UNIVERSITÀ DEGLI STUDI FIRENZE

FLORENCE center for DATA SCIENCE

# $D^2$ SEMINAR SERIES

*Florence Center for Data Science 'Double' Seminar Series*

Florence Center for Data Science is happy to present the next seminar of the Series on **March 17th**, from **2.30 - 4 pm**

Click on the link to register online:
https://us02web.zoom.us/webinar/register/WN_mHAHeMr0RkKgv-eXiUsyzQ

## SPEAKERS, TITLES, ABSTRACTS:

**Alberto Cassese - Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence.**

Title: "Bayesian negative binomial mixture regression models for the analysis of sequence count and methylation data"

Abstract: A Bayesian hierarchical mixture regression model is developed for studying the association between a multivariate response, measured as counts on a set of features, and a set of covariates. We have available RNASeq and DNA methylation data on breast cancer patients at different stages of the disease. We account for heterogeneity and over-dispersion of count data by considering a mixture of negative binomial distributions and incorporate the covariates into the model via a linear modeling construction on the mean components. Our modeling construction employs selection techniques allowing the identification of a small subset of features that best discriminate the samples, simultaneously selecting a set of covariates associated to each feature. Additionally, it incorporates known dependencies into the feature selection process via Markov random field priors. On simulated data, we show how incorporating existing information via the prior model can improve the accuracy of feature selection. In the case study, we incorporate knowledge on relationships among genes via a gene network, extracted from the KEGG database. Our data analysis identifies genes that are discriminatory of cancer stages and simultaneously selects significant associations between those genes and DNA methylation sites. A biological interpretation of our findings reveals several biomarkers that can help to understand the effect of DNA methylation on gene expression transcription across cancer stages."

**Chiara Bocci - Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence.**

Title: "Sampling design for large-scale geospatial phenomena using remote sensing data"

Abstract:In many fields of application, it is common to be interested in spatially-related phenomena and, in particular, to deal with attributes that, being defined on continuous spatial domains, are observed on a fine grid. For such kind of data, selecting the units spatially well spread over the study area allows to collect more information and consequently provides a better estimation of the population parameters. Moreover, technological advances have led to a growing availability of ready-to-use low-cost spatial data, like remote sensing data, which can be used as auxiliary information in the sampling design development process in addition to the units' spatial location.
Several sampling methods in literature simultaneously implement the selection of well-spread samples and the use of auxiliary variables in the selection process. These designs implicitly assume that, besides having a spatial pattern, the study variable shows a relationship with some auxiliary variables. However, this relationship is never known exactly, and for large-scale phenomena, it is often unrealistic to assume a unique relationship that holds everywhere. Therefore, we propose a two-step sampling design to identify when and how it is useful to exploit the auxiliary information, in addition to the units' spatial location, in the second step of the sampling selection process. We evaluate the performance of our proposal through Monte Carlo experiments in two simulation studies: one on pseudo-real datasets and one on synthetic datasets.
This talk is based on joint work with Saverio Francini and Emilia Rocco.