A novel Big Data analytics and intelligent technique to predict driver's intent

Lech Birek[1], Adam Grzywaczewski[2], Rahat Iqbal[3], Faiyaz Doctor[4], Victor Chang[5]

[1,2]**Jaguar Land Rover, UK**
[3]**Coventry University, UK**
[4]**University of Essex, UK**
[5]**Xi'an Jiaotong-Liverpool University, China**

Abstract: Modern age offers a great potential for automatically predicting the driver's intent through the increasing miniaturization of computing technologies, rapid advancements in communication technologies and continuous connectivity of heterogeneous smart objects. Inside the cabin and engine of modern cars, dedicated computer systems need to possess the ability to exploit the wealth of information generated by heterogeneous data sources with different contextual and conceptual representations. Processing and utilizing this diverse and voluminous data, involves many challenges concerning the design of the computational technique used to perform this task. In this paper, we investigate the various data sources available in the car and the surrounding environment, which can be utilized as inputs in order to predict driver's intent and behaviour. As part of investigating these potential data sources, we conducted experiments on e-calendars for a large number of employees, and have reviewed a number of available geo referencing systems. Through the results of a statistical analysis and by computing location recognition accuracy results, we explored in detail the potential utilization of calendar location data to detect the driver's intentions. In order to exploit the numerous diverse data inputs available in modern vehicles, we investigate the suitability of different Computational Intelligence (CI) techniques, and propose a novel fuzzy computational modelling methodology. Finally, we outline the impact of applying advanced CI and Big Data analytics techniques in modern vehicles on the driver and society in general, and discuss ethical and legal issues arising from the deployment of intelligent self-learning cars.

Keywords: driver's intent prediction, Big Data, Big Data Analytics, Computational Intelligence, e-calendar, geo referencing.

1. Introduction

The automotive industry is an ever-evolving competitive sector, which was propelled forward through human ingenuity and scientific advances. A comparison between today's vehicles with the ones that drove off the assembly line just 100 years ago would demonstrate that: today's cars last longer, are more fuel efficient, are able to provide their owners with a comfortable driving experience, while offering a variety of services through the integration of networking and ITC solutions. Vehicle manufacturing is one of the many industrial success stories of the 20th century and set to be one of the most rapidly evolving industries in the 21st century. The freedom offered by a car is now of paramount importance to modern citizens and the modern vehicle is emerging to be an extension of peoples' everyday living and working environments. This growing reliance and interdependency means that there is high demand for a wide range of quality services to be seamlessly offered by the vehicles. As it was stated by Jonathan Ive the head of design of Apple Computers "The car has become an article of dress without which we feel uncertain, unclad, and incomplete in the urban compound". The exponential growth of

miniaturization in computing technologies and the rapid advancements in communication technologies have led a variety of sensing equipment and interactive information systems to rapidly find their way into modern vehicles. This wealth of information offered through the invisible web of smart electronics, along with the opportunities offered by state of the art features installed in the cabins and powertrain systems of modern vehicles, deliver the promise of providing an enhanced driving experience with the help of specifically designed embedded computing applications. Modern research and every day experience, demonstrates that being in a good mood is one of the best precondition for safe driving, and that happy drivers produce fewer accidents [1]. On the contrary, when the car, or the environment causes the driver to feel negative emotions such as aggressiveness, anger and stress, this influences their concentration and skills, as for example preventing the driver from concentrating on the road ahead, which can consequently cause accidents, and life threatening situations [2].

In recent years, developing novel applications, which can promote driver's experience by effectively anticipating and reacting to their intent, have gained an immense boost from the Big Data and Internet of things revolution. Accumulating and utilizing data from the various data sources available in the modern hi-tech environment allows for gaining detailed insights on the drivers needs, preferences and habits, thus offering opportunities for the development of improved driver assistive systems. The importance of Big Data to this and wider application areas is highlighted in a recent statement by IBM's chief executive officer that "Big Data is the new oil" [3]. The analogy of data to the liquid that has fueled our cars and our lives in the previous century has been identified and highlighted by various studies demonstrating the power and impact of Big Data to modern economy and societies. As pointed out by Hashem et al. Big Data has three main characteristics. Firstly, the data itself is numerous and high dimensional. Secondly, it is not possible to categorize the data into regular relational databases, and finally data streams are created, captured, and analyzed rapidly [4]. As Gerhard mentions Big Data is a revolutionary leap forward from traditional analysis, which possesses three main characteristics volume, variety, and velocity [5]. Volume refers to the amount of data, which are created and stored. Variety is related to the various types of and sets of data collected, and velocity can be defined as the speed of data generation, streaming and aggregation [6]. In Kaissler et al.'s study, data value, and complexity are also proposed as Big Data characteristics. Data value is a measure of the usefulness of data in decision-making process, while complexity is a measure of the degree of interdependence and interconnectedness in Big Data structures. The inherent difficulty concerning the handling of these large amounts of data results in major challenges concerning its storage and analysis, as well as cost and time associated to the efficient delivery of results. Moreover, the results of this analysis should be delivered in an interpretable and easily visualized way [7]. Big Data analytics refers to the techniques utilized in order to examine and process the data so that hidden underlying patterns are revealed and interesting relations and other insights concerning the application context under investigation are exposed.

In modern times advances in software and hardware technologies provide a wealth of diverse data sources, which generate a wealth of Big Data information relating to the drivers intent and preferences, allowing for patterns of behavior of the driver to be revealed. It is common knowledge that most people use the same routes when commuting to work or other everyday locations (i.e. home, petrol station, supermarket etc.) at the same times of the day for the same days of the week. Similarly, although a person may shop on different days or at different times, they will often visit the same grocery store(s) [8]. Therefore, it is possible to predict a person's future preferences and actions if they are correlated with past behavior [9]. User profiling, and data provided by GPS, or geographical data accumulated by mobile devices can be utilized in order to reveal the driver's preference concerning specific roads and destinations thus allowing for the necessary adjustments to the vehicle's configurations. Data streams containing physiological signal information such as (EEG, HR, GSR) captured with the use of sensory equipment installed in the car, or from unobtrusive wearable sensors can provide insights of the driver's physical and mental state, which can aid the car in anticipating the driver's intent and

actions. Smart phones and modern mobile devices are another platform that poses embedded sensors (such as GPS, WiFi, accelerometers etc.) and can collaborate with on-board computer systems to offer valuable information in order to reveal the user's needs and behavioral patterns [10]. Video input through cameras inside the car monitoring the driver's eyes and upper body posture can provide video streams which can reveal different aspects of the driver such as their sleepiness or fatigue levels [11], while cameras outside the car can provide information concerning the position of the car on the lane thus contributing in revealing the driver's intents. For example, this information may be used to identify the driver's intent to stop the vehicle and take a break to rest, or their intentions to overtake a slow moving vehicle ahead. Audio equipment can be utilized for capturing audio signals relating to the driver's behavioral patterns or psychosomatic state. Previous research have detected yawning to monitor the driver's drowsiness levels [12] and other audio signals can be utilized so that the car can anticipate the driver's state of mind in context of their vehicle usage at a given point in the day, i.e. an intention for a relaxed drive home after a tiring day at work. A rich source of information to reveal driver's intent can be provided by the ever-expanding use of social networks. Before entering their cars, the drivers interact with social networks, where they post comments and status updates which if harvested can be utilized by the car to automatically detect their driver's desire to drive to a certain destination or to reveal their overall affective state. Personal data from electronic calendars and mailing lists is another information source, which can be used in order to provide insights of the driver's driving patterns. This can be combined with data concerning the volume of traffic, which can be utilized to inform the driver and allow the car to plan an optimal route to the desired location. This data can be provided from RSS feeds or other related publicly available news websites. The behavior and intent of the driver can also be related to the way they use their car. Controller Area Network (CAN bus) data captured by modern cars, relating to the actions of the driver, such as the amount of pressure applied in the acceleration or brake pedals, the angle rotation of the steering wheel, and others parameters, can also be utilized to discover behavior patterns and construct effective predictive models of the driver's response to different situations and their expectations from the vehicle.

Most of the data sources identified above are related to the ability of modern car to connect with other smart networks and devices and has led the automotive industry to a very important point of its history. Connectivity is not new to the car and it has been introduced as a premium feature to vehicles for years. It is only now that the world wide regulation for emergency call and stolen vehicle tracking is mandating connectivity across all new vehicles introducing the need for automotive companies to compete even more in this market. The smart phone revolution has also created a big market for infotainment systems, which rely on this connectivity infrastructure, to provide the same experience within vehicles. Vehicular Ad Hoc Networks (VANET) for example is a powerful technology which has emerged from the need to support the utilization of the various wireless products and devices in the car and it has been used by applications in intelligent transportation systems (ITS), surveillance systems, and safety alerts on the road [13,14]. The introduction of vehicle connectivity to the modern automotive industry creates vehicle specific opportunities. Current luxurious cars have significantly complex computer systems with an array of internal computer networks, embedded devices as well as hundreds of non-trivial sensors gathering gigabytes of data every minute. State of the art vehicles are being equipped with significant amounts of networking infrastructure to provide connectivity to other vehicles, road infrastructure, home networks, power grid data networks and end user mobile devices (Bluetooth, NFC, Wi-Fi and its derivatives). The self-learning car developed by Jaguar Land Rover is an example of a state of the art vehicle exploiting some of these capabilities [15]. This emerging technological environment creates a massive opportunity for vehicles to take full advantage of the concept of the Internet of Things by collecting, processing and aggregating this information and transforming it into knowledge required by their cars, other drivers, vehicles, and the wider society. For example having a distributed sensor network of cars being able to assess the density of traffic on the road can be used to optimize the traffic within a region, maximizing the road and public transport throughput.

For the advantages of Big Data and Internet of Things to be exploited there is an increasing need for implementing intelligent systems and computational techniques which can reduce the complexity and cognitive burden on accessing and processing the large volumes of data generated in both embedded hardware and software based data analytics [16,17]. Big challenges stem from the utilization of Big Data in real world problems such as the prediction of the driver's intent, since the implementation of real time applications is becoming increasingly complex. This complexity derives from a variety of data related factors. One factor is the high dimensionality degree which a dataset may possess thus increasing the difficulty of processing and analyzing the data. The interactions, co-relations and causal effects of these high dimensional data parameters in relation to the behaviors and specific outcomes of these systems are often too complex to be analyzed and understood by human users. Additionally data can be accumulated from diverse sources and input channels, making the online processing very demanding due to the variety of signal input which need to be synchronized and diverse data types which need to be analyzed simultaneously. Furthermore, the collected data is often comprised of multiple types of inputs, which are also not always precise or complete due to various sources of imprecision, and uncertainty paired with missing data (e.g. malfunctioning or inaccurate sensors). Moreover, there is an inherent need in applications existing in modern cars for high speed storage, processing of data and retrieval of the corresponding analysis results. Another factor that should be taken into account is that the method utilized for Big Data analytics should extract knowledge from data in an interpretable way. The computational techniques deployed to perform this task should make the underlying patterns, existing in the data, transparent to the person who tries to utilize and understand them. Finally, there is a need for techniques performing online adaptation, to incorporate contextual and user specific elements in their design, and decision making mechanism, in a user friendly and computationally feasible manner. All the above factors should be reflected in the computational and machine learning techniques utilized in order to process and analyze Big Data so that successful applications and models can be constructed [18].

This research paper is structured as follows. Initially in section 2 we identify various signals and inputs, which can be utilized by computer systems in order to automatically identify the driver's intent, and behavior. In section 3, we examine one of these data sources in detail. More specifically using a number of statistical measures and with the help of a classification performance analysis, we investigate the potential of utilizing e-calendar and geo-location information to identify the driver's location. Through this analysis, we highlight that e-calendar data can be used in order to reveal the driver's intention to travel to certain destinations. In section 4, we propose a novel fuzzy computational modelling methodology capable of handling the challenges of exploiting large amounts of data originating from multiple and diverse data sources. In section 5, we discuss the impact of this research, to the driver and the society, and discuss the legal and ethical challenges arising from the deployment of this kind of technologies. Finally, in section 6 we summarize and discuss our research findings.

