**UT Summer Statistics Institute**  
Department of Statistics and Data Sciences  
The University of Texas at Austin

Instructor: Tal Yarkoni

Course: Introduction to Data Science in Python

Course Description: Modern data scientists have a bewildering array of tools at their disposal. In recent years, Python has emerged as a language of choice for many data scientists due to its appealing combination of flexibility, power, and extensive community support. This short course surveys the Python software ecosystem and familiarizes participants with cutting-edge data science tools. Topics include interactive computing basics; data preprocessing and cleaning; exploratory data analysis and visualization; and machine learning and predictive modeling. Participants will explore core concepts in data science and Python via hands-on, interactive exploration and analysis of sample datasets.

**Day 1** focuses on setting up an ideal environment for data science in Python. Topics include:
- Overview of the Python programming language, including pros and cons of Python relative to other common data science languages (R, Matlab, etc.)
- Introduction to the core scientific computing packages (Numpy/Scipy/Matplotlib)
- Setting up a Python data science environment with Anaconda and pip
- Conducting reproducible, interactive analyses using the IPython/Jupyter notebook
- Maintaining best practices for data science (version control, code profiling, etc.)

**Day 2** focuses on preprocessing and cleaning data in preparation for analysis. Topics include:
- Importing and exporting data from common formats (.csv, .xls, etc.)
- Basic text processing in Python
- Data structures and data manipulation in pandas
- Common data munging operations (aggregating, reshaping, etc.)
- Sanitizing and cleaning data (outlier detection, missing values, etc.)

**Day 3** focuses on data exploration, visualization, and statistical analysis. Topics include:
- Visualizing data in Python using matplotlib and seaborn
- Descriptively summarizing data
- Basic inferential statistics (t-tests, linear regression) in Python using statsmodels
- Cursory overview of Python packages for advanced or domain-specific data analysis (time series analysis, Bayesian modeling, geospatial analysis, etc.)

**Day 4** covers machine learning and predictive modeling in Python. Topics include:
- Introduction to scikit-learn
- Feature selection and transformation
- Classification and regression analysis
- Data decomposition and clustering
- Cross-validation and model evaluation
- Constructing and applying fully reproducible data science workflows