Mining Heterogeneous Urban Data at Multiple Granularity Layers

PhD Candidate:
Antonio Attanasio

1. Introduction
In the last few years, the capability to both generate and collect data of public interest within urban areas has increased at an unprecedented rate, to such an extent that data rapidly scale towards big urban data. The abundance of information collected through ad-hoc sensor networks in the city context provides a remarkable opportunity to tackle interesting urban challenges and to add intelligence in the urban environment. However, for each data source and type, different spatial and temporal references are potentially used. Moreover, the high volume and heterogeneity of data increases the complexity of the analysis and new suitable algorithms should be devised. When massive volumes of data are considered, alternative efficient data storage and processing technologies are required.

The research activity aims at the design and development of innovative data mining solutions and it focuses on data analysis algorithms, suitable to mine useful insights by exploring large and heterogeneous data collections, deployed on cloud-based platforms to guarantee good performance in the mining process [4]. The PhD thesis proposes novel data mining algorithms and it enhances the existing ones by fitting them to novel big data architectures. The adoption of high performance and cloud computing infrastructures is considered to leverage on parallel computations. Real datasets are used to assess the proposed approaches.

2. Methods and results
In the PhD research, data mining solutions have been studied and developed to address the following issues.

![Figure 1. Architecture of the BI2CITY engine with space-time data aggregation framework.](image)

A distributed business intelligence engine (BI2CITY) has been developed to efficiently support the integration and analysis of huge and various data collections. BI2CITY stores fine-grained data collected in the urban area. Then, it computes the temporal and spatial data aggregation on the fly to transpose the original data into the proper resolution as required by the analysis performed by the user. BI2CITY allows to correlate different urban data (weather, pollution, traffic, energy consumption) and to forecast their expected values for every spatio-temporal aggregation level. To efficiently deal with huge data collections, BI2CITY is based on MongoDB and the MapReduce paradigm. BI2CITY has been suitably applied to different use cases, where many types of data are used to provide insightful descriptive and predictive analyses. It has been validated on real data collected in a major Italian city [5].

![Figure 2. Regression analysis at 1 hour resolution and algorithm execution time vs number of nodes.](image)

3. Conclusion
The proposed data mining approaches proved to be effective solutions to get useful knowledge from heterogeneous data in complex urban application domains. The described results confirm the importance of analyzing data with suitable granularity levels, in order to extract patterns and relationships among variables that are significant for the purposes of the analysis. Moreover, the implemented algorithms exhibited a good scalability with big datasets as well.

4. References