2. Data sources for predicting driver's intentions and behaviour

The driver's intent predictive model is often built up through the analysis of the historical data (vehicle time and position). Research reports that the percentage of repeated trips increases over time of observation and reaches about 60% after 40 days [19]. Furthermore, only 10 routes on average per driver account for half of the trips taken. Though this means that a lot of the routes could be correctly predicted just by analyzing the historical data, there are still about 40% of these which deviate from the pattern and could benefit from additional sources of information. Authors of [20] report that approaches based only on a particular set of features fail to achieve peak performance. Instead, they proposed and evaluated an approach that combines different mobility features and treats the next place prediction task as a supervised learning problem.

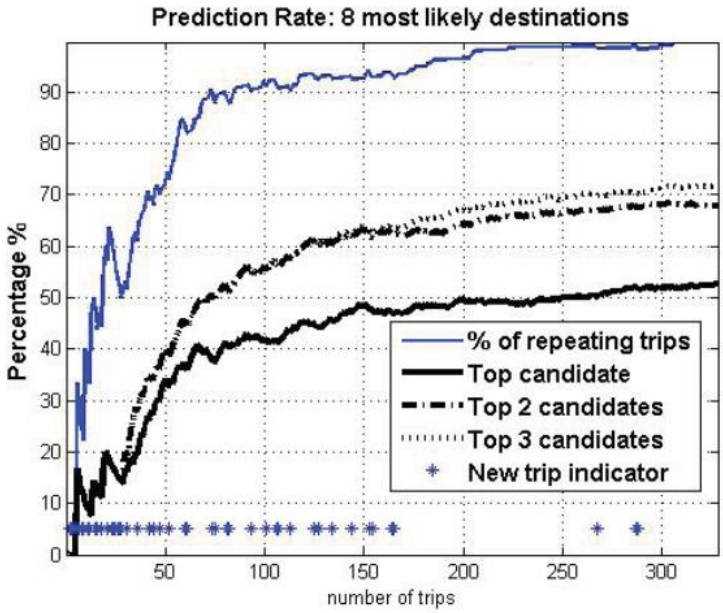

Figure 1. Prediction rate, Percentage of repeating trips and occurrences of new trips [21].

Figure 1 presents the prediction results obtained by applying the model from [21], where a generic method combining fuzzy and Markov models has been proposed. It can be clearly seen, that the prediction rate increases exponentially with the number of trips. We believe that including additional sources of information can improve the prediction during the period with a smaller number of trips. Research shows, that if properly implemented and used, groupware calendars can be a valuable source of information on the user's day plan. In [22], the author looked at the use of the on-line calendars in an organization of around 100,000 employees. The conclusion was that the on-line calendars are a massive self-logging system that has a great potential of being the organizational source of knowledge and learning about user mobility habits. It is important to note, that different systems implement different ways of sharing the calendar information. Users usually have the option to keep their details hidden from others, allowing one to only see if the person is busy or free (FreeBusy Reader mode). Another way is to allow open access, where the details about certain events including date and time, location, description and participants are visible to anyone in the company (Reader mode). The latter option, although causing some privacy issues, is considered to be a preferable solution. As the authors of [23] point out, the benefits of an open system became clear over time, where it filled a need to communicate one's whereabouts and availability. As we discovered through our analysis, most of the employees (95%) kept their calendars in the open reader mode, which provided us with plenty of data to analyze.

A number of papers looked into calendars as one of the sources used in the tasks related to time and location prediction. In [24], a system was proposed that estimated the number of people in front of an outdoor advertising billboard. The prediction was made based on the number of mobile phones near the billboard. Moreover, the system inferred people's preferences (e.g. do they like a specific ad shown on the billboard?), by combining location estimations from the mobile phones with listings of social events that are freely available online. Online calendars were used to mine the times and locations of the events happening in a particular area. The problem of unpredictability in daily movement patterns has been researched in [25]. It focused on the change in user's behavior and suggested a real time estimator for when unpredictability occurs (i.e. unusual day). The electronic calendar usage was one of the elements used to explore this problem. In [26,27] the times of visits to the place which matched start time of the calendar entries were used for semantic place prediction, based on an identified set of features. The possibility of exploiting the correlation between movement data and social interactions (which

can include calendar events) in order to improve the accuracy in forecasting of the future geographic position of a user was presented in [28]. Sharing information between the users is also at the core of research presented in [29], where the authors tackled the problem of sparse location datasets by using probabilistic graphical models. Sharing among the users is explored with the aim to learn the quantitative relationship between week-hours, which allows for accurate predictions on days with no observation history. None of those papers, however, look in detail at the content of the calendars to what extent the calendars are used and what kind of location based information is available.

Historical data can be enriched by other data sources in order to identify more effectively the intentions of the driver as it concerns their destination, preferable routes, or usage of in-car features. For example, driver's affective state can be utilized as an additional tool in order to recognize driver's intent and satisfy their preferences. Different affective and physical states may correlate more strongly with different destinations, and the car users may require the car to adjust its cabin or engine features to match their needs. It is logical for a driver to feel stressed on their way to an important business meeting, and happy on their way to the cinema to watch their favorite movie. Towards achieving the goal of recognizing the driver's state a variety of different inputs have been utilized. One of the most popular inputs is physiological signals captured by sensors. Electroencephalogram (EEG) and Electrooculography (EOG) recordings are physiological signals, which have been utilized by recent research to detect the driver's alertness levels. In the work by Chen et al. a system utilizing these signals was proposed, achieving high classification accuracy in a sensible computation time [30]. Other physiological signals which can be used to provide estimates on the driver's overall affective state, thus allowing an insight at the driver's future destination, preferences and needs, are the heart rate (HR), galvanic skin response (GSR), and skin temperature (ST). Aspects of the emotional state of individuals have been found to correlate strongly with these signals. More specifically in the work by Rainville et al. in 2006, it was demonstrated that HR increases when an individual is presented with a positive stimuli, thus relating the HR signal with the valence aspect of the person/driver's affective state [31]. The relation of an individual's valence (how positive or negative they feel) is also found to relate to changes in ST as it was shown in [32]. GSR signal on the other hand relates to another aspect of a person's affective state, namely arousal [33]. Arousal is defined as the levels of activation of an individual and is considered to be one of the key ingredients of emotional processes. These signals have been used in developing systems, which automatically detect the driver's affective state. A demonstrative example is the work by Nasoz et al. where a system responsible for identifying the driver's affective state from multimodal inputs including GSR, HR, and ST was developed. Their system was able to differentiate and recognize efficiently four driving related emotion categories (neutrality, panic/fear, frustration/anger, and boredom/sleepiness) [34].

In-vehicle computer systems can be used in order to capture a variety of signals related to the manner the driver is utilizing the car, and this data can provide valuable information for identifying the driver's behavior. These signals may include the pressure, which is applied on the acceleration and brake pedals, the angular velocity with which the driver is turning the wheel, frequency of gear shift, and well as other dynamic measurements of the drivers interactions with the vehicle. For example, as stated in ref. [35] the movement of the driver's foot before and after they press a pedal can facilitate an understanding their driving behavior, state, and style. In this study, the researchers have proposed and implemented a framework, which utilizes Hidden Markov Models to predict brake and acceleration pedal presses. They have achieve a notable performance which demonstrates the potential for developing driving assist, and behavior prediction systems based on these kinds of inputs. Another example is the work by Huang et al. where Support Vector Machine (SVM) was utilized in order to predict the driver's intentions of changing a lane. The researchers utilized a combination of inputs including steering angle, gas pedal pressure, velocity and acceleration of the vehicle, and eye movement as inputs to the developed system, achieving a high accuracy rate of 88.78% [26,27].

Video input is another data source that can be used to facilitate the automated prediction of the driver's intent and behavior. Video cameras have been utilized as an information source towards a variety of driver related predictive tasks. This includes the anticipation of driving maneuvers from the driver, intention of changing the lane, or adjustment to the position of the car in the lane, recognizing the driver's fatigue, and emotional state etc. In the work by Jain et al. [36], the researchers installed cameras inside to a car and through applying an HMM model they aimed to predict driving maneuvers before they actually occur. They have tested their method on a dataset containing hours of natural driving (both freeway and in-city) and they have achieved a notable accuracy at predicting driving maneuvers in real time, delivering the results a reasonable amount of time before these maneuvers happen, thus allowing the systems of a car to respond accordingly. Agrawal et al. utilized cameras to detect eye and lips motion, along with facial expressions related to four basic emotions, in order to propose a combined approach to deal with recognizing fatigue or bad mood of the driver. The recognition accuracy of their fuzzy rule based method that used both of these features achieved a very high accuracy of 94.58% [37]. Video input has also aided in providing environmental contextual information to the driver such as detecting traffic lights [38] or even perform rainy weather recognition [39]. Despite the vast spectrum of applications using video input, which can provide a wealth of information and improve the accuracy of a system's predictions pertaining to driver's current location, intended destinations, or facilitate in discovering the driver's needs and preferences at a specific point in time, there are some inherent difficulties concerning the heavy computational burden of storing and processing video input in real time.

The introduction of wireless technology in vehicular environments has set the stage for the development of applications that enable information sharing amongst moving vehicles thus promoting road safety and reducing the number of fatalities directly linked to car accidents. Such technology has been used in Mobile ad hoc networks (MANETs), and more specifically in the automotive area, in Vehicular ad hoc networks (VANETs), which utilize dedicated short range communication (DSRC) to facilitate communication between vehicles moving at close distances, and their surroundings. In the study by Al–Sultan et al. a novel unobtrusive system was developed, able to detect the driver’s behavior and communicate with nearby vehicles warning them of possible dangers and sharing the available information. Their system utilized a five layer context aware architecture which once having collected data from the driving environment, is able to distinguish between certain and uncertain information, and perform appropriate actions. The system integrates contextual information including the driver, the vehicle, and the surrounding environment in order to classify behavior using the labels: “normal, drunk, reckless or fatigue”. To achieve this goal the model utilizes a real-time probabilistic approach based on Dynamic Bayesian networks (DBNs). Moreover in terms of behavior recognition and classification, the suggested model deploys a dynamic approach, accounting for both the static and the temporal aspects of behavior thus constructing a more complete and accurate model. The work of Al-Sultan et al highlights the importance of employing both contextual, environmental and driver specific data in the behavior recognition process, in order to develop a robust and valid model [40].

Another interesting information platform where high volumes of data are generated is social networks. People’s engagement with social networks in modern societies has grown exponentially. Research has shown that especially young adults are most likely to incorporate social media in their lives. In the US, it is estimated that 90% of adults between the ages of 18-29 use social networks. Social networks enable communication across the globe, and provide a direct means for sharing thoughts, feelings and ideas. People and consequently drivers, generate vast amounts of data containing rich information through their posts, comments and status updates. This data is possible to be utilized towards predicting the driver's destination intentions or for automatically assessing their current state. For example a driver can reveal their intended destination, current location or preferable route in a Facebook post before the time they start their journey. Automatically recognizing this intention as for example relating certain words in Facebook posts to specific destinations (home, work etc.) and feeding these information to the

car, shall allow for the delivery of personalized and high quality services based on anticipating driver needs. The goal of delivery high end services can also be achieved through the utilization of social media data and Sentiment Analysis to reveal the driver's evaluations and feelings before commencing their journey. Sentiment analysis of social media data has been used efficiently as a tool for assessing the stance of people towards various services and products [41]. The effectiveness of sentiment analysis by utilizing social network data is highlighted in the work by Tan et al. Their experiments have demonstrated that through utilizing social media data classification accuracy of the user's sentiment significantly increases [42].

