Ph.D. in Data Analytics and Decision Science:

Thesis Defense May 19th, 2022 Aula Alario (Building 21) – at 14.30

Fabio Azzalini – XXXIV Cycle

Advisor: Prof.ssa Letizia Tanca

Data Integration and Ethical Quality: Fundamental Steps of the Data Analysis Pipeline

Abstract

Data Science plays a very important role in the current society. In many scenarios, it allows to obtain insights that have a critical impact on our daily lives (e.g. precision medicine, fraud detection or autonomous vehicles), that otherwise would be impossible to achieve. Unfortunately, often the data sources used in data science applications are very heterogeneous and this prevents us from easily using them in data analytic tasks. In this context, before getting to the actual data modeling phase, it is necessary to apply a series of methods to provide the data science algorithms with correct and reliable data. Specifically, often times the data comes from different sources that need to be integrated. Additionally, the data provided by the sources are often of poor quality, and can present ethical problems which, if not solved, would affect the final decisions of the prediction algorithms.

This thesis presents a collection of methods and tools to improve the quality of datasets and to prepare them for being used in data science tasks.

In the data integration process, Entity Resolution has the role of identifying records that refer to the same data item. Due to the huge size of today's data sources, modern methods, use the so-called Blocking techniques to improve their efficiency by partitioning the initial dataset into smaller, quicker to execute, blocks. Traditional blocking techniques fails at identifying semantically similar values since they only consider the syntactical aspects of the data. To overcome these challenges, in this thesis we propose, LSH-Embeddings and Clust-Embeddings, two automatic blocking strategies that aim at capturing the semantic properties of data by means of recent Deep Learning frameworks.

Another step of the Integration pipeline, Data Fusion, addresses the problem of discovering the true values of a data item when multiple sources provide different values for it. In this thesis we propose STORM, a novel domain-aware algorithm for data fusion designed for the multi-truth case, that is, when a data item can also have multiple true values. To determine the true values STORM assesses the trustworthiness of the sources by taking into account their authority: here, we define authoritative sources as those that have been copied by many other ones. As a further support to the Data Fusion phase, the thesis also proposes Deep-Fusion, a multi-truth data fusion method, specifically designed to work with data sources containing dirty values or text written in natural language, very frequent in current integration problems.

Another issue that arises in current data science has to do with ethics, since, for an application to be reliable, it should be associated with tools to discover bias in data, in order to avoid (possibly unintentional) unethical behavior and consequences. In this thesis we propose E-FAIR-DB, a novel solution that, exploiting the notion of Functional Dependency - a type of data constraint - aims at enforcing data ethics by discovering and solving discrimination in datasets.