



H2020 - INDUSTRIAL LEADERSHIP - Leadership in enabling and industrial technologies - Information and Communication Technologies (ICT)

ICT-11-2017 Collective Awareness Platforms for Sustainability and Social Innovation – Innovation Action (IA)



CHILD
RESCUE

Collective Awareness Platform for Missing Children Investigation and Rescue

D2.2 Multi-source Analytics Methodological Foundations

Workpackage: WP2 – Grassroot Collective Intelligence in the Missing Children Investigation

Authors: S5, SLG, SoC, NTUA

Status: Final

Date: 31/10/2018

Version: 1.00

Classification: Public

Disclaimer:











The ChildRescue project is co-funded by the Horizon 2020 Programme of the European Union. This document reflects only authors' views. The EC is not liable for any use that may be done of the information contained therein.

ChildRescue Project Profile

Grant Agreement No.: 780938

Acronym:	ChildRescue
Title:	Collective Awareness Platform for Missing Children Investigation and Rescue
URL:	http://www.childrescue.eu
Start Date:	01/01/2018
Duration:	36 months

Partners

	National Technical University of Athens (NTUA), Decision Support Systems Laboratory, DSSLab <u>Co-ordinator</u>	Greece
	European Federation for Missing and Sexually Exploited Children AISBL - Missing Children Europe (MCE)	Belgium
	The Smile of the Child (SoC)	Greece
	Foundation for Missing and Sexually Exploited Children – (Child Focus)	Belgium
	Hellenic Red Cross (REDCROSS)	Greece
	Frankfurt University of Applied Sciences (FRA-UAS)	Germany
	SINGULARLOGIC ANONYMI ETAIREIA PLIROFORIAKON SYSTIMATON KAI EFARMOGON PLIROFORIKIS (SLG)	Greece
	Ubitech Limited (UBITECH)	Cyprus
	MADE Group (MADE)	Greece
	SUITE5 DATA INTELLIGENCE SOLUTIONS LIMITED (S5)	Cyprus

Document History

Version	Date	Author (Partner)	Remarks
0.10	28/05/2018	Minas Pertselakis, Konstantinos Tsatsakis (S5)	ToC
0.20	29/08/2018	Minas Pertselakis, Konstantinos Tsatsakis, Ioanna Michael (S5)	First Draft
0.40	12/10/2018	Nefeli Bountouni (SLG)	SLG contribution integration
0.50	23/10/2018	Minas Pertselakis, Anastasios Tsitsanis (S5)	Final Draft
0.60	25/10/2018	Nefeli Bountouni (SLG)	Internal Review
0.70	25/10/2018	Antonia Tsirigoti, Nikos Yannopoulos (SoC)	Internal Review
0.80	30/10/2018	Minas Pertselakis, Anastasios Tsitsanis (S5)	Final version
0.90	31/10/2018	Dimitris Varoutas (NTUA)	Quality Review
1.00	31/10/2018	Christos Ntanos (NTUA)	Final Review

Executive Summary

Deliverable D2.2 is part of the ChildRescue WP2 entitled "Grassroot Collective Intelligence in the Missing Children Investigation" and presents the results of "Task 2.2: Multi-source Analytics for Missing children Investigation".

The purpose of Task 2.2 is to study in depth the state-of-play regarding the most important research domains related to ChildRescue Data Analytics and propose, based upon this knowledge, the methodological foundations to be adopted by the project's data processing engine. The ChildRescue methodology should be able to deal successfully with various heterogeneous data sources and assist in the whole ChildRescue workflow process by making useful predictions and evidence-based recommendations.

In this context, D2.2 presents a sound methodological approach revolving around three axes: the profiling of human behaviours and activities, the evaluation and validation of evidence, and the estimation of routes and POIs based on human mobility patterns. The recommended methods and techniques are derived and scientifically supported by a thorough examination of the related literature, which consists of three major domains of multi-source data analytics: Profiling of human behaviour, spatiotemporal data analysis and social media analytics. The analytical review of these domains results in a number of methods and algorithms that deal with behavioural patterns and can eventually be applied to all ChildRescue aspects that require data analytics.

