**EGRMGMT 590-01: Sourcing Data for Analytics**

Fall 2020: Course Syllabus

**Course Times**

TBD

**Instructors**

**Jon Reifschneider**

**Title:** Director, MEng in AI for Product Innovation

**Affiliation:** Pratt School of Engineering

**Contact:** [jon.reifschneider@duke.edu](mailto:jon.reifschneider@duke.edu), 919-660-7403

**Office:** Fitzpatrick Center 3559

**Office Hours:** TBD

**Daniel Egger**

**Title:** Executive in Residence

**Affiliation:** Pratt School of Engineering

**Contact:** [Daniel.egger@duke.edu](mailto:Daniel.egger@duke.edu), 919-660-5443

**Office:** Fitzpatrick Center 3575

**Office Hours:** TBD

**Course Description**

In practice within industry, one of the main activities, and challenges, in implementing machine learning applications is collecting data to use in modeling. This course introduces students to the technical and also the non-technical (business, regulatory, ethical) aspects of collecting, cleaning, and preparing data for use in analytics applications. Technical aspects covered will include types of data, methods of sourcing data via the web, APIs, and from domain-specific sensors and hardware (IoT devices), an increasingly common source of analytics data in technical industries. The course also introduces methods and tools for evaluating the quality of data, performing exploratory data analysis, and pre-processing data for use in analytics. Students will build skills in numerical programming via hands-on work with industry-standard Python libraries for cleaning and manipulating data. Legal and ethical aspects covered include an introduction to data privacy, GDPR, regulatory issues, bias, and industry-specific concerns regarding data usage.

**Pre-Requisites**

Students are expected to understand the main concepts of calculus, linear algebra and probability & statistics, as well as possess a foundational level of proficiency in Python programming.

**Learning Objectives**

Through this course, students will be expected to:

* Understand the different types of data and their applications in modeling
* Understand the various sources for data (sensors/hardware, APIs/web, etc) and be knowledgeable in methods to collect data from each source
* Demonstrate skills in working with data in Python, via the Numpy and Pandas libraries
* Develop experience in collecting and pre-processing data for use in analytics models, via hands-on programming
* Be able to evaluate the usefulness of datasets for analytics purposes, including measures of quality as well as quantity
* Demonstrate skills in analyzing data via exploratory data analysis
* Build an appreciation for important regulatory and ethical considerations when sourcing data for use in AI
* Gain experience in the end-to-end process of sourcing data for use in AI modeling – from identifying data needs, determining potential sources, assessing legal and ethical concerns, evaluating potential sources, cleaning and pre-processing data, and performing exploratory data analysis

**Course Materials**

Required textbook:

* “Python Data Science Handbook: Essential Tools for Working with Data”, by Jake VanderPlas, O'Reilly Media; 1 edition (December 10, 2016), ISBN-13: 978-1491912058, full text and code freely available at https://jakevdp.github.io/PythonDataScienceHandbook/.

Required free software:

* Python 3.7.x (suggest installing Python via the Anaconda distribution (<https://www.anaconda.com/distribution/>)
  + Python is today the pre-dominant language used in industry data science and ML teams
* The following libraries must also be installed (can be installed using pip or conda):
  + Numpy
  + Pandas
  + Jupyter Notebook
  + Matplotlib

**Course Grading**

* 40% Homework assignments (4 assignments)
* 30% Project (5% midterm checkup, 20% final report, 10% final presentation)
* 20% Final Exam
* 10% Class Participation

**Course Schedule**

|  |  |  |  |
| --- | --- | --- | --- |
| Session | Module | Topics | Deliverable Due |
| Session 1 | **Introduction** | - Value of data  - Types and sources of data  -Evaluating data needs |  |
| Session 2 | **Tools for working with data** | - Introduction to Numpy, Pandas, Matplotlib  - Loading & manipulating data |  |
| Session 3 | **Web data sources** | - Collecting and processing web data via web scraping and APIs |  |
| Session 4 | **Sensor data sources** | - Introduction to time series  - Collecting & processing sensor data | Assignment 1 due |
| Session 5 | **Labeling data** | - Defining target values for modeling  - Target labeling methods | Project proposals due |
| Session 6 | **Exploratory data analysis** | - Summary statistics  - Histograms and KDE  - Feature correlation | Assignment 2 due |
| Session 7 | **Dealing with messy data** | - Assessing data quality  - Dealing with outliers  - Methods of handling missing data |  |
| Session 8 | **Pre-processing data for modeling** | - Scaling & standardizing  - Using categorical features in modeling  - Dimensionality reduction / PCA | Assignment 3 due |
| Session 9 | **Data lineage & replicability** | - Modeling pipeline  - Data lineage / replicability |  |
| Session 10 | **Data privacy & regulation** | - Industry-specific regulations  - Data privacy and GDPR | Assignment 4 due |
| Session 11 | **Ethical & bias considerations** | - Understanding data bias  - Methods of dealing with bias |  |
| Session 12 | **Project presentations** |  | Project reports due |
| Session 13 | **Final exam** |  |  |

**Class Policy**

Students are expected to follow the Duke community standard:

1. I will not lie, cheat or steal in my academic endeavors, nor will I accept the actions of those who do
2. I will conduct myself responsibly and honorably in all my activities as a Duke student

Collaboration on homework assignments is encouraged at the “whiteboard” level. However, all answers, code etc is to be original. Any sources used must be cited – plagiarism is a violation. Homework, projects or exams that violate this class policy will receive zero credit.

**Class Project**

A significant portion of the grade for this course comes from a team-based course project in which students will put their learning into practice in sourcing data for a real-world industry modeling problem. Students will form teams of 3-5 members and choose a topic area from one of the Program’s focus industries: Energy & Environment, Healthcare/BioTech, or Manufacturing. Students will submit a one-page project proposal document which must be approved prior to proceeding. The focus of the project will be on identifying an industry problem that can be solved via machine learning or other analytics, and applying the skills learned in the course to identify data sources, collect and integrate data, evaluate data quality, analyze and visualize the datasets, and conduct pre-processing in preparation for modeling. Students will also provide an analysis of the relevant regulatory issues related to use of the datasets proposed, and identify potential ethical concerns that may arise by using the data in AI algorithms to solve the industry problem identified.