DSC 510: Introduction to Data Science and Analytics

Instructor: Professor George Pallis, Computer Science Department
Lab instructor: Dr. Pavlos Antoniou, Teaching Special Staff
ECTS: 8
Fall Semester

This course will examine how data analysis technologies can be used to improve decision-making. The aim is to study the fundamental principles and techniques of data science, and we will examine real-world examples and cases to place data science techniques in context, to develop data-analytic thinking, and to illustrate that proper application is as much an art as it is a science. In addition, this course will work hands-on with the Python programming language and its associated data analysis libraries.

Content

Thanks to advances in Internet computing, large-scale distributed infrastructures, and modern software tools, it is now possible to process and analyze data at an unprecedented scale. This capability enables us to extract invaluable insights from the vast amounts of heterogeneous data generated daily, ranging from business transactions and scientific experiments to social media activity and sensor networks. Consequently, both industry and academia are experiencing a revolution fueled by one of the most sought-after professional profiles of the 21st century: the **data scientist**.

This course introduces students to the **fundamental steps of the data science pipeline**, combining theoretical foundations with practical applications and hands-on use of modern tools:

- **Data Wrangling**
  - Techniques for data acquisition (scraping, crawling, parsing, API integration, open data portals).
  - Data manipulation and preprocessing for structured, semi-structured, and unstructured datasets.
  - Common sources of data problems, including missing values, noisy or inconsistent data, and strategies for data cleaning and reconciliation.
  - Ensuring data quality through systematic testing.
- **Data Handling**

  - **Handling Text**: preprocessing textual data, tokenization, stemming, lemmatization, vector space representation, embeddings, and topic modeling.
  - **Handling Networks**: representing, analyzing, and visualizing graphs and social networks, including community detection, centrality measures, and influence analysis.
- **Data Interpretation**
  - Approaches for working with “found data” and the design of observational studies.
  - Text mining techniques, including the vector space model, topic modeling, and word embeddings for natural language understanding.
  - Social network analysis methods for identifying influencers, community structures, and diffusion patterns.
- **Data Visualization**
  - Principles of effective visualization: plot types for one, two, and three variables; design and layout best practices.
  - Using visualization to detect data quality issues, scaling methods for large datasets, and techniques for visualizing uncertainty.
- **Machine Learning**
  - Introduction to supervised and unsupervised learning, model selection, and evaluation.
  - Feature engineering, data vs. algorithm trade-offs, and the challenges of high-dimensional data.
  - Practical considerations, such as interpretability, overfitting, and computational efficiency.
- **Observational Studies and Causality**
  - Methods to distinguish correlation from causation.
- **Fairness, Anonymization, and Ethical Concerns**
  - Ensuring fairness and mitigating bias in data-driven models and decision-making systems.
  - Data privacy, anonymization techniques, and compliance with ethical and legal requirements.
  - Responsible use of data science and AI with respect to fairness, accountability, and transparency.

Students will be introduced to these concepts in lectures and reinforced through lab sessions using the Python programming language and its data analysis ecosystem (e.g., Pandas, scikit-learn, and visualization libraries). In parallel, they will engage in a **semester-long project**, working in agile teams to design, implement, and evaluate a data-driven solution to a real-world problem. The outcome will be a public, open-source project portfolio. The course concludes with a final two-hour exam, ensuring a balance between theoretical understanding and practical expertise.

Learning Outcomes

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

- **Demonstrate a coherent understanding** of the principles, techniques, and software tools required to perform all fundamental steps of the Data Science pipeline.
- **Acquire and integrate data** from multiple sources and formats (e.g., structured, semi-structured, unstructured), using Web scrapers, REST APIs, open data repositories, and big data platforms.
- **Wrangle and preprocess data** effectively, addressing issues such as missing values, incorrect entries, inconsistent representations, and assessing overall data quality.
- **Interpret and analyze data** by applying machine learning and exploratory methods, while demonstrating critical thinking, collaboration in team discussions, and the ability to create ad-hoc visualizations for insights.
- **Communicate and disseminate results** through well-structured reports, effective visualizations, reproducible workflows, and critical reflection on ethical, societal, and privacy considerations in data science practice.

Teaching methods:

- Physical in-class recitations and lab sessions
- Homework assignments
- In-class quizzes

Course project Transversal skills:

- Evaluate one's own performance in the team, receive and respond appropriately to feedback.
- Give feedback (critique) in an appropriate fashion.
- Demonstrate the capacity for critical thinking
- Write a scientific or technical report.

Expected student activities:

- Attend the lectures and lab sessions
- Complete 3-4 quizzes (held during lab sessions)
- Conduct the class project
- Read/watch the pertinent material before a lecture
- Engage during the class, and present their results in front of the other

colleagues

Assessment methods:

- 10% continuous assessment during the semester
- 60% final exam
- 30 % final project, done in groups of 3 students (max)

Students who accumulate more than four (4) unexcused absences from lectures will be ineligible to take the course examinations.

Similarly, students who accumulate more than three (3) unexcused absences from lab sessions will be ineligible to submit the final project.

Course Book:

Data Science for Business: What you need to know about data mining and data analytic thinking. Provost & Fawcett (O'Reilly, 2013) (Updated 2019) http://data-science-for-biz.com/

This book covers the fundamental material that will provide the basis for you to think and communicate about data science and business analytics. We will complement the book with discussions of applications, cases, and demonstrations, and possibly some additional readings or notes for material that is not covered in the book.

One particularly useful book for those interested in the "hands-on" component of the class:

(OPTIONAL)
Python Machine Learning: Machine Learning and Deep Learning with Python, scikit- learn, and TensorFlow, 2nd Edition by Sebastian Raschka & Vahid Mirjalili

Use of AI tools: Students should follow the following recommendations.

https://www.ucy.ac.cy/graduateschool/wp-content/uploads/sites/45/2023/10/ENG-Recommendations-for-the-use-of-

Artificial-Intelligence-in-teaching-processes-at-UCY-starting-Fall-Semeter-2023-2024-.pdf