Naturally, anything related to human behaviour modelling suffers from complexity, heterogeneity and ethical issues. All of these challenges and the conclusions drawn are discussed in depth, while the perspectives for ChildRescue are outlined, considering the next steps of development and implementation.

The present deliverable will be merged with the results of D2.1 "Profiling Methodological Foundations" and D2.3 "Privacy and Anonymisation", in order to form the deliverable D2.4 - "Profiling, Analytics and Privacy Methodological Foundations, Release I". All the work conducted in WP2, along with any updates in methodologies, will be collected and presented in D2.5 "Profiling, Analytics and Privacy Methodological Foundations, Release II", which is the final deliverable for WP2.

Table of Contents

1	Introduction	8
1.1	Purpose & Scope	8
1.2	Structure of the Deliverable	8
1.3	Relation to other WPs & Tasks	9
2	Landscape Analysis	11
2.1	Computational Learning in Human Profiling	13
2.1.1	Research Literature Study.....	14
2.1.2	Key points extracted from the Literature Analysis.....	15
2.1.3	Challenges and Perspectives.....	17
2.2	Exploiting Spatiotemporal Data coming from Multiple Sources.....	18
2.2.1	Research Literature Study.....	19
2.2.2	Key points extracted from the Literature Analysis.....	20
2.2.3	Challenges and Perspectives.....	23
2.3	Social Media Analytics	24
2.3.1	Research Literature Study.....	25
2.3.2	Key points extracted from the Literature Analysis.....	25
2.3.3	Challenges and Perspectives.....	29
2.4	Discussion and key-takeaways.....	30
3	Methodological Approach for Multi-source Analytics in ChildRescue.....	31
3.1	Predictions based on Behavioural and Activity Profile Data	36
3.1.1	Possible sources.....	37
3.1.2	Methods and algorithms	38
3.2	Evidence Analysis and Evaluation.....	39
3.2.1	Possible sources.....	40
3.2.2	Methods and algorithms	40
3.3	Real time Route/Destination Estimation.....	42
3.3.1	Possible sources.....	43
3.3.2	Methods and algorithms	44
3.4	Discussion & Limitations.....	45
4	Conclusions & Next Steps	48
	Annex I: References	49
	Annex II: Research Literature	56
II.1	Computational Learning in Human Profiling	56

II.2 Exploiting Spatiotemporal Data coming from multiple sources 61
II.3 Social Media Analytics 67
Annex III: Past cases list of reference 74

List of Figures

Figure 1-1 Relation to other WPs/tasks.....	10
Figure 2-1 Main categories of Data Analytics techniques.....	11
Figure 2-2 General methodology for predictive analytics in social media	27
Figure 3-1 The proposed preliminary Data Model for ChildRescue	35
Figure 3-2 A simple data analytics flow.....	36
Figure 3-3 Behavioural Prediction process.....	37
Figure 3-4 Evidence Evaluation process	39
Figure 3-5 Route estimation process	43

List of Tables

Table 2-1 List of computational learning algorithms for human profiling	16
Table 2-2 Variables based on information in police files describing missing persons [33]	17
Table 2-3 Methods and Algorithms for spatiotemporal data analysis	21
Table 2-4 Most widely used data sources in spatiotemporal data analysis	22
Table 2-5 Methods and Algorithms for social media analytics	27
Table 2-6 Most widely used data types in social media analytics	28
Table 3-1 List of data fields included in the Profiling Template	32
Table 3-2 List of data fields for the Events Template	34
Table 3-3 Relation of ChildRescue investigation phases and Analytics types	36
Table 3-4 Data sources to be used for profiling	38
Table 3-5 Summary of algorithms for profile modelling and predictions	38
Table 3-6 Data sources to be used for evidence evaluation.....	40
Table 3-7 Algorithms related to Evaluation Steps	42
Table 3-8 Data sources that will be used as input for the recommended methods.....	43
Table 3-9 Methods and algorithms for route/destination estimation.....	44
Table 3-10 Summary of data collection and analysis limitations.....	47

1 Introduction

1.1 Purpose & Scope

A primary goal of ChildRescue is to assist decision-making in the complex missing children investigation process by exploiting a plethora of sources: testimonials from relatives and other acquaintances, the experience and scientific knowledge of psychologists and social workers, but also significant details and insights from social media, as well as complementary bits of information from open and linked data. The analysis of these elements will compose a much more detailed profile of the missing or unaccompanied child, which can then be compared to past cases profiles in order to infer useful correlations and predict possible behaviours.