A demonstrative example of utilizing multiple data sources can be found in the patent by Jaguar Land Rover limited [43]. This system utilizes a combination of user data (e.g. GPS data from the car or a mobile device) and activity data (e.g. e-calendar data) to predict the next destination of the driver. The system processes the data and determines next non-routine events (e.g. a meeting which does not occur according to a regular pattern) and routine events (e.g. a regular start time of work in the office). An overview of the proposed system can be seen in figure 2.

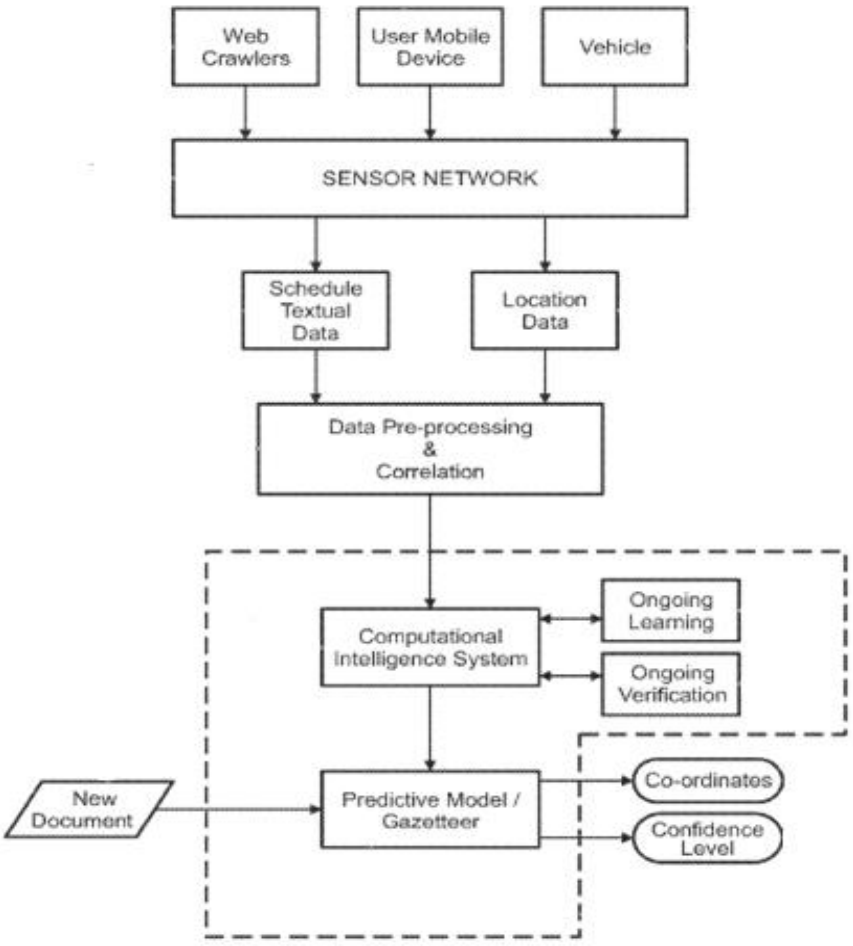

Figure2. Predicting destination based on user and activity data [43].

3. Preliminary Experiments on utilizing calendar data

The scope of our research involves looking into possible sources of information, which can be incorporated into a predictive model to increase the accuracy of predicting the driver's intent.

In order to achieve this, we investigated the groupware calendars used by Jaguar Land Rover employees. We address the question of whether the content of the calendar can provide meaningful information to help determine the time and location of the user. In order to achieve this, we conducted a statistical analysis to gain an insight on the general use of the e-calendars by the employees. Moreover, we reviewed and evaluated a number of available geo referencing systems. We explored the correlations between each location event identified previously from GPS data, with the overlapping calendar events. Finally, we explored the potential to learn the vocabulary used to describe a specific location event, and to be able to later identify that location based only on the description provided by the user in the calendar description, or content of their e-mails.

3.1 Basic statistics for electronic calendar utilization.

We conducted a statistical analysis on the general use of the electronic calendars by a sample group of users. Focus has been put on the analysis of 3 different fields: location, summary, and description. We analyzed these fields in terms of the number of unique and empty entries, and the most commonly used words and expressions. We looked into the time patterns of the events: the duration, and distribution during the day, and week. We then summarized our findings. Data was collected from 329 users who agreed to have their publicly available company calendars processed. The calendars have been downloaded and processed so that each event was a separate entry in a database. We downloaded the past year and the future month of events for this study, which resulted in an overall number of 574,614 events, out of which 77.87% were unique (as per event ID). The gathered statistical calendar data can be seen in Table 1.

Table 1. Calendar in numbers.

Data:	
Users	329
Events in all calendars	574,614
Number of unique events (based on event ID)	447,476
Average events / user	1746.55
Lowest number of events	147
Highest number of events	5,611
Number of unavailable events (hidden)	50,250
Percentage of conflicting events	35.90 (30.48*)%
Where (Location) field:	
Number of empty entries	22.93%
Number of unique non-empty entries	68.29%
Number of unique words. No stop words included, no stemming applied	10,200
Summary (Title) field:	
Number of empty entries	4.41%
Number of unique non-empty entries	86.85%
Description (Body) field:	
Number of empty entries	33.04%
Number of unique entries	58.18%

⋆ not including all day events

A number of statistical tests have been performed and the following results have been obtained:

Access mode

The Access Mode defines the way the users are sharing their calendars. In FreeBusy Reader mode other users can only view the periods when the owner of the calendar is busy (i.e. has a scheduled event), without having access to any details of the event (description, location, participant, etc.). The study shows that for the sample group only 5% of the events were shared

in the closed, FreeBusy mode. The other 95% had their calendars in Reader mode, which allows access to the details of the events.

Event duration

The majority of the events were shorter than 1 day, with only 7% (Figure 3) of them being scheduled as all day events. However, they are important as very often users use the all day events to schedule holidays or other activities which change the location context (i.e. visit to a company in India, holidays in USA, etc.).

Availability of information in each of the field (summary, description, location)

The analysis showed that majority of the fields were being filled, with only 38% being left empty in case of Description. The Summary field had the highest percentage of non-empty entries, with 91%. The most relevant, Location field had a fill rate of almost 60%, which favors this study as it is believed to contain the most useful location information.

Time span between occurrence of the event and creation of the last update

Pre-processing: Removing empty entries and not counting entries with negative duration entries (difference between creation and start of the event should not be negative where the same is also true for the difference between updated and start and created and updated). This eliminates what seems to be invalid entries, or some recurring events for which update time is set for every event of recurrence, even those that already occurred. In 12904 out of 16718 events (77.19%) the events had been updated. The majority of the events had their final update done within a day of the creation. The rest had final updates spanning from 1 month to over a year after the event has been created. This time period can be explained by the recurring events, for which the creation time was set when the recurrence started, but the final update reflects the date of update of the most recent event. The time difference between the start and the final update of the event has also been measured. Almost 33% of all events are finally updated no later than a day before the start, and more than 60% have their final update a week before the occurrence of an event. It is worth noting, that most updates in that period occur 10 and 70 minutes before the start of the event. This can cause issues for the predictive features, but only if the update details are related to time and location of the event. As for the time between the creation and start of the event, over 64% of them are created a week or more before their start. This is useful information as it indicates that in general users plan their work related activities well in advance.

Analysis of time and date of the events

This includes duration of the events, frequency of occurrence in different times of the day, days of the week and months of the year. The results show almost equal distribution of the events throughout the year (on monthly basis) and throughout the working days of week (Monday - Friday). The daily schedule shows a bimodal (double-peaked) distribution, which is characteristic for human related activities, with most of the events happening before noon. The events usually last no more than 2 hours (for almost 77% of events).

Calendar coverage

There is a difference on a per user basis as to how often the calendar is used. Some of the usage covers up to 75% of the working day, but others have as little as 6% of their time covered.

Dictionary analysis

The evolution of the word dictionary as the time progresses can be seen on figure 3. The 20 most commonly used words in summary location and description fields were calculated. The results for the location field can be found in figure 4.

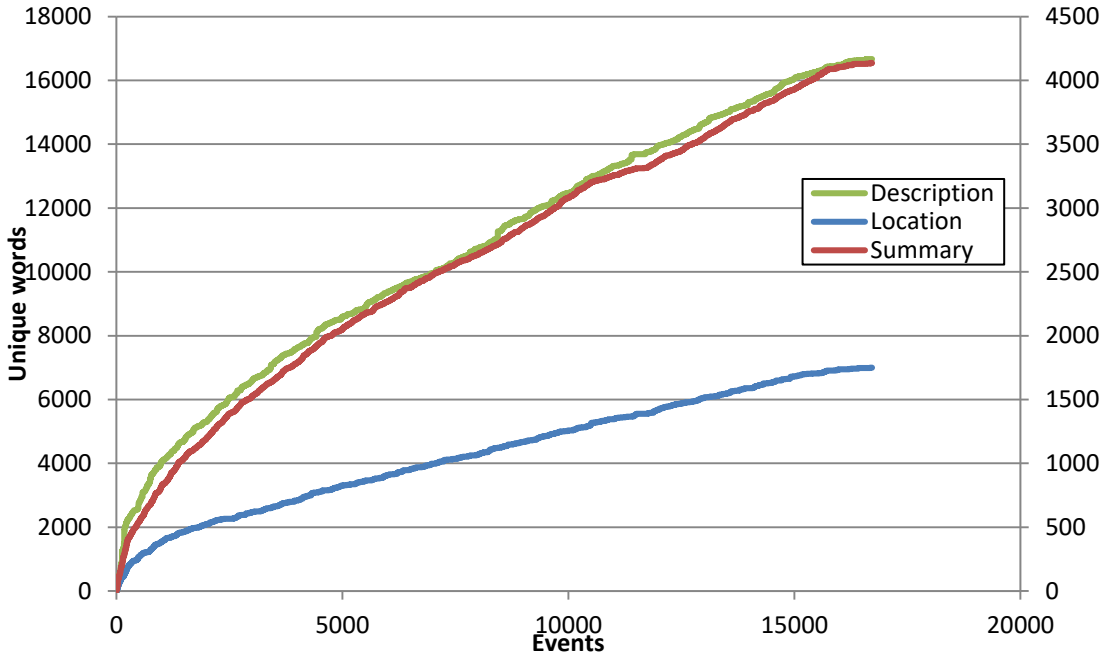

Figure 3. The number of unique words in the population dictionary against the increasing number of events. The events were grouped together from all available calendars and were sorted by date. The left Y-axis represents the number of unique words for the *Description* field, the right Y-axis represents the number of *Location* and *Summary* fields.

