STEVENS INSTITUTE OF TECHNOLOGY

FE-582: Foundations of Financial Data Science

Syllabus

FE582 Instructor: Dragos Bozdog
Office: Babbio 429A
Email: dbozdog@stevens.edu
Phone: (201) 216-3527

FE513 Co-Requisite Instructor: Xiaodi Zhu (Coco)
Email: xzhu@stevens.edu

Time: FE-582: Tuesday (6:00pm-7:50pm)
FE-513 Co-Requisite: Tuesday (8:00pm-8:50pm)

Room: Hanlon Financial Systems Lab (Babbio 4th floor) and Blackboard Collaborate (Online)

Office Hours: By appointment

Description: This course will provide an overview of issues and trends in data quality, data storage, data scrubbing, data flows, and data encryption. Topics will include data abstractions and integration, enterprise level data issues, data management issues with collection, warehousing, preprocessing and querying. Furthermore, the Hadoop based programming framework for big data issues will be introduced along with any governance and policy issues. These concepts will be applied to areas such as digital marketing and computational advertising, energy and healthcare analytics, social media and social networks, and capital markets financial data. A one credit Hanlon lab course, FE-513: Practical Aspects of Database Design will be attached to this course in order to facilitate learning of the practical side of data management.

Objective: This course is the first course for the certificate in Financial Services Analytics. Financial services analytics is the science and technology of creating data-driven decision making analytics for the financial services industry. This can lead to more effective business operations, enhanced customer services and product offerings, and improved risk analysis and risk management. This course is the key building block in this certificate as good data and the understanding of data is critical to the creation of robust financial services analytics. The financial services analytics certificate has four key areas making up its knowledge base. They are
Foundations of Financial Data Science (FE-582)
Introduction to Knowledge Engineering (FE-590)
Financial Systems Technology (FE-595)
Data Visualization Applications (FE-550)

Prerequisite FE 513 – Practical Aspects of Database Design

Textbooks: No single textbook covers all the topics. Several references will be used and supplementary notes will be provided whenever appropriate.

General References:

1. Charu C. Aggarwal, *Data Classification: Algorithms and Applications.* CRC Press, *2015.* (ISBN: 978-1-4665-8674-1)
2. Charu C. Aggarwal, *Data Mining.* Springer, 2015. (ISBN: 978-3-319-14141-8)
3. Deborah Nolan and Duncan T. Lang, *Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving*, CRC Press, 2015. (ISBN: 978-1-4822-3481-7)
4. Norman Matloff, *The Art of R Programming,* No Starch Press, 2011. (ISBN: 978-1-59327-384-2)
5. Cathy O'Neil and Rachel Schutt, *Data Science,* O'Reilly, 2014. (ISBN: 978-1-449-35865-5)

Outcomes:

After taking this course, the students will be able to:

1. Have a working knowledge of the issues of data quality, data storage, data scrubbing, data flows, and data encryption and their potential solutions.
2. Understand and design various schemas needed for the representation of financial data.
3. Tackle problems dealing with data management issues such as collection, warehousing, preprocessing and querying.
4. Develop and evaluate strategic data initiatives around governance and policy requirements.
5. Will get a primer on database management as well as advantages and disadvantages from the attached lab course FE 513.
6. Understand how to write applications using the map-reduce feature of Hadoop clusters.
7. Have a working understanding of all the databases available for them through the Hanlon lab.
8. Apply the newly acquired data management and database skills to financial data from the capital markets, social media, and the financial services sector.

Grading:

Assignments 50%
Final Assignment 50%

Graduate Student Code of Academic Integrity:

All Stevens, graduate students promise to be fully truthful and avoid dishonesty, fraud, misrepresentation, and deceit of any type in relation to their academic work. A student's submission of work for academic credit indicates that the work is the student's own. All outside assistance must be acknowledged. Any student who violates this code or who knowingly assists another student in violating this code shall be subject to discipline.

All graduate students are bound to the Graduate Student Code of Academic Integrity by enrollment in graduate coursework at Stevens. It is the responsibility of each graduate student to understand and adhere to the Graduate Student Code of Academic Integrity. More information including types of violations, the process for handling perceived violations, and types of sanctions can be found at www.stevens.edu/provost/graduate-academics .

	Topic	Reference
Week 1	Introduction to Data Science	
Week 2	Financial Data Quality Issues and Data Scrubbing. Data Preparation.	
Week 3	Similarity and Distances.	
Week 4	Introduction to Relational Databases	
Week 5	Using SQL Basics for Preprocessing and Querying Data	
Week 6	Outlier Analysis: Extreme Value Analysis. Probabilistic Models. Clustering and Distance-Based Outlier Detection. Density-Based Methods.	
Week 7	NoSQL and Hadoop Cluster principles and applications Principles and categorization. MongoDB	
Week 8	Hadoop. Pig. Hive. HBase.	
Week 9	Case Study (Data Manipulation and Modeling):	Ref. 4: pp. 45-100, pp. 105-164, pp. 217-236
Week 10	Case Study (Simulation):	Ref. 4: pp. 241-276, pp. 367-394
Week 11	Case Study (Data and Web Technologies):	Ref. 4: pp. 399-415
Week 12	Case Study (Twitter Data)	
Week 13	Case Study (High-Frequency Big Data Warehouse and Querying Solutions)	
Week 14	Final Assignment Presentations	