
- Zillow Real Estate Research data (http://www.zillow.com/research/data/)
- IMF Global Housing Watch and Dallas Fed's International House Price Database
- BIS Property Price Statistics
- First 5 LA/UCLA Anderson Forecast City Human Capital Index (CHCI)
- Real-time search data Google Trends and Google Domestic Trends tracking sectors in economy
- Google Books Ngram Viewer (word frequency in books from 1800 to date)
- Angus Maddison data, two thousand year GDP per capita data around the world.
- Citylab's Maps (http://www.citylab.com/posts/maps/)