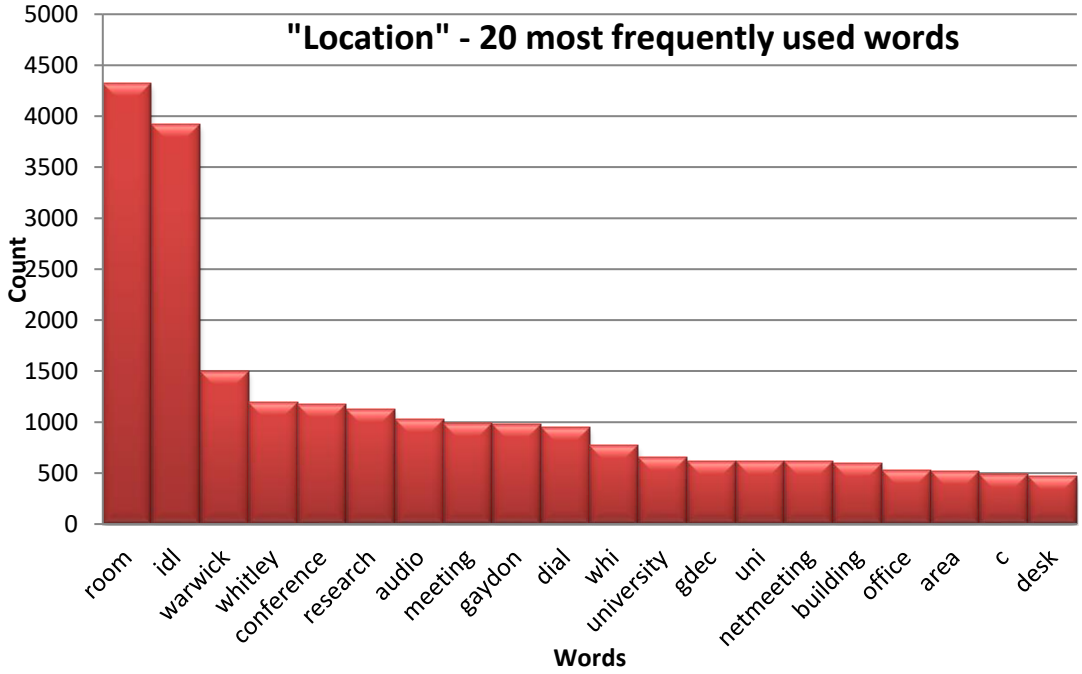

Figure 4. 20 most frequently used words in the *Location* field.

3.2 Evaluation of basic geo-referencing systems

3.2.1 Geo-referencing systems

Geo referencing systems usually comprise of two parts:

- Geo tagger (Geo parser) – processes the document in order to identify the entities which represent possible place names. This usually involves typical information retrieval and text processing steps, such as tokenization and sentence splitting, Part of Speech tagging, lemmatization and Named Entity Recognition in order to obtain the geo tagged file.
- Geo resolver (Geo coder) - once the document is geo tagged, i.e. the place names are identified, they are processed against the available Gazetteers – dictionaries of place name and location pairs. It is often the case that the same place name can point to different locations; therefore, the results are usually ranked based on some criterion, like the distance from the defined point, country, type of place, etc.

There are numerous existing geo referencing systems on the market, some of which will be briefly described below. The performance of those systems is evaluated over the test data described in the previous section. We focused our evaluation on the "Location" field. After removing duplicate, empty and unavailable entries, we were left with 1746 unique entries. We queried a number of available geo services with the location fields from the calendars. The entries were not pre-processed in any way, other than using lower cases and removing unnecessary spaces, tabs and new line signs. The average length of the location field in that form was 27.91 characters and it consisted of 4.77 tokens on average.

3.2.2 Overview of the tested services

Table 2. Basic information about the tested geo-referencing solutions.

Name of the provider	Name of the service	Limits of free queries	Comments
Yahoo!	PlaceFinder PlaceMaker Places	2000/day	
Google	Places	100/day	limited query number for free accounts
Edina Unlock	Places Text	unlimited	requires source of the text to be on the web
MapQuest	Nominatim	unlimited	can't handle longer queries

3.2.2.1 Yahoo! Geo Services

Yahoo! offers free, limited access to some of their geo services, such as PlaceMaker (PlaceSpotter in Yahoo! BOSS Geo) and PlaceFinder. The free access is limited to 2000 queries a day per app ID.

- **PlaceFinder** provides access to geocoder and reverse geocoder, and the database can be queried by providing the address, place name, or the coordinates. Tests proved that the service returns many results, but very often they are not the actual locations of the event. This is mainly due to company jargon, use of people's names, phone numbers and web addresses in the location field. For example, the use of abbreviations such as IDL, WHI is very often confused with airport codes.
- **PlaceMaker** acts as the geo-entity extractor. The service aims at identifying places in unstructured text data and returning geographic metadata. This solution seems to be more suitable for the purpose of analyzing the Location fields as it is expected that these fields will contain words and expressions which can distort normal geo coders, such as names and phone numbers. The results, however, were not that promising. Service had issues with some obvious name places, such as c*astle Bromwich*, or office – *gaydon*, which was correctly identified by PlaceFinder.
- **Places** was another service similar to PlaceFinder, but the number of returned results was much smaller.

Figure 5 presents the information on how many entries returned a place name, and also how many of those identified place names could point to more than one location.

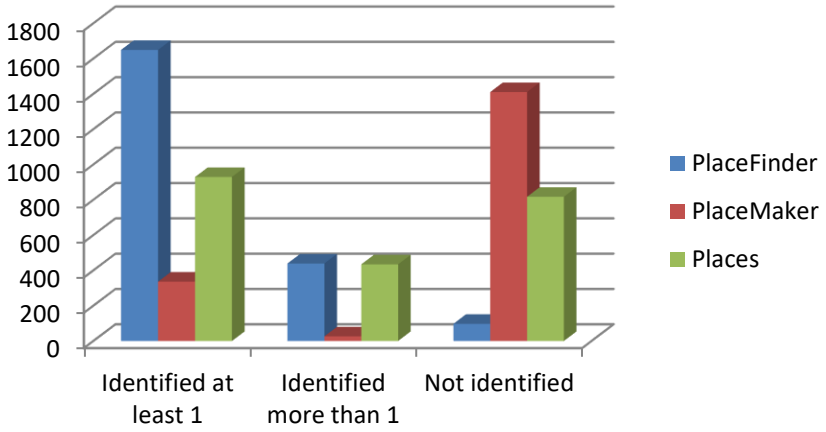

Figure 5. Results of the evaluation of Yahoo! geo services.

Additionally, the identified locations were put on the map to visualize the results of the search (see Figure 6). It can be seen that many of the places are positioned in the locations, which were very unlikely to be intended locations of the events (Africa, North America, etc.) given that the majority of the locations should be identified in the UK.

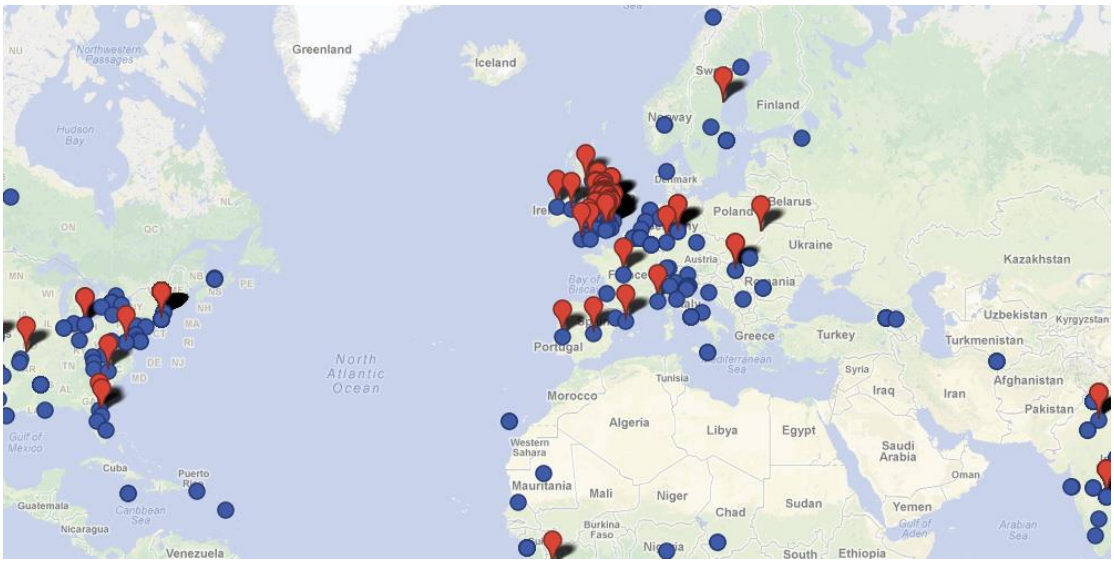

Figure 6. Map with identified places from Yahoo! Geo services.

3.2.2.2 Google Places

Google Places service provides information about places defined as establishments, geographic locations, or points of interest. Several requests are available, out of which the Text Search Request, is of our interest, as it returns list of places based on user's search string. The service responds with a list of places which match the string and any location bias that has been sent (for example “pub near Warwick University” will return a list of pubs around the location found by parsing “Warwick University”). The response includes geographical coordinates, names and addresses if available, along with the tags associated with the places. Unfortunately, the service could not be tested thoroughly as the free requests for Text Searches to Places were limited to

only 100 queries per day. Nevertheless, the results for 100 location entries are presented below in figure 7.

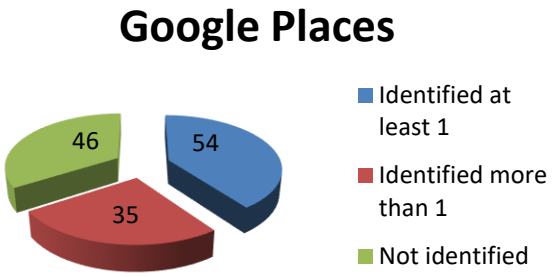

Figure 7. Results of the evaluation of Google Places service.

Although this service seemed to provide correct prediction, even for entries with abbreviations, such as "Warwick uni" or "Warwick idl", the process is still problematic with more complex queries (for example: "Warwick university idl - boardroom"). For efficient use, it would be necessary to pre-process the queries before submitting them in order to obtain more meaningful results.

3.2.2.3 Edina Unlock

Edina Unlock offers two services:

- Unlock Places, which is used to search across different sources of geographical data in order to find the locations of place names. It can be used to provide geographical information as points or larger boundaries. Although it can provide the information about bigger places around the world, more in depth knowledge of the UK places is available through the use of data from Ordnance Survey.
- Unlock Text is a geo parsing service which is able to search the text for references to geographical locations. It then uses Unlock Places to provide geographical information, and generates a set of geo referenced results.

The limitation of that service is that only the text available as a web resource can be processed. This causes additional issues, as the retrieved locations of the events have to be stored on the server. Additionally, after submitting the texts for processing, it was found that it took a great deal of time to obtain the results. In the end, Unlock Text had difficulties in identifying geographical locations which were easily identified by other services, probably due to lack of the use of capital letters. As a result this service was not able to identify any geographical location.

3.2.2.4 MapQuest

Another map service which has been briefly tested was MapQuest Nominatim Search Service, which is very similar to Yahoo! PlaceFinder Geo service, in the way it provides the results. The main issue with this service was that it could not handle longer queries though the results produced were similar to those returned by Yahoo.

3.3 Calendar and location correlation

In this section, we illustrate the correlations existing between geo-location data and the collected e-calendar data. In order to achieve this, each location event identified previously from GPS data, was correlated with the overlapping calendar events. For example, in figure 8,

it can be observed that location 25 is correlated to three calendar events (one-page presentation, preparation-security, and security briefing). Similarly, the user can utilize the software to see the correlation of any location, with the corresponding calendar events (Figure 8).