The purpose of this deliverable, entitled "D2.2 – Multi-source Analytics Methodological Foundations", is:

- to investigate the state-of-the-art landscape in respect to predictive analytics and relevant domains of interest,
- to study the available sources of data employed in literature, and
- to recommend a well-founded methodology to cope with the multi-source data analytics dimension of ChildRescue.
- The suggested approach should also take into account the availability of sources and datasets from the pilot partners of the consortium, and prepare the conditions required for their analysis.

The present deliverable is released under the framework of Work Package 2 "Grassroot Collective Intelligence in the Missing Children Investigation", and reports the work conducted in Task 2.2 "Multi-source Analytics for Missing Children Investigation". It contains the 1st iteration of the data analytics methodology, which will be further refined and optimised in the following months relying on the feedback from the end-users, as well as the technical specifications of the ChildRescue platform. The finalised proposed methodology will be delivered at the end of the second year of the project [M24] in D2.5 "Profiling, Analytics and Privacy Methodological Foundations, Release II".

1.2 Structure of the Deliverable

The deliverable is comprised of the following sections:

Section 1 provides an introductory description of the document, presenting its purpose, structure and relation to other work packages.

Section 2 contains the landscape analysis on three broad areas of research, namely human behaviour profiling, multi-source geospatial data exploitation and social media analytics. Key points, comparison of different methods, challenges, as well as possible applications and perspectives are discussed in this section. An analytical presentation of all studied articles can be found in the relevant tables in ANNEX II, for easier examination and collation.

Following the literature overview, Section 3 focuses on building a sound methodological approach for multi-source analytics related to the ChildRescue objectives. The proposed methodology consists of a primal data model and three main application fields, each one supported by a set of relevant data sources and methods for their exploitation, state-of-art algorithms and techniques. The key-takeaways, along with possible challenges and risks, are discussed at the end of the section.

Section 4 wraps up the most important findings of this deliverable and outlines a plan for the next steps.

1.3 Relation to other WPs & Tasks

The reporting document at hand is the result of the first iteration of Task 2.2 - "Multi-source Analytics for Missing Children Investigation". It explores the vast field of data analytics using multiple heterogeneous sources and proposes a methodology to deal with them in the most appropriate and sound manner.

In order to achieve this, Task 2.2 builds on top of Task 1.1 "User Requirements" and Task 1.3 "ChildRescue Integrated Methodology, Release I". Furthermore, Task 2.2 receives crucial input from the research conducted in D2.1 - "Profiling Methodological Foundations" which provides definite directions on the nature and importance of profile data. At the same time, Task 2.2 is affected by the results reported in D2.3 "Privacy and Anonymisation" where several technical, as well as ethical aspects should be considered carefully in order to preserve the privacy and security of the data. The outcome of Task 2.2, reported in this deliverable (D2.2), will be merged with the results of D2.1 and D2.3, in the context of D2.4 - "Profiling, Analytics and Privacy Methodological Foundations, Release I".

Task 2.2 constitutes the scientific algorithmic background regarding multi-source analytics and provides the necessary input to WP3 - "ChildRescue Platform Architecture Definition and Implementation". The methodological approach proposed in D2.2, along with the user requirements and integrated methodology from WP1, will be translated into the ChildRescue architecture of T3.1 "ChildRescue Architecture and Platform Design", which will be then implemented in T3.2 and T3.3. As a next step, the technical verification in T3.4 will provide the necessary feedback on the performance of the suggested algorithms.

The overall approach, initially designed in D2.2 and later on updated in D2.4, will also be tested by the pilots through the evaluation framework of WP4. After completion of the first platform implementation iteration and the pilot evaluation, the algorithms and methodology will be revised accordingly. These updates will be part of D2.5 - "Profiling, Analytics and Privacy Methodological Foundations, Release II" [M24], which concludes the iterations and covers the final methodological approach of WP2.

Lastly, D2.2 will be publicly available after delivery, through the ChildRescue site and other project communication channels, as determined in the dissemination plan and the data management plan of WP5.

