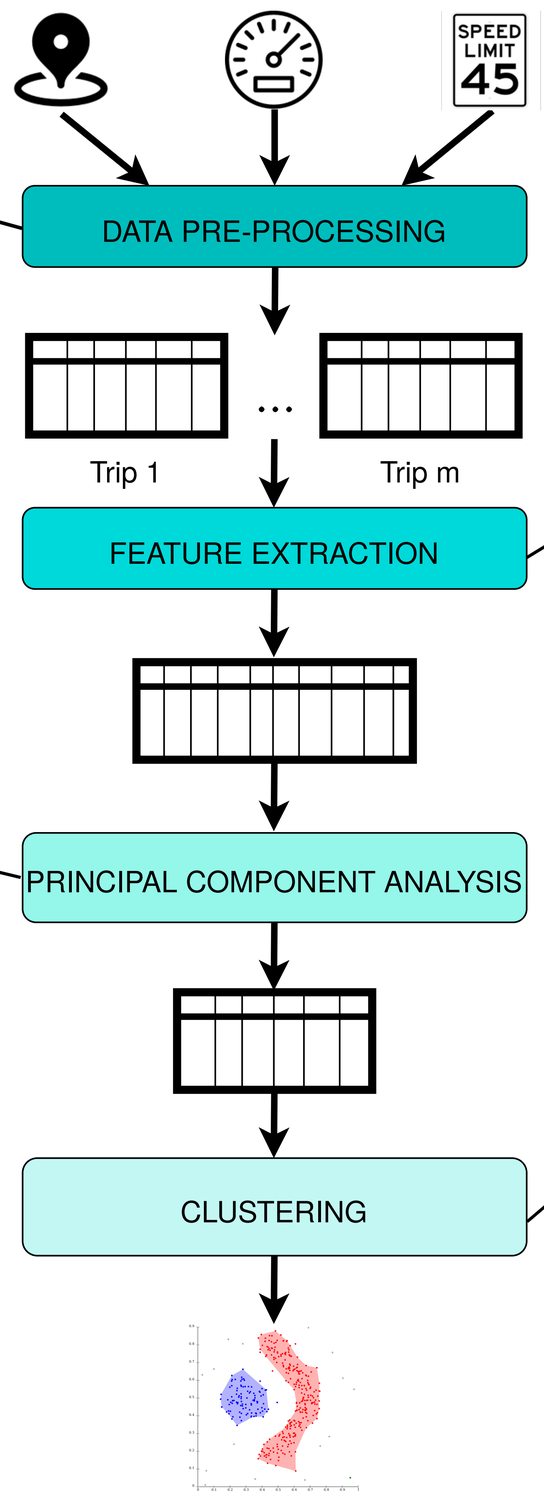
An Unsupervised Approach for Automotive Driver Identification

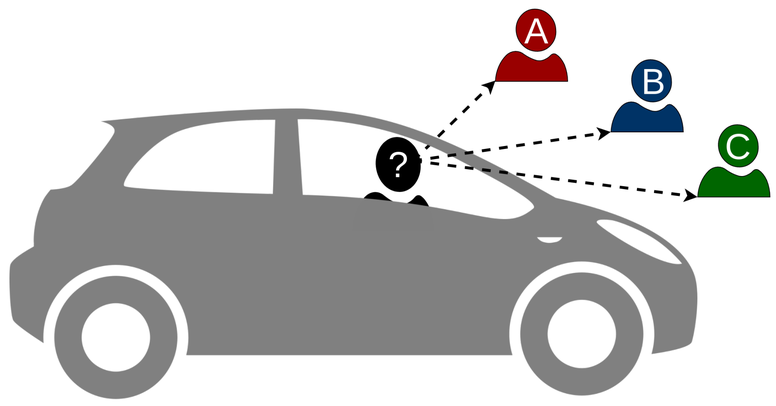
Nicholas Mainardi, Michele Zanella, Federico Reghenzani,
Niccolò Raspa, Carlo Brandolese *({name.surname}@polimi.it)*

The adoption of on-vehicle monitoring devices allows different entities to gather valuable data about driving styles, which can be further used to infer a variety of information for different purposes, such as fraud detection and driver profiling. In this work, we focus on the identification of the **number of people usually driving the same vehicle**, proposing a data analytic work-flow specifically designed to address this problem.
Most of the works found in literature exhibit two relevant limitations: (a) they rely on input data mainly retrieved with *invasive* methodologies (i.e., reading from the vehicle Electronic Computed Board or the CAN bus) and (b) they leverage supervised techniques. Conversely, our approach is based on **unsupervised learning algorithms** working on **non-invasive data gathered from a specialized embedded device**.

The Proposed Workflow

In our approach, non-invasive monitoring device (plugged to OBD for power supply only) gathers:
(a) **Motion data** (accelerations and angular velocity)
(b) **GPS data** (position, altitude, speed)
(c) **Reverse geocoded information** (Type of road and speed limit)
Then data are partitioned in **trips,** i.e., sets of measurements collected from the engine switch-on to the engine switch-off. Finally, we discard the GPS position and altitude, since they have no relation withthe driving style.

