**Data Science Foundations Curriculum**

Master in Big Data Solutions

# Subject description

**Subject: Data Science Foundations**

**Year:** 2017-18

**Quarter:** 1st

**Degree:** Master in Big Data Solutions

**Number of credits: 8 / Hours of class: 84 / Hours of homework: 84**

**Teaching Staff:**

Name: Liana Napalkova

Email: liana.napalkova@bts.tech

Linkedin: <https://www.linkedin.com/in/liananapalkova/?ppe=1>

Name: Ludovico Boratto

Email: ludovico.boratto@bts.tech

Linkedin: <https://www.linkedin.com/in/ludovico-boratto-774367a4/?ppe=1>

Name: Francisco Sacramento Gutierres

Email: francisco.gutierres@bts.tech

Linkedin: [ttps://www.linkedin.com/in/francisco-gutierres-2303a511/?ppe=1](https://www.linkedin.com/in/francisco-gutierres-2303a511/?ppe=1)

**Schedule:**

* 4 hour sessions from 9:00 – 13:00
* Weekly homework: 4 hours

# Subject introduction and goals

Nowadays the world's most valuable resource is no longer oil, but data. We are surrounded by digital artefacts, for example, an e-library, e-mail, e-shopping, e-ticket, e-payment, e-governance, etc. The cost of data collection and storage goes down and the ability to process large amounts of data steadily increases. In such settings the Data Science becomes a crucial component that defines the competitiveness of the organizations. This science deals with applying automated methods to analyse massive amounts of data and to extract knowledge from them. The availability of Big Data technologies allows a wide application of the Data Science methods in a business domain.

Additionally, researchers in academia and industry are investing a lot of efforts to improve the value of geospatial big data as well as take advantage of its value. During the last years, there has been a significant change in the way researchers are developed novel big data solutions based on geospatial data. In fact, arises several new emerging platforms for sharing the collected geospatial big data and for tracking human mobility via mobile devices, traffic monitoring, mobile phone data for tourism, urban statistics and other application domains.

At the end of this course students will acquire the following knowledge and skills:

* Manage data using Python
* Use Python to load, process and save structured data to external files in different formats, such as JSON, CSV, TXT.
* Perform basic operations with data using Pandas DataFrame's and NumPy arrays.
* Learn the basics of relational databases and how to write Structured Query Language (SQL) queries to access the data.
* Interact with a database using Python.
* Perform time series analysis and forecasting in Python
* Learn how to fit data to probability distributions and regression models
* Use Python/Pandas to perform general statistical analysis in their data.
* Apply appropriate and useful Geographic Information System (GIS) tools and Geostatistics in Python to find spatial data patterns.

# Teaching methodology

The “Part I: Practical Data Manipulation in Python” applies the learn-by-doing methodology. Students will practice the data handling and Python programming skills by working with realistic data sets. The data handing assignments will involve the tasks that are common for the majority of real-world Data Science projects.

The “Part II: ​ Probability and Statistics” methodology have been devised and planned so that this theoretical/practical approach is adopted throughout BTS programs. Groups will be small, allowing maximum interaction between teachers and students and among the students themselves. Our teaching methodologies include the following:

· Lectures and analysis of teaching notes.

· Lecture-demonstration by teacher.

· Class discussion conducted by teacher.

· Analysis and discussion of case studies in small groups as well as in class.

· Presentations by student panels from the class: class invited to participate.

· Student reports by individuals.

· Student-group reports by committees from the class.

· Individual reading and research.

· Reading assignments in journals, monographs, etc.

The practical approach adopted for part I (reflection, analysis, action and evaluation) aims to help students acquire and consolidate their knowledge, focusing always on decision-making and professional practice.

Everyday classes are complemented by a series of extracurricular activities (lectures, practical training, etc.) to help students expand their knowledge, focus their careers, gain professional experience, and try out initiatives and proposals drawn up by experts in different fields of action.

# Contents

**1.1. Part I: Practical Data Manipulation in Python**

The practical data manipulation means common actions that should be performed with data prior to further analysis (descriptive, predictive or prescriptive one). Such actions involve cleaning, filtering, joining, etc. A special attention is given to the unstructured data that should be converted into a structured (tabular) data in order to enable the knowledge extraction.

1. Basic data operations on relational data

1.1. Loading/saving a data file (unless already covered by pre-course)

1.2. Cleaning noisy data, normalizing

1.3. Selecting columns

1.4. Filtering rows

1.5. Joining data from multiple files on an attribute

1.6. Working with text

2. Loading and filtering text

2.1. Text pre-processing, normalization, stemming, stopword removal

2.2. Converting text to vectors

2.3. Computing text similarity

**1.2. Part II: Probability and Statistics**

Probability and Statistics are two of many statistical techniques that can be used to analyze big data data to find useful patterns.