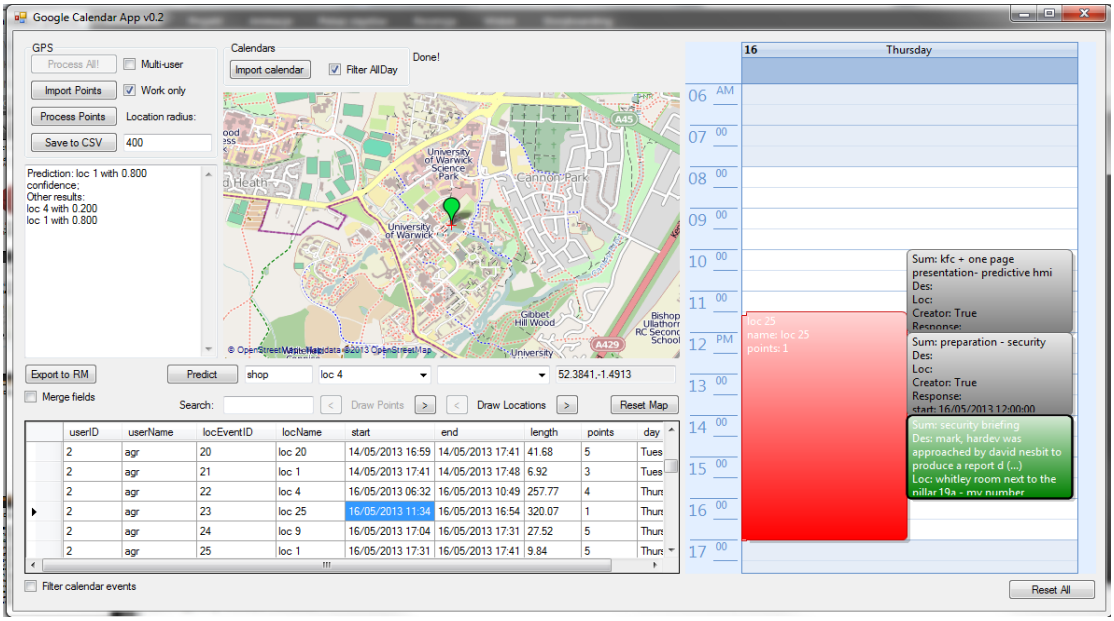

Figure 8. Correlated location (in red) and calendar events (in green and grey).

In order to use the correlated location, and for calendar events to learn the vocabulary, the location and calendar event pairs need to be specified. Here it may be difficult to establish which events are in fact the ones describing the identified location, due to conflicts and overlapping of calendar events (figure 9.E), and locations (figure 9.B).

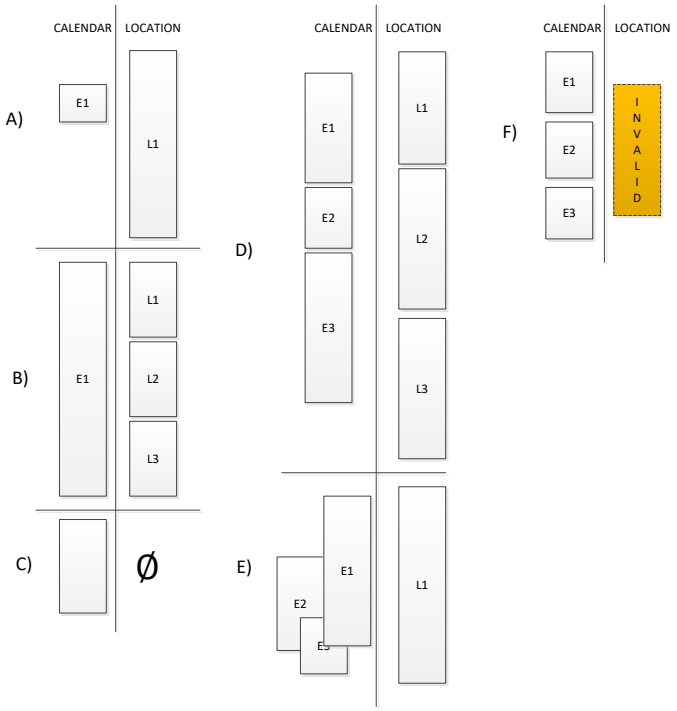

Figure 9. Various possibilities of conflicts between identified geographic locations and calendar events.

To resolve the conflict we filter the events based on:

- The content (or more precisely the lack of it) in the location, description and summary fields of the calendar event.
- If there are a number of events overlapping each other, we take the one that spans the longest over the duration of the location event.

The end results include an array of data samples which consist of the text included in the description of the calendar event with a classification label indicating the location of the user throughout the duration of the event.

3.4 Learning user's vocabulary for place name geo tagging and recognition

The aim of this step is to learn the vocabulary used to describe the location event, and to be able to later identify the location based only on the description provided by the user in the calendar description, or content of e-mails. In order to achieve this, with the help of an open source predictive analytics platform called Rapidminer we performed text preprocessing, and applied a number of machine learning techniques including K-nearest neighbor (K-NN), Naive Bayes, Decision and Random Trees and Support Vector Machine (SVM).

KNN is a one of the simplest machine learning algorithms. KNN is based on the idea that data samples, which are similar are in close proximity in the feature space. When a new data sample is presented, the classifier identifies the k-data samples, which are closest to the new sample. This can be achieved by calculating a measure of distance, such as the Euclidian distance. The new sample is then assigned to the class that matches the most frequent class of its k nearest neighbors. Simplest case is when k=1 meaning that the new sample is assigned to the class of its closest neighbor [44].

Naive Bayes are statistical classifiers, which are based on Bayes' Theorem. They are able to predict the probability that a data sample belongs to a class and the data sample is classified to the most likely class. Naive Bayes are simple classifiers that are shown to have a similar performance to more complicated techniques [45]. However, Naive Bayes classifiers make strong independence assumptions. More specifically, Naive Bayes classifiers make the conditional independence assumption, that the effect of a specific feature on a class is independent of the values of the other features.

A decision tree is a tree-structured representation of a decision procedure, which can be considered as a divide and conquer strategy to identify the class of an instance. Every node of the tree defines either a class name (leaf/answer node), or a test, which divides the space of instances based on the possible outcomes of the test in the node (non leaf/decision node). A new data sample is assigned to a class based on the values of its features, starting from the first root node, and then moving down the tree structure crossing the corresponding decision nodes, until it finally reaches a leaf node that contains the class label [46]. A Random Forest comprises of a number of simple decision trees, which are used to produce a final decision. When a new training is presented to a Random Forest, the most popular class among the classes calculated by every simple decision tree, is assigned to the sample. It is shown that by using random forests, prediction accuracy can be improved compared to simple decision trees, and that there is a significant improvement in the tendency of simple decision trees to over fit to their training set.

Support Vector Machine (SVM) is a very popular ML technique widely used by modern researchers in a variety of different application areas. SVM is based on the idea of "margin". A margin can be defined as the distance between the hyper-plane that separates the classes describing a problem, and the instances belonging to each class. The basic principal, which drives the computational mechanism of the SVM methodology is to maximize the "margin" [47].

The results obtained by applying the aforementioned basic ML algorithms can be seen in table 3. These initial results are promising and demonstrate that more research effort is required to exploit e-calendar data as a source for predicting the driver's future destination. It must be noted, that these results are computed based on only one aspect of predicting vehicle usage namely calendar and geo location data. Therefore, the application of simple ML techniques as the ones described above, to model this problem is justified. However, in order to exploit the multitude of data sources available to predict driver's intent and behavior, there is a need for more efficient ML techniques. In the following section, we propose a holistic fuzzy computational modeling methodology able to handle the volume of the data and the computational requirements associated with the automated prediction of the driver's intent and behavior.

Table 3. Results for the applied machine learning methods.

Algorithms	Parameters	Accuracy	W Mean Recall	W Mean Precision	Correlation	Margin	Kappa	g-mean
kNN	k=5, weights, cosine similarity	53.50%	5.28%	4.62%	0.156	0.000	0.122	4.94%
	k=4	55.00%	5.38%	4.64%	0.155	0.000	0.137	5.00%
Naïve Bayes	greedy, min band 0.1, ker. 10	54.00%	4.94%	4.76%	0.091	0.000	0.128	4.85%
Decision Tree	cryt. - acc, rest default	60.50%	4.55%	2.79%	0.036	0.001	0.018	3.56%
Random Forest	cryt. - acc, rest default	60.50%	4.55%	2.75%	0.000	0.000	0.000	3.54%
SVM	default	60.50%	4.55%	2.75%	0.000	0.000	0.000	3.54%

4. Computational Intelligence for predicting driver's intent.

In order for diverse data sources, such as the calendar data described in the section 3, and sources discussed in section 2, to be exploited for predicting driver's intent, there is a need for robust and flexible machine learning approaches, which are able to deal with the computational requirements of processing large amounts of diverse data. Machine learning (ML) approaches offer a means for modelling patterns and correlations in data in order to discover relationships and make predictions based on unseen events. ML approaches consist of supervised learning (learning from labeled data), unsupervised learning (discovering hidden patterns in data or extracting features), and reinforcement learning (goal oriented learning in dynamic situations) [48]. As such ML approaches can also be categorized into regression techniques, clustering approaches, density estimation methods, and dimensionality reduction approaches. Non-exhaustive examples of these approaches are Decision tree learning, Associate rule learning, support vector machines, and Bayesian networks. Computational Intelligence (CI) is a subclass of ML approaches where algorithms have been devised to imitate human information processing and reasoning mechanisms for processing complex and uncertain data sources. CI techniques form a set of nature inspired computational methodologies and techniques which have been developed to address complex real-world data driven problems for which mathematical and traditional modelling are unable to work due to: high complexity, uncertainty, and stochastic nature of processes. Fuzzy Logic (FL), Evolutionary Algorithms (EA) and Artificial Neural Networks (ANN) form the triad of core CI approaches that have been developed to handle this growing class of real-world problems. FL is an established methodology to deal with imprecise and uncertain data [49]. FL provides an approach for approximate reasoning and modelling of qualitative data and adaptive control [50,51] based on the use of linguistic quantifiers (fuzzy sets) for representing uncertain real-word, data and user

defined concepts and human interpretable fuzzy rules that can be used for inference and decision making [52]. EAs are based on the process of natural selection for modelling stochastic systems [53] and approaches such as genetic algorithms, genetic programming, and swarm intelligence optimization algorithms [54,55,56] can be used for optimizing complex real-world systems and processes. Finally ANNs enable feature extraction and learning from experiential data [57], and are based on imitating the parallel processing and data representation structure of neurons, in animal, and human brains. CI techniques have been used successfully in the past to tackle the challenges of dealing with high volumes of data and modelling complex real-world problems. A demonstrative example is the work by Maniak et al., where the researchers proposed a novel biologically inspired approach called Hierarchical Spatial-Temporal State Machine (HSTSM). The research team's approach relied on a combination of soft-computing techniques such as: deep belief networks, auto-encoders, agglomerative hierarchical clustering and temporal sequence processing [58] The approach was designed to mimic the structure and functions of the mammalian brain based on a theory proposed by Jeff Hawkins [59,60]. From the team's results, it was demonstrated that the suggested CI methodology was able to handle high volumes of data characterized by complex correlations between input values and temporal consequences between different input states of the system.

In order to predict driver's intent we propose a novel fuzzy computational modelling methodology. The proposed approach is based on a hybrid method, incorporating a combination of popular Computational Intelligence techniques such as adaptive Fuzzy Logic systems and genetic algorithms. Our approach is able to handle the high volume and complexity of data associated with the driver's intent and propose intervention strategies applied to the configurations of the car, to offer an optimized and personalized driving experience. Our approach consists of the following steps:

The first step towards the construction of the computational model responsible to model driver's behavior and predict their intentions and future destinations is to identify and collect all available and relevant data sources. As described in the previous section, this data could be obtained from various data sources, such as historical data of previous destinations and routes, physiological signal data from EEG, EOG, HR, GSR, ST sensors, video and audio input, data provided by the car such as pedal pressure, social media data and others.

Following data collection is applying a suitable feature selection technique. Feature selection plays a vital role in our methodology for several reasons. A model comprising of a variety of inputs results in a very large feature space, adding to the complexity of the model and obstructing its applicability in a real setting. Additionally, by removing redundant and insignificant features from the data set, the correlations existing between the diverse inputs applied in this problem, become more obvious, thus resulting into a more transparent and comprehensible model. In order to achieve this, we propose the feature extraction method proposed by Luukka et al. [61]. This method relies on fuzzy entropy measures, and it was tested successfully with four large medical datasets. As it was demonstrated by the team's results, the suggested method managed to improve the classification accuracy, while using a significantly smaller set of features.

