"Essentials of Data Science" Bootcamp

January 23rd-February 20th (Saturdays 10am - 6pm)

Instructors: S. Samsonau, Ph.D, A. Shehu, Ph.D.

Invited lecturers: TBA

Organized with a support from NYCASCENT
Location: TBA
Syllabus is subjected to change

The bootcamp addresses the industry need for data scientists. The design of the course aims for graduate students/postdocs in STEM field without previous exposure to data analysis techniques. The material introduced by combining theory and practice using "coding by example" paradigm. After each module students will be able to implement in practice certain methods. The practical experience will be sufficient to develop a deeper and wider understanding by reading a recommended literature and working on projects.

Week 1. Getting Started. Data Science with R

Class Objectives

- The class will start with a discussion about Data Science field in general. After that students will practice to use basic data structures in R. Such topics as reading data, reshaping data, split-apply-combine paradigm, plotting will be addressed. Basic git operations using R studio The following packages will be emphasized: readr, readxl, stringr, lubridate, plyr, dplyr, reshape2, ggplot2.

Recommended reading

- Data Manipulation with R - Second Edition, Abedin, Jaynal
- R Graphics Cookbook

Week 2 Data pre-processing. Regression.

Class Objectives

- It will begin with cleaning and modifying original data. After that linear regression will be practiced. Such ideas as bias-variance tradeoff, over-fitting, creation of new variables will be addressed. The next part will include estimating a model performance, re-sampling methods, selection of variable subset, regularization methods. Regression trees will be discussed as an example of an alternative algorithm.
  The following package will be emphasized: caret

Recommended reading

- An Introduction to Statistical Learning: with Applications in R
- Caret package website

Week 3 Classification. Unsupervised Learning.

Class Objectives

- Students will practice classification algorithms such as logistic regression, KNN, classification trees, SMV, as well as methods for model performance evaluation. The second part of the class will address unsupervised learning.
  The following package will be emphasized: caret

Recommended reading

- An Introduction to Statistical Learning: with Applications in R
- Caret package website

Week 4. Data mining as a whole process.

Class Objectives

- This class will consider the process of creating data science product as a whole. Rattle package will be used as an additional option for use in data mining tasks. Reporting tools such as knitr, R presentation, Slidify, and Shiny will be addressed.
  The following packages will be emphasized: Rattle, caret, slidify, shiny.

Recommended reading

- Data mining with Rattle and R
- Data Science for Business

Week 5. Text mining

Class Objectives

- Deriving high quality information from large texts via pattern recognition. We introduce the tm package to present methods for corpus handling, preprocessing, metadata management and creation of term-document matrices. Text mining methods will be used for sentiment analysis and fraud detection.
  The following packages will be emphasized: tm

Recommended reading

- Introduction to text mining in R