Course Title: Statistical and Machine Learning for Big Data

Dates: 06 – 10 July 2020

Tuition fees:
  Professionals: 1,350 CHF
  External students: 1,000 CHF
  UNIGE students: 600 CHF (limited number of places)

Early bird fee: no

Course description:

The main learning objective of the one-week summer course is to provide advanced undergraduate students, graduate students and researchers with formal training in analytics and skills necessary to apply popular statistical and machine learning methods to high dimensional data. This course includes several multivariate analytical methods and some new univariate methods. Health and business research examples particularly large medical insurance, environmental exposure, cancer, and suicide datasets will be comprehensively analyzed. This context based short course provides skills to analyze different types of data using primarily R programming. Programming instructions are also provided in Python.

Students can work independently on big data analysis by the end of this one week course and also collaborate with other researchers on projects that require statistical expertise for large and complex data analyses. Topics of the course include:

- **Matrix algebra review using introductory R**
- **Introductory topics** (analytics and big data, predictive models, introduction to Python, machine and deep learning packages, bootstrapping, cross validation, GLM, GAM, high dimensional data presentation, Monte Carlo methods)
- **Supervised Learning** (k-NN, naive Bayes, SSVM, EM algorithm, Copula based methods, regularization methods, unbalanced data, undersampling and oversampling)
- **Unsupervised Learning** (cluster analysis, PCA, latent variables and factor analysis, decision trees, bagging, random forests, boosting)
- **Artificial Intelligence** (feedforward neural nets, backpropagation algorithm, contrastive divergence (CD), stochastic MLE, score/ratio matching, pseud-likelihood estimation, annealed importance sampling, deep belief networks)
- **Case Studies**

Course notes and materials. The e-copy of lecture notes with computer programs will be designed and provided by the presenter before the start of the course.

Course credits: equivalence of 3 ECTS, based on a personal/group project

Target Audience: students with interests in analytics methods and numerical sciences.
**Prerequisites:** A knowledge of fundamental probability and statistics concepts (e.g. acquired at a non-specialized bachelor level) as well as a beginner-level familiarity with the R statistical programming language are strongly desirable.

**Faculty & Staff:**

**Program Directors:**
- Prof. Maria-Pia Victoria-Feser, GSEM, University of Geneva
- Prof. Stephane Guerrier, GSEM, University of Geneva

**Instructors:**

**Prof. Si Konda:** Si Konda is an assistant professor in Biostatistics at University of Illinois at Chicago (UIC). He has considerable experience in developing and teaching statistical and machine learning courses. He provided multiple two day analytics and machine learning seminars for Society of Actuaries from 2016-2019 and conducted one-day machine learning workshop at Abbott Labs, Chicago, USA in 2019. He is the recent golden apple award winner for excellence in teaching & leadership at UIC. Si Konda was a visiting faculty at University of Waterloo in 2012 and University of California at Santa Barbara from 2013 to 2015. Si Konda has a Ph.D. in Statistics from the Case Western Reserve University in Cleveland, USA.