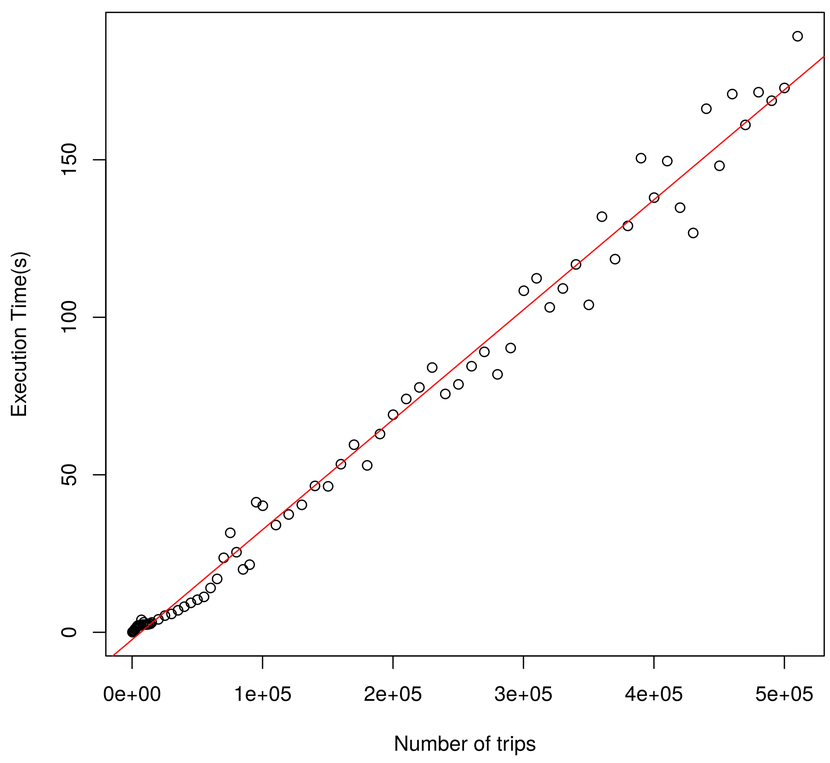




We devise a set of **features** for a trip to characterize the driving style following two approaches:
(a) aggregating the motion data samples with some **statistical measures** – mean, variance, skewness and kurtosis – providing a set of purely statistical domain-agnostic features;
(b) designing a set of **specific features** in order to characterize the driving style.

In order to avoid the *curse of dimensionality issue* due to the high dimensionality of the clustering space, we need to shrink the number of features. To this purpose, we apply **Principal Component Analysis** (PCA). In our case, by retaining 75% of the statistical significance of the data, we reduce the number of features from 40 to 5–7 features depending on the trip.

The last step performs the clustering. We choose to employ a density-based approach, in particular the **DBSCAN** [1] algorithm because:
(a) it does not require to specify the number of clusters to be built as an input parameter
(b) it labels as noise points those ones not belonging to any cluster, which may be useful for anomaly detection.



Preliminary Evaluation

The evaluation of the proposed workflow is focused on two key aspects:
(a) **Accuracy analysis**. In order to check the accuracy of our workflow, due to the unavailibaility of labelled dataset, we compute the average number of drivers per vehicle for the UK region, being this value strongly dependent on the geosocial conditions. We get the average number of adult people per vehicle [2] and the number of drivers per adult people obtaining an average value of drivers per vehicle of 1.095. Applying the proposed workflow to trips of the same vehicle, the results show that the number of cars with more than one driver is around 10%.
(b) **Scalability analysis**. Given that the computational complexity of PCA is linear in the number of trips n, the average-case complexity of the overall program is dominated by DBSCAN algorithm, i.e. $O(n \log n)$. To measure actual execution times, we perform an experimental evaluation on a single-core R implementation on AMD Opteron 8435 cores and 128 GB of RAM. The plot shows that the actual value of execution time can be approximated by a linear trend, due to the minimal impact of the $\log n$ term.

References

[1] M. Ester, H.P. Kriegel, J. Sander, and X. Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *In Proceedings of the 2 nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, Oregon, USA, Evangelos Simoudis, Jiawei Han, and Usama M. Fayyad (Eds.). AAAI Press, 226–231. http://www.aaai.org/Library/ KDD/1996/ kdd96-037.php
[2] UK Government. 2016. Driving licence holding and vehicle availability (NTS02). National Travel Survey.
[3] H2020 M2DC: https://m2dc.eu
[4] HEAPLab Politecnico di Milano: www.heaplab.deib.polimi.it

Acknowledgments

This work was supported in part by the European Community under grant agreement no. 688201 (M2DC – EU H2020 [3]).

INTelligent Embedded Systems Architectures and applications (INTESA) Workshop 2018 Oct 4
ESWEEK 2018 Sep 30 - Oct 5, Turing, ITALY