Following this step, data clustering is applied on the derived feature space in order to extract clusters representing different possible states of the modeled environment. We propose the use of the popular Fuzzy-C means algorithm [62]. Fuzzy C-means partitions a finite collection of elements into a collection of fuzzy clusters. In contrast to hard clustering, fuzzy clustering enables the elements under investigation to belong to more than one cluster with a given degree of membership. This membership value illustrates the strength of the relation between an element and a specific cluster. The primary goal of this clustering step is to define a finite number of modelled states, which are more probable to occur, thus resulting in a reduction of the input space.

Temporal sequence learning is then utilized to train a model, in order to predict the next state of the modelled environment based on the temporal relations between the modelled input states. Through this process, the relations among input variables and different states of the behavior of the driver can be revealed, thus enabling the discovery of hidden patterns in the data pertaining to the driver's behavior or their intended destinations.

A fuzzy control system would be responsible for associating, the modeled states relating to future destinations, selected routes, or the driver's cognitive/affective state, to the adjustment of in-vehicle features in a manner that influences the driver in a positive way. The driver themselves could facilitate the construction of fuzzy rule base of the system by providing their opinion concerning actions that they would expect their car to make given certain scenarios. The fuzzy rule base could be further enriched by taking into account expert's advice or existing literature describing the optimal actions to be performed in the cabin and power train system of the car, given the context the individual is driving in.

Personalization of the fuzzy control system could be enhanced and optimized, by enabling the adaptation of fuzzy rules comprising the rule base of the aforementioned fuzzy control system. This can be achieved by utilizing fuzzy adaptation methods similar to the ones described in the work by Karyotis et al. and Doctor et al. [63,64]. By applying the aforementioned adaptation techniques, whenever the driver performs an action inside the car, which contradicts to the action decided and performed by the fuzzy control system, then changes would be applied to the fuzzy rules that contributed to the output of the system in that particular situation.

With the help of a genetic algorithm, the fuzzy sets and rules could be optimized in order to enhance the performance and effectiveness of the fuzzy control system. Optimization of fuzzy systems by utilizing GA is a proven technique. For example, Bernardo et al. utilized a genetic algorithm to optimize the internal parameters of a fuzzy logic based system for financial prediction, while in the work by Karyotis et al., the researchers optimized a fuzzy rule base describing a computational model of emotion [65,66].

Data fusion can be achieved by using the approach developed by Chang [67] and Sun et al. [68] that different data sources can be put into the integrated MapReduce framework, processed, analysed and returned with the final output. They have explained how to achieve fusion with their system architecture to allow concurrent computation and optimization. Figure 10 shows the percentage of data fusion success against the data size. These were close to 100% success when the size of data was 5 GB. While the size of the data increased up to 40 GB, the percentage of data fusion success dropped but it was still maintained above 92%. The success rate dropped when the data size increased.

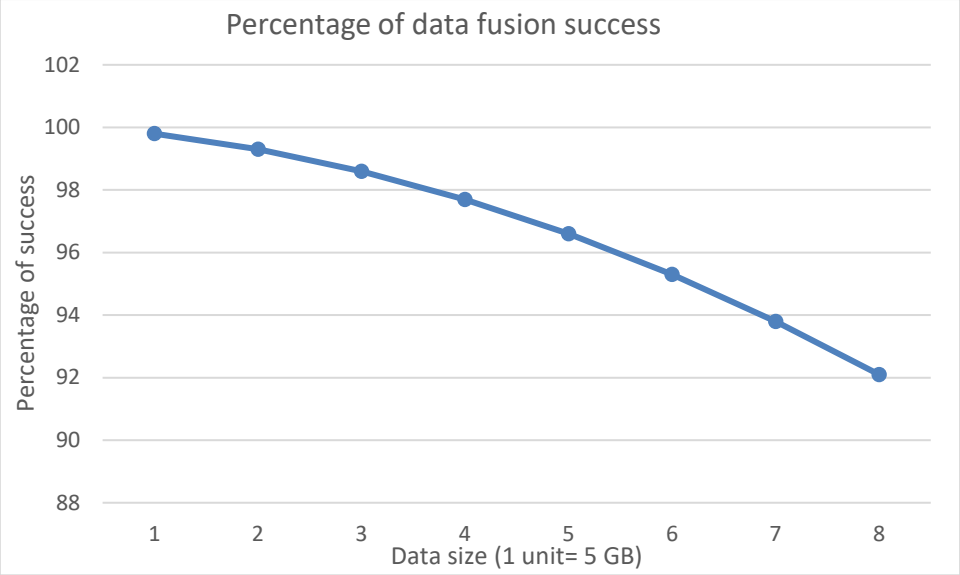

Figure 10: Percentage of data fusion success between 5GB and 40 GB

The proposed service will be evaluated by the application of a requirement-based evaluation framework, specifically designed to statistically assess the prediction capabilities of the system, together with its usability aspects. Standard techniques, such as randomly partitioning of the data into two independent sets will be used to provide a good estimate of the system's forecasting accuracy (training set (e.g. 80% of the data) for forecasting model construction, and test set (e.g. 20% of the data), for accuracy estimation of the proposed system). Special care will be provided in constructing a user-friendly system, with advanced visualization features that will be easy to use by a novice user. In order to achieve this goal, the proposed computational models will be evaluated by academics, industrial experts and future users. This evaluation will include acceptance testing of the new technology by the industry, usability testing by the average user, and investigation of its applicability in real world contexts. This process will identify potential implementation problems, and prepare suitable mitigation actions such as the modification of the initial design, to overcome those problems. Given the commercial potential of the computational models discussed in this paper, the authors will explore wider application of system in different case studies, and identify application areas, where the new computational mechanisms can be used effectively.

The proposed approach has significant advantages compared to other research efforts to predict the driver's intentions. First advantage is the use of multiple data sources. Contrary to other efforts that utilize a limited number of inputs, the presented framework takes into account and exploits data originated from a combination of diverse sources such as video, audio sensors, calendar data, historical data, and data provided by the car itself (e.g. pedal pressure). This way based on more information and aspects of the behavior of the driver, a more holistic view of the driver's intent can be constructed. Second advantage is the potential of the methodology to process Big Data and produce accurate forecasting results in real time. This is achieved by the proposed intelligent feature selection method that has a proven record for reducing effectively the input space [61]. The spatial and temporal modeling of the proposed approach also consists of a methodology that has provided state of the art classification results and has the ability to deal with data of high volume and complexity [58]. Third advantage is the ability of the proposed approach to deal with the inherent uncertainty concerning the behavior of the driver that relates to the nature of the human behavior. Fuzzy Logic systems are shown to be able to account and model this uncertainty very effectively [63]. Finally, as shown in previous research the proposed framework allows for easy adaptation and optimization of the constructed model in order to be tailored to the personal characteristics of a specific driver, or the needs of a situation and achieve improved classification accuracy [66]. Considering the aforementioned advantages, and the classification results achieved by utilizing only calendar data and basic ML

techniques in Figure 11, we can conclude that the proposed framework is a very promising approach for forecasting the driver's intent.

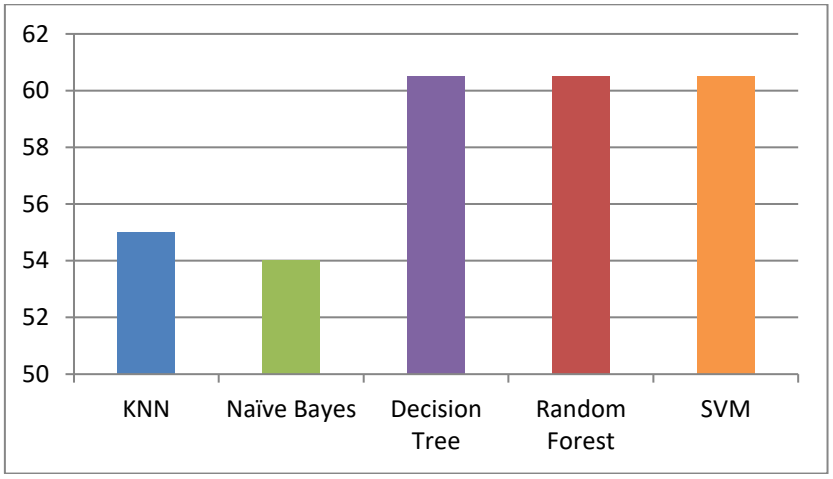

Figure 11. Results for basic ML methods based on calendar data

Blending methods in Figure 11 will be a step forward, since it can improve the accuracy and strengthen our research contributions. We followed a related work by [69], in which execution time and the number of vehicles were measured. Although four algorithms were developed, only MixGroup and DGD algorithms were used. MixGroup algorithm combined all and took the best one. DGD algorithm observed all, picked the best performing one in training data, and then asked all simulated cars to follow that. Computationally, simulations to demonstrate different locations to predict driver's intents can be used. We put 60 simulated cars under the prescribed scenario to predict each driver's intents and compare the expected and actual percentage of accuracy. We can use recall, precision and F-measure to determine the accuracy. Each time 10 simulated cars were added, until it reached 0 simulated cars. All results were taken three times to get the mean values for accuracy. Figure 12 shows that DGD algorithm has a better accuracy, staying 75% for up to 40 simulated cars and then 70% between 50 and 60 simulated cars. MixedGroup has 70% accuracy for up to 40 simulated cars and then 65% between 50 and 60 simulated cars.

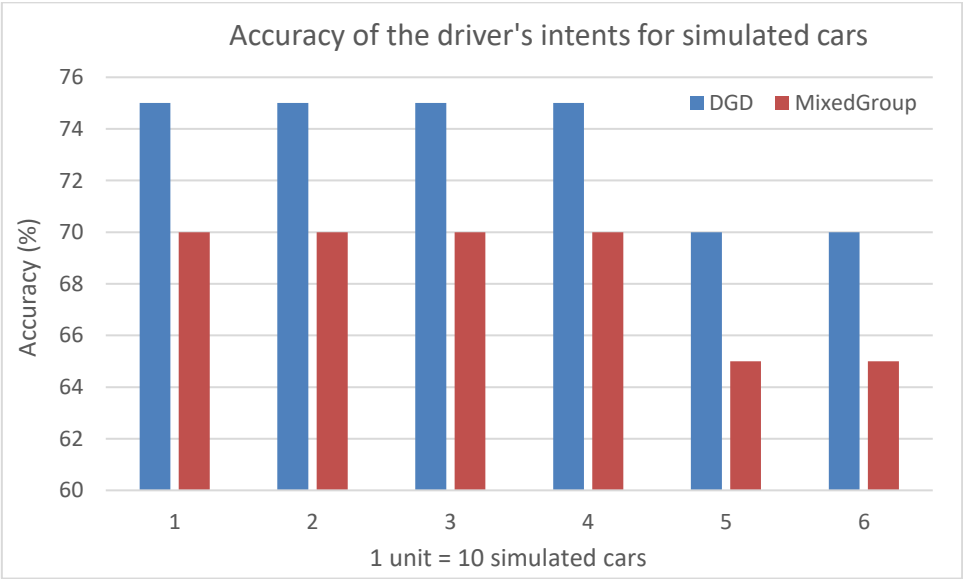

Figure 12: Accuracy of the driver's intents for up to 60 simulated cars

5. Discussion

In this research, we proposed the utilization of e-calendar data, in combination with data obtained from various heterogeneous sources, which are available in abundance in the modern environment, to automatically predict driver's intent and behavior, through the deployment of an advanced fuzzy computational modelling technique. This methodological framework bears a significant impact on drivers as individuals and on societies that choose to accommodate this kind of technologically enhanced services. At the same time, there are critical ethical and legal issues arising from the application of smart vehicles that possess the ability to collect and analyze large amounts of user specific data to provide intelligent decisions. These issues are related to the application of advanced artificial intelligence, and to the manipulation of personal sensitive data. In this section, we outline the benefits of our research to the driver and society and discuss the ethical and legal issues related to the application of this technology.