Therefore, these traditional statistics can be used by students in two different ways. In the first case, the non-spatial statistics considering a large set of data values that one wants to understand, and using descriptive, predictive and prescriptive statistics to try to summarize them.

In the second case, through the spatial statistics, which are designed for use with spatial data-with geographic data. These methods essentially use space-area, length, proximity, direction, orientation, or some conception of how the objects in a dataset interact with each other- just like in mathematics. In fact, this characteristics makes spatial statistics different from traditional statistical methods. However there are descriptive spatial statistics alike to descriptive traditional statistics. For example, if we have several lots of Geo-tagged tweets on the map, we might want to identify where the center of those points is located. The equivalent traditional statistic would involve computing the mean for a set of Twitter data values. We also might also want to see how spread out those points are around the center. So, this is similar to calculating the standard deviation for a set of Twitter data values.

3. Statistics re-cap

3.1. Probability

3.2. Estimation

4. Introduction to data analytics

4.1. Analytic models: descriptive, predictive and prescriptive

4.2. The process of Data Analytics

4.3. Overview of analytic techniques and application

4.4. Libraries, frameworks and services for Data Analytics

4.5. Related disciplines

**1.3. Part III: Basic Data Analysis**

5. Basic data types

5.1. Vectors, matrices, and tensors

6. Structured data types

6.1. Graph processing analysis examples

6.2. Time series analysis examples

6.3. Text processing example

7. Overview of main analytic techniques

7.1. Statistics-based modelling

7.2. Principal component analysis

# Schedule of contents and activities

|  |  |  |
| --- | --- | --- |
| **Session** | **Activity at class (4 hours)** | **Activity at home (4 hours)** |
| **Session 1.1.1**  2-oct    Liana Napalkova | **Content**  Brief introduction into common data formats, such as JSON, CSV and TXT.    Description of how to manage data using Pandas DataFrames.    **Activity**  Load data (given in CSV and TXT formats) into Pandas DataFrames.  - Perform basic manipulations with data, such as view data and describe a quick statistic summary  - select data by label and position  - Boolean indexing    Commit scripts to Git. | **Individual Assignment**  Select any open source data in JSON, CSV or TXT formats:  - Load data into Pandas DataFrame  - Answer a series of questions related to the data  - Commit scripts to Git |
| **Session 1.1.2**  5-oct    Liana Napalkova | **Content**  Description how to work with SQLite database    **Activity**  Create SQLite database:  - Import data (given in CSV and TXT formats) into SQLite database  - Load data from database into Pandas DataFrame and perform basic manipulations with data    Commit scripts to Git. | **Individual Assignment**  Select any open source data in JSON, CSV or TXT formats:  - Create SQLite database  - Load data into SQLite database  - Read data from database into Pandas DataFrame  - Answer a series of questions related to the data  - Commit scripts to Git |
| **Session 1.1.3**  9-oct    Liana Napalkova | **Content**  Description of how to perform cleaning, filtering and joining of data using Pandas DataFrames and NumPy arrays.    **Activity**  Download city bike stations or weather data (JSON format):  - Clean and store data  - Analyse an hour of city bike data    Commit scripts to Git. | **Individual Assignment**  Given city bike stations or weather data, answer a series of questions and solve challenges.    Commit scripts to Git. |
| **Session 1.1.4**  11-oct    Ludovico Boratto | **Content**  Introduction to text processing in Python and to the NLTK library.    **Activity**  Loading text from file, sentence tokenize and word tokenize, normalizations, and part-of-speech tagging.    Commit scripts to Git. | **Individual Assignment**  Get data that is not raw text (JSON or CSV formats), perform text loading, filtering and pre-processing with the NLTK library.    Commit scripts to Git. |
| **Session 1.1.5**  16-oct    Ludovico Boratto | **Content**  Text stopping, stemming and lemmatization. TF-IDF.    **Activity**  Load text from a file, and perform text stopping, stemming and lemmatization with the NLTLK library. TF-IDF with the scikit-learn library.    Commit scripts to Git. | **Individual Assignment**  Take a different text file, perform text stopping, stemming and lemmatization with the NLTLK library and produce some basic statistics of how the text changed after these operations (e.g., number of final tokens w.r.t. the initial ones).  Perform TF-IDF on the text.    Commit scripts to Git. |
| **Session 1.2.1**  18-oct    Francisco Gutierres | **Content**  Introduction to descriptive statistics using Pandas’ DataFrame, Python and the NumPy library.    **Activity**  Perform Descriptive statistics with data using DataFrames, such as Means and averages. Variance. Distributions. Representing histograms. Plotting histograms. Representing Probability mass function (PMFs). Plotting PMFs. Outliers. Conditional probability. | **Individual Assignment**  Select any open source data:  - Load data into Pandas DataFrame and NumPy array  - Develop Descriptive Statistics for Pandas Dataframe and with NumPy  - Answer a series of questions related to the data  - Commit scripts to Git |
| **Session 1.2.1**  23-oct    Ludovico Boratto | **Content**  Basic metrics for text similarity (Euclidean distance, cosine similarity).    **Activity**  Load text from a file, and perform text similarity with NLTK.    Commit scripts to Git. | **Individual Assignment**  Take two text files and detect plagiarism (i.e., how similar the two documents are).    Commit scripts to Git. |
| **Session 1.2.3**  25 - oct    Francisco Gutierres | **Content**  Cumulative distribution functions and Continuous distributions.    **Activity**  Implementation of Python algorithms to calculate the class size paradox, the limits of PMFs, Percentiles, Cumulative distribution functions, Representing CDFs, Conditional distributions, Random numbers.    Operations on distributions: Skewness, Random Variables, Probability density function (PDFs), Convolution, Central limit theorem, The distribution framework. Hypothesis testing.    Commit scripts to Git. | **Individual Assignment**  Exercises with different data sources.    Commit scripts to Git. |
| **Session 1.2.4**  30-oct    Francisco Gutierres | **Content**  Continuous distributions.    **Activity**  Implementation of Python algorithms to:  - calculate the exponential distribution  - the Pareto distribution  - the normal distribution  - the normal probability plot  - the lognormal distribution    Commit scripts to Git. | **Individual Assignment**  Exercises with different data sources.    Commit scripts to Git. |
| **Session 1.2.5**  3-nov    Francisco Gutierres | **Content**  Introduction to Probability: rules of probability, Binomial distribution, Bayes’s theorem.    **Activity**  Fitting data to probabilistic models in Python.    Commit scripts to Git. | **Individual Assignment**  Exercises on Probability with different sources of data.    Commit scripts to Git. |
| **Session 1.2.6**  6-nov    Ludovico Boratto | **Content:**  Basic data types: vectors, matrices, and tensors.  Main operations performed on them.    **Activity**  Load data from a file and represent in vector-, matrix-, and tensor-form.    Perform simple operations on them.    Commit scripts to Git. | **Individual Assignment**  Using a different source of data, practice with the data structures and their manipulation through exercises.    Commit scripts to Git. |
| **Session 1.2.7**  8-nov    Ludovico Boratto | **Content**  Introduction to graphs (definition, finding paths, degree).    **Activity:**  Implementation of algorithms to build a graph in Python, find a shortest path in it and find its degree.    Commit scripts to Git. | **Individual Assignment**  Practical exercises to modify the algorithms implemented in class.    Commit scripts to Git. |
| **Session 1.3.1**  13-nov    Francisco Gutierres | **Content**  Estimation and Correlation.    **Activity**  Estimation and Correlation in Python.    Estimation:  - Exponential distributions  - Confidence intervals  - Bayesian estimation    Correlation:  - Standard scores  - Covariance  - Correlation  - Making scatterplots in pyplot  - Spearman’s rank correlation  - Least squares fit  - Goodness of fit  - Correlation and Causation    Commit scripts to Git | **Individual Assignment**  Exercises on estimation and correlation with different sources of data.    Commit scripts to Git. |
| **Session 1.3.2**  15-nov    Francisco Gutierres | **Content:**  Data analytics in Phyton and Geographic Information Systems (GIS).    **Activity**  Load geospatial data (given in Shapefile, CSV and TXT formats) into ArcMap environment. Summary Statistics (Analysis). Analytic models: descriptive, predictive and prescriptive.    Commit scripts to Git and ESRI Git. | **Individual Assignment**  Take a different geospatial file, perform summary statistics with the ArcMap and ArcGIS Pro Python window script to demonstrate how to use the Statistics tool in immediate mode.    Commit scripts to Git and ESRI Git. |
| **Session 1.3.3**  23 – nov    Ludovico Boratto | **Content**  Graph density, graph connection, distance and diameter of a graph.    **Activity**  Implementation of Python algorithms to calculate graph density, graph connection, distance and diameter of a graph.  Commit scripts to Git. | **Individual Assignment**  Practical exercises to modify the algorithms implemented in class.    Commit scripts to Git. |
| **Session 1.3.4**  27 – nov    Ludovico Boratto | **Content**  Time Series introduction.    **Activity**  Loading and handling time series in Pandas. Check stationarity of a time series. How to make a time series stationary.    Commit scripts to Git. | **Individual Assignment**  Exercises on data series handling and stationarity with different sources of data.    Commit scripts to Git. |
| **Session 1.3.5**  29– nov    Ludovico Boratto | **Content:**  Time series forecasting.    **Activity:**  Time series forescasting in Pandas.    Commit scripts to Git. | **Individual Assignment**  Time series forescasting in Pandas with a different source of data.    Commit scripts to Git. |
| **Session 1.3.6**  4 – des    Ludovico Boratto | **Content**[1]  Text classification.    **Activity**  Using Stanford Text Analysis Tools in Python. Preliminary study on text classification.    Commit scripts to Git. | **Individual Assignment**  Text classification exercises.    Commit scripts to Git. |
| **Session 1.3.7**  11 – des    Ludovico Boratto | **Content:**  External Maximum Entropy modelling. Introduction to sentiment analysis    **Activity:**  Using External Maximum Entropy Modelling libraries for text classification. From text classification to sentiment analysis.    Commit scripts to Git. | **Individual Assignment**  Exercises on External Maximum Entropy Modeling for text classification and sentiment analyisis.    Commit scripts to Git. |
| **Session 1.3.8**  14 – des    Ludovico Boratto | **Content:**  Statistics-based modelling. Fitting data to probability distributions.    **Activity:**  Fitting data to probability distributions in Python.    Commit scripts to Git. | **Individual Assignment**  Exercises with different data sources.    Commit scripts to Git. |
| **Session 1.3.9**  18 – des    Ludovico Boratto | **Content:**  Principal component analysis. Fitting data to regression models. | **Individual Assignment**  Exercises with different data sources.    Commit scripts to Git. |