The first to benefit from the deployment of "deep learning" cars that are able to predict their driver's intent and behavior, is the driver themselves. Anticipating the driver's intent to drive to a certain destination enables the car to perform adjustments to the cabin, and the configuration of the power train systems, to match the needs of a particular route, and the preferences of the driver. Therefore, the driver's emotional wellbeing can be promoted by actions such as: creating a friendly cabin-environment with the help of the car's infotainment systems and climate control, to help the driver enjoy a relaxing drive, after a stressful day at work; creating a personalized environment in the cabin, or through changing the configuration of the engine to the terrain of an intended route or satisfy the driver's personal style; learning the entertainment preferences of the driver and the passengers and responding accordingly; providing reminders and alternate routes, and traffic alerts to enable the driver to arrive in time for appointments in different parts of the city; notifying the appropriate recipients in case of an unexpected delay; suggesting refueling options for reducing carbon emissions, or for avoiding stressful situations; predicting who the driver might want to call next from previous call patterns etc. This type of actions would allow cars to become facilitators of modern and dynamic individuals, by enabling them to fulfill their goals, improving driving performance and utility. Another aspect of driving which is greatly enhanced by deploying smart car technologies is driver's safety. Driver's safety can also be promoted, by equipping the car with the tools to exploit data sources, such as physiological, sensory, or video input. By utilizing physiological signals, or video input capturing the face and upper body of the driver, the car shall be able to automatically recognize the driver’s fatigue, and drowsiness levels, and could deliver feedback to alert the driver. This feedback may include visual or audio delivered cues for taking a break through the vehicles infotainment system, such as changing the radio station or audio track, or performing other actions aiding the driver to remain safe, and drive more mindfully.

Modern societies that support the development, and deployment of smart cars, may harvest a large number of advantages related to these technologies, which span over a broad spectrum, from promoting the environment to reinforcing the economy. Firstly, enabling the automatic prediction of the driver's intent and behavior, shall boost the development of a large number of services and products, thus creating a large and vibrant market. These products and services could be relevant to the recognition aspect itself (creation of improved and unobtrusive sensors, or development of software), or they could be utilized for exploiting the predicted driver's preferences, through the delivery of high-end feedback (e.g. digital dashboards with extended capabilities). Applying the proposed framework for predicting the driver's intent lead us to realizing the vision of smart cities. An intelligent car, able to predict the driver's intended destination and routes, and then communicate this knowledge to networks of interconnected cars or government authorities, has the potential to facilitate the smooth operation of traffic within smart cities. One way this can be achieved is through supporting decisions and actions aimed to manage traffic, and public transport in the city, based on the provided predicted routes

and destinations of smart vehicles. Another aspect that can benefit from the deployment of “deep learning cars” is the management of various city resources. One example is the improvement that can be achieved at managing the available and allocated parking spaces in a busy city center. Another example is the optimization of the power grid management to support electric vehicle charging. Effective power grid management will also influence the environment in a positive way since it promotes the use of electric and hybrid vehicles. The environment will also benefit from the presented framework since it allows for the optimization of engine configurations in modern vehicles in order to reduce CO2 emissions. It is important to note that under a broad context these technologies will also propel the creation of innovative research aiming to contribute to predicting driver's intent, or exploiting this prediction. This could benefit scientific disciplines such as mechanical and electrical engineering, pattern recognition, signal processing, and other specialized application domains.

A self-learning car would be responsible to make decisions to promote the driver's safety and satisfaction by relying on various factors. Based on these factors the algorithm would infer the driver's behavior. It is possible, in cases of machine learning techniques such as deep learning networks or genetic algorithms, to be unable to fully understand the reasoning behind the machine's decision [70]. How is the AI judging the driver? Are we certain that this judgment is not affected by factors like race or religion simply based on found correlations? Therefore, there is a need for transparent computational models, where the factors contributing to the car's decisions could be clear to an external observer. When the machine is imitating humans in making complex decisions, which integrate social or personal dimensions it needs to inherit the social requirements as well. Moreover, the transparency of the computational mechanisms of the machine learning technique used should also be accompanied by robustness of the developed AI system. The AI algorithms should be resilient against ill-intended and malicious manipulations. There are also significant issues concerning assuming responsibility in the face of accidents, or other unfortunate events, resulting from the application of intelligent technologies like self-learning cars. For example, it is possible that an accident is caused, when the self-learning car decides to perform a change at the cabin or the engine's configuration. In this case who is to blame for the accident, the engineers developing the hardware or the software of the system, the driver, or the system itself? The aforementioned moral and ethical issues are demonstrated by the unfortunate events of the 7th of May 2016, when a Model S Tesla vehicle, while in autopilot mode, crashed in Williston Florida causing the death of its driver. The car’s sensory system failed to recognize a truck's white trailer because of a very bright sky. In such kind of cases, where a self-driving car is involved, things become even more complex. One can imagine a case, when an instant decision should be made by the artificial intelligence of the car concerning the safety of its driver against the safety of the driver/passengers of another vehicle or pedestrians, or even against the machine's own existence. Nevertheless, despite the aforementioned issues, there is also a very optimist and positive ethical side at the creation of intelligent cars. Self-learning cars are not only tools to get to a destination but they could also become companions for humans in modern society. This vision is demonstrated in the statement by Isaac Asimov in the book I, Robot for robots: "There was a time when humanity faced the universe alone and without a friend. Now he has creatures to help him; stronger creatures than himself, more faithful, more useful, and absolutely devoted to him. Mankind is no longer alone" [71].

Ethical and legal issues arise from the way driver-specific data are collected, and communicated in the context of an interconnected self-learning car. This data, as outlined earlier in this paper, may include sensitive personal information such as location history, social media interactions, physiological signal information relevant to the driver's health and affective state, and others. These data is utilized by the car itself to predict the driver's intentions, but it could be also communicated to and used by organizations, governments or other intelligent cars which are connected to the original vehicle, in several ways. Research has shown that people are not always aware that Big Data is being gathered, processed, and shared with other parties through their interaction with intelligent devices, such as a self-learning car. Moreover, the existing

privacy frameworks are insufficient especially when it comes to secondary use of this kind of data [72]. In order to address these issues, the corresponding legal frameworks must be designed to protect personal information such as information concerning religion, ethnicity, political opinions, personal health of the driver. Personal data collected by self-learning cars must be treated lawfully towards specific goals, which are clear to the recipients and must be secured, stored in safe data centers or cloud infrastructures, where the collected data should not be kept longer than necessary. These principles are reflected to the Data Protection Act, which controls the way personal information, similar to the ones collected by self-learning cars, should be used by organizations, businesses, or the government [73]. It is imperative, that intelligent technologies respect the privacy of individuals, and legal frameworks ensure sensitive data protection, otherwise the reality vividly described in the work by Warren et al. by the statement: "what is whispered in the closet shall be proclaimed from the house-tops" [74], may become a harsh reality.

6. Conclusions

In this paper, we investigated the challenge of predicting the driver's intentions and behaviors under the scope of the emerging Big Data and Internet of Things technological revolutions. This challenge can be addressed through the technologically enhanced and interconnected environment, both inside and outside modern vehicles, which provide multiple data sources containing a wealth of information. We have identified and highlighted a variety of these data sources, focusing on the area of predicting the driver's intended destination, behavior, affective/cognitive state and preferences. Part of our research scope was to explore the potential of different input sources to be utilized through applying a proposed computational methodology. Towards achieving this goal, we investigated in detail one of those sources namely: e-calendars. E-calendars were explored as a data source, which could aid in predicting the next destination of a driver. Towards this goal, we analyzed data from the groupware calendars used by Jaguar Land Rover employees that provided a number of interesting statistical facts concerning general e-calendar usage. Though the use of different basic machine learning techniques we also investigated the potential to accurately recognize geographic locations with the help of the vocabulary used to describe the location of events taken from e-calendar entries. The results from the analysis are promising and combined with the fact that there is a lack of solutions to cope with the use of group, and user specific language for location identification and geotagging, they demonstrate that more research effort is required in this area. Summarizing on these results we can conclude that utilizing e-calendar information may result in improving the accuracy of predicting the driver's next destination. We also perform experiments for the percentage of data fusion success and all results stayed above 92% for up to 40 GB of data. We combined all the methods and developed two algorithms from a related work. The combined accuracy test for the driver’s intent from 60 simulated cars were performed, and stayed 70% and 65% for up to 60 simulated cars by two algorithms. Results also support our analysis.

E-calendar data when combined with the variety of diverse data sources available in modern vehicles can further enhance our ability to predict drivers’ intentions and preferences. To achieve this, we proposed and presented a state of the art fuzzy computational modeling technique. The aims are to integrate multiple data sources, such as the e-calendar data, and exploit the advantages of the increased connectivity aspects of modern age, which enable vehicles to have access to rich information sources such as social networks, other intelligent cars etc. This technique is designed to address the challenges arising from analyzing and processing large volumes of these diverse data parameters, in order to efficiently address the problem of automatically recognizing the driver's intention and its consequences on improving vehicle utility. The proposed framework has the potential to lead to applications with a significant impact to the society, economy and the individual driver. This can be achieved

through developing applications, which take advantage of the predicted driver's intent, the features and capabilities of the modern vehicles, and the ability of the car to communicate with other external, digital entities. These applications promise to revolutionize the way drivers interact with their vehicles, optimize the vehicle overall performance, even contribute to an improved management of traffic, and energy resources inside a smart city. These can be achieved through features such as: pre-emptive satellite navigation configuration, optimization of electric power train operation in hybrid vehicles (to reduce CO2 emissions), pre-emptive vehicle preconditioning (heating/cooling/de-icing), electric vehicle charging optimization and effective power grid management, driver recognition and vehicle personalization, information and multimedia personalization and contextualization. However, the deployment of super intelligent technologies poses serious ethical and legal challenges, which should be addressed in order for these technologies to provide better services to their users and a higher level of human machine interaction, instead of ultimately ending up with a high tech dystopia.

References

[1] L. James, D. Nahl, Road Rage and Agressive Driving, Prometheus Books, Amherst, N.Y., U.S.A, 2000.

[2] E. Wells-Parker, J. Ceminsky, V. Hallberg, R.W. Snow, G. Dunaway, S. Guiling, M. Williams, B. Anderson, An exploratory study of the relationship between road rage and crash experience in a representative sample of US drivers, Accid. Anal.Prev. 34 (3) (2002) 271–278.

[3] D.D. Hirsch, Glass house effect: big data, the new oil, and the power of analogy, Maine Law Rev. 66 (2013) 373.

[4] I.A.T. Hashem, I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Gani, S.U. Khan, The rise of big data on cloud computing: review and open research issues, Inform. Syst. 47 (2015) 98–115.

[5] B. Gerhardt, K. Griffin, R. Klemann, Unlocking Value in the Fragmented World of Big Data Analytics, Cisco Internet Business Solutions Group, 2012 June http://www.cisco.com/web/about/ac79/docs/sp/Information-Infomediaries.pdf.

[6] S. Kaisler, F. Armour, J.A. Espinosa, W. Money, Big Data: issues and challenges moving forward, 46th Hawaii International Conference on System Sciences (HICSS), IEEE, January 2013, 2013, pp. 995–1004.