The Schedule of activities can be modified according to the program needs.

# Qualification system

The evaluation system is based on role-playing, classroom work, presentation of hands-on GitHub assignments and final exam. In order to successfully complete the Subject “Data Science Foundations”, participants should meet all BTS requirements.

**Participation: 30%**

Class participation

In-class activities in individual and/or groups

Collaborative initiatives

**Assignments: 40%**

Continuous assessment by delivery of individual exercises and group exercises.

**Final exam/project: 30%**

Theoretical evaluation

Practical evaluation

**Conditions to recover the subject**

If the final mark is below 5, students will be able to deliver a small project with an real application of the presented techniques.

**85% of attendance is required.**

# Bibliography

**Recommended:**

|  |  |
| --- | --- |
| **Kaggle (pay attention to Kernels)** | <https://www.kaggle.com/c/titanic> |
| **The Python Tutorial** | <https://docs.python.org/2/tutorial/> |
| **Tutorialspoint Python Tutorial** | <http://www.tutorialspoint.com/python/> |
| **Code Academy** | <http://www.codecademy.com> |
| ***Think Stats. Probability and Statistics for Programmers* (*book*)** | <http://greenteapress.com/thinkstats/> |
| **Data Science from Scratch**  **First Principles with Python (*book*)** | [http://math.ecnu.edu.cn/~lfzhou/seminar/[Joel\_Grus]\_Data\_Science\_from\_Scratch\_First\_Princ.pdf](http://math.ecnu.edu.cn/~lfzhou/seminar/%5bJoel_Grus%5d_Data_Science_from_Scratch_First_Princ.pdf) |
| **Python Scripting for ArcGIS 2nd Edition (*book*)** | https://www.amazon.com/gp/product/1589483715/ref=as\_li\_qf\_sp\_asin\_il\_tl?ie=UTF8&camp=1789&creative=9325&creativeASIN=1589483715&linkCode=as2&tag=gislounge-20&linkId=2RFYDRIAOWXUWJCW |
| **Python in ArcGIS Pro documentation** | <http://pro.arcgis.com/en/pro-app/arcpy/get-started/installing-python-for-arcgis-pro.htm> |

**Complementary:**

|  |  |
| --- | --- |
| **Geospatial Big Data: Challenges and Opportunities (*research paper*)** | <http://www.sciencedirect.com/science/article/pii/S2214579615000040> |
| **Python and GIS Resources (website)** | <https://www.gislounge.com/python-and-gis-resources/> |
| **Python Scripting for ArcGIS (website)** | https://www.gislounge.com/python-scripting-arcgis/ |

**Other academic resources:**

|  |  |
| --- | --- |
| **Geonet GitHub** | <https://github.com/geonet> |
| **Esri GitHub** | <https://github.com/esri> |

Bibliography and other academic resources will be detailed and updated in the Campus.

# Software requisites

● Students should have a Github account

● Students should have Python installed, including:

o Anaconda 2.7, 3.6 (Jupyter)

o Python libraries: SQLite, NLTK, NumPy, Pandas, Scikit-Learn, Stanford text analysis tools, Scipy, Seaborn, Matplotlib.

● ArcGIS:

o Students should request a trial ArcGIS organization account (<http://www.esri.com/en/arcgis/products/arcgis-pro/DesktopFreeTrial>):

▪ ArcGIS Online

▪ ArcMap

▪ ArcGIS Pro