[7] S. Sagiroglu, D. Sinanc, Big Data: a review, 2013 International Conference on Collaboration Technologies and Systems (CTS), IEEE, May, 2013, pp. 42–47.

[8] R. Simmons, B. Browning, Y. Zhang, V. Sadekar, Learning to predict driver route and destination intent, IEEE Intelligent Transportation Systems Conference, 2006. ITSC'06, IEEE, September, 2006, pp. 127–132.

[9] A. Grzywaczewski, R. Iqbal, Task-specific information retrieval systems for software engineers, J. Comput. Syst. Sci. 78 (4) (2012) 1204–1218 Elsevier.

[10] M. Reininger, S. Miller, Y. Zhuang, J. Cappos, A first look at vehicle data collection via smartphone sensors, 2015 IEEE Sensors Applications Symposium (SAS), IEEE, April, 2015, pp. 1–6.

[11] X. Wang, C. Xu, Driver drowsiness detection based on non-intrusive metrics considering individual specifics, Accid. Anal. Prev. (2015).

[12] S. Abtahi, B. Hariri, S. Shirmohammadi, Driver drowsiness monitoring based on yawning detection, 2011 IEEE Instrumentation and Measurement Technology Conference (I2MTC), IEEE, May, 2011, pp. 1–4.

[13] N. Kumar, R. Iqbal, S. Misra, J.J. Rodrigues, An intelligent approach for building a secure decentralized public key infrastructure in VANET, J. Comput. Syst. Sci. 81 (6) (2015) 1042–1058.

[14] N. Kumar, R. Iqbal, S. Misra, J.J. Rodrigues, Bayesian coalition game for contentionaware reliable data forwarding in vehicular mobile cloud, Future Gener. Comput. Syst. 48 (2015) 60–72.

[15] http://www.landrover.com/experiences/news/self-driving-car.html. As accessed online on 24 May 2016.

[16] R. Iqbal, F. Doctor, B. More, S. Mahmud, U. Yousuf, Big data analytics: computational intelligence techniques and application areas, Techniol. Forecasting Social Change (2018).

[17] T. Maniak, C. Jayne, R. Iqbal, F. Doctor, Automated intelligent system for sound signalling device quality assurance, Inform. Sci. 294 (2015) 600–611.

[18] S. Suthaharan, Big data classification: problems and challenges in network intrusion prediction with machine learning, ACM Sigmet. Perform. Eval. Rev. 41 (4) (2014) 70–73.
[19] J. Froehlich, J. Krumm, Route Prediction from Trip Observations, SAE SP, 2008.
[20] S. Noulas, N. Lathia Scellato, C. Mascolo, Mining user mobility features for next place prediction in location-based services, 2012 IEEE 12th International Conference on Data Mining (2012) 1038–1043 Noulas, 2012.
[21] D. Filev, F. Tseng, J. Kristinsson, R. McGee, Contextual on-board learning and prediction of vehicle destinations, 2011 IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems (CIVTS) Proceedings (2011) 87–91.
[22] J. Grudin, A case study of calendar use in an organization, ACM SIGOIS Bull. 17 (no. 3) (1996) 49–51.
[23] L. Palen, Social, individual and technological issues for groupware calendar systems, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems the CHI Is the Limit – CHI '99 (1999) 17–24.
[24] D. Quercia, G. Di Lorenzo, F. Calabrese, C. Ratti, Mobile phones and outdoor advertising: measurable advertising, IEEE Pervasive Comput. 10 (2) (2011) 28–36.
[25] J. McInerney, S. Stein, A. Rogers, N. Jennings, Exploring Periods of Low Predictability in Daily Life Mobility, Nokia Mobile Data Challenge, 2012.
[26] C. Huang, Mining Users Behavior and Environment for Semantic Place Prediction vol. 2012, Nokia Mobile Data Challenge, 2012.
[27] X. Huang, Driver Lane Change Intention Recognition by Using Entropy Based Fusion Techniques and Support Vector Machine Learning Strategy, M.S., Thesis, Computer System Engineering Dept., Northeastern University, Boston, Massachusetts, 2012 December.
[28] M. De Domenico, A. Lima, M. Musolesi, Interdependence and Predictability of Human Mobility and Social Interactions vol. 2012, (2012) arXiv preprint arXiv:1210.2376, p. 21, October.
[29] B. Kapicioglu, Place models for sparse location prediction, 7th Annual Machine Learning Symposium (2012).
[30] L.L. Chen, Y. Zhao, J. Zhang, J.Z. Zou, Automatic detection of alertness/drowsiness from physiological signals using wavelet-based nonlinear features and machine learning, Expert Syst. Appl. 42 (21) (2015) 7344–7355.
[31] P. Rainville, A. Bechara, N. Naqvi, A.R. Damasio, Basic emotions are associated with distinct patterns of cardiorespiratory activity, Int. J. Psychophysiol. 61 (1) (2006) 5–18.
[32] R.A. McFarland, Relationship of skin temperature changes to the emotions accompanying music, Biofeedback Self-regul. 10 (3) (1985) 255–267.
[33] M.E. Dawson, A.M. Schell, D.L. Filion, The electrodermal system, in: J.T. Cacioppo, L.G. Tassinary, G. Berntson (Eds.), Handbook of Psychophysiology, Cambridge University Press, Cambridge, 2007.
[34] F. Nasoz, C. Lisetti, A. A.Vasilakos, Affectively intelligent and adaptive car interfaces, Inform. Sci. 180 (2010) 3817–3836.
[35] C. Tran, A. Doshi, M.M. Trivedi, Modeling and prediction of driver behavior by foot gesture analysis, Comput. Vis. Image Understand. 116 (3) (2012) 435–445.
[36] A. Jain, H.S. Koppula, B. Raghavan, S. Soh, A. Saxena, Car that knows before you do: anticipating maneuvers via learning temporal driving models, Proceedings of the IEEE International Conference on Computer Vision (2015) 3182–3190.
[37] U. Agrawal, S. Giripunje, P. Bajaj, Emotion and gesture recognition with soft computing tool for drivers assistance system in human centered transportation, 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, October, 2013, pp. 4612–4616.
[38] M. Diaz-Cabrera, P. Cerri, P. Medici, Robust real-time traffic light detection and distance estimation using a single camera, Expert Syst. Appl. 42 (8) (2015) 3911–3923.
[39] H. Kurihata, T. Takahashi, I. Ide, Y. Mekada, H. Murase, Y. Tamatsu, T. Miyahara, Rainy weather recognition from in-vehicle camera images for driver assistance, Proceedings of IEEE Intelligent Vehicles Symposium, 2005, IEEE, June, 2005, pp. 205–210.
[40] S. Al-Sultan, A.H. Al-Bayatti, H. Zedan, Context-aware driver behavior detection system in intelligent transportation systems, IEEE Trans. Veh. Technol. 62 (9) (2013) 4264–4275.
[41] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inform. Retriev. 2 (1–2) (2008) 1–135 (1).
[42] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, P. Li, User-level sentiment analysis incorporating social networks, Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM August, 2011, pp.1397–1405.
[43] Jaguar Land Rover Limited, User Content Analysis. GB 2522733, (2015).

[44] H. Bhavsar, A. Ganatra, A comparative study of training algorithms for supervised machine learning, Int. J. Soft Comput. Eng. 2 (4) (2012) 2231–2307.
[45] I. Rish, An empirical study of the naive Bayes classifier, IBM New York, August, IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol. 3, 2001, pp. 41–46 (no. 22).
[46] P.E. Utgoff, Incremental induction of decision trees, Mach. Learn. 4 (2) (1989) 161–186.
[47] S.B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised Machine Learning: A Review of Classification Techniques, (2007).
[48] T.M. Mitchell, Machine Learning, (1997) Boston, et al..
[49] L.A. Zadeh, Fuzzy sets, Inform. Control 8 (3) (1965) 338–353.
[50] F. Doctor, C.H. Syue, Y.X. Liu, J.S. Shieh, R. Iqbal, Type-2 fuzzy sets applied to multivariable self-organizing fuzzy logic controllers for regulating anesthesia, Appl.Soft Comput. 38 (2016) 872–889.
[51] Y.X. Liu, F. Doctor, S.Z. Fan, J.S. Shieh, Performance analysis of extracted rule-base multivariable type-2 self-organizing fuzzy logic controller applied to anesthesia, BioMed Res. Int. (2014).
[52] S. Mahmud, R. Iqbal, F. Doctor, Cloud enabled data analytics and visualization framework for health-shocks prediction, Future Gener. Comput. Syst. (2015).
[53] D. Whitley, An overview of evolutionary algorithms: practical issues and common pitfalls, Inform. Softw. Technol. 43 (14) (2001) 817–831.
[54] F. Dreier, Genetic Algorithm Tutorial, (2002).
[55] R. Poli, W.B. Langdon, N.F. McPhee, J.R. Koza, A Field Guide to Genetic Programming, (2008).
[56] R.S. Parpinelli, H.S. Lopes, New inspirations in swarm intelligence: a survey, Int. J. Bio-Inspired Comput. 3 (1) (2011) 1–16.
[57] S.S. Haykin, Neural Networks and Learning Machines vol. 3, Pearson Education, Upper Saddle River, 2009.
[58] T. Maniak, R. Iqbal, F. Doctor, A Method for Monitoring the Operational State of a System, (2016) Interactive Coventry Ltd Pending Patent Reference No: 1607820.6.
[59] J. Hawkins, On Intelligence, Times Books, New York, 2004.
[60] J. Hawkins, D. George, Hierarchical Temporal Memory: Concepts, Theory, and Terminology, (2006).
[61] P. Luukka, Feature selection using fuzzy entropy measures with similarity classifier, Expert Syst. Appl. 38 (4) (2011) 4600–4607.
[62] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorythms, New York, Plenum, 1981.
[63] C. Karyotis, F. Doctor, R. Iqbal, A. James, An intelligent framework for monitoring students affective trajectories using adaptive fuzzy systems, IEEE International Conference on Fuzzy Systems, 2–5 August, 2015, Istanbul, Turkey, 2015.
[64] F. Doctor, V. Hagras, V. Callaghan, A fuzzy embedded agent-based approach for realizing ambient intelligence in intelligent inhabited environments, IEE Trans. Syst. Man Cybern.–Part A: Syst. Hum. 35 (1) (2005) 55–56.
[65] D. Bernardo, H. Hagras, E. Tsang, A Genetic Type-2 fuzzy logic based system for financial applications modelling and prediction, 2013 IEEE International Conference on Fuzzy Systems (FUZZIEEE), IEEE, July, 2013, pp. 1–8.
[66] C. Karyotis, F. Doctor, R. Iqbal, A. James, V. Chang, A Fuzzy Modelling Approach of Emotion for Affective Computing Systems, (2016).
[67] V. Chang, Computational intelligence for medical imaging simulations, J. Med. Syst. 42 (1) (2018) 10.
[68] G. Sun, V. Chang, G. Yang, D. Liao, The cost-efficient deployment of replica servers in virtual content distribution networks for data fusion, Inform. Sci. 432 (2018) 495–515.
[69] D. Liao, H. Li, G. Sun, M. Zhang, V. Chang, Location and trajectory privacy preservation in 5G-Enabled vehicle social network services, Journal of Network and Computer Applications (2018).
[70] N. Bostrom, E. Yudkowsky, The ethics of artificial intelligence, The Cambridge Handbook of Artificial Intelligence, (2014), pp. 316–334.
[71] I.I. Asimov, Robot, Oxford University Press, 2000.
[72] N.M. Richards, J.H. King, Big data ethics, Wake For. Law Rev. 49 (2014) 393.
[73] Gov.uk/data-protection/the-data-protection-act.
[74] S. Warren, L. Brandeis, The right to privacy, Harv. Law Rev. IV (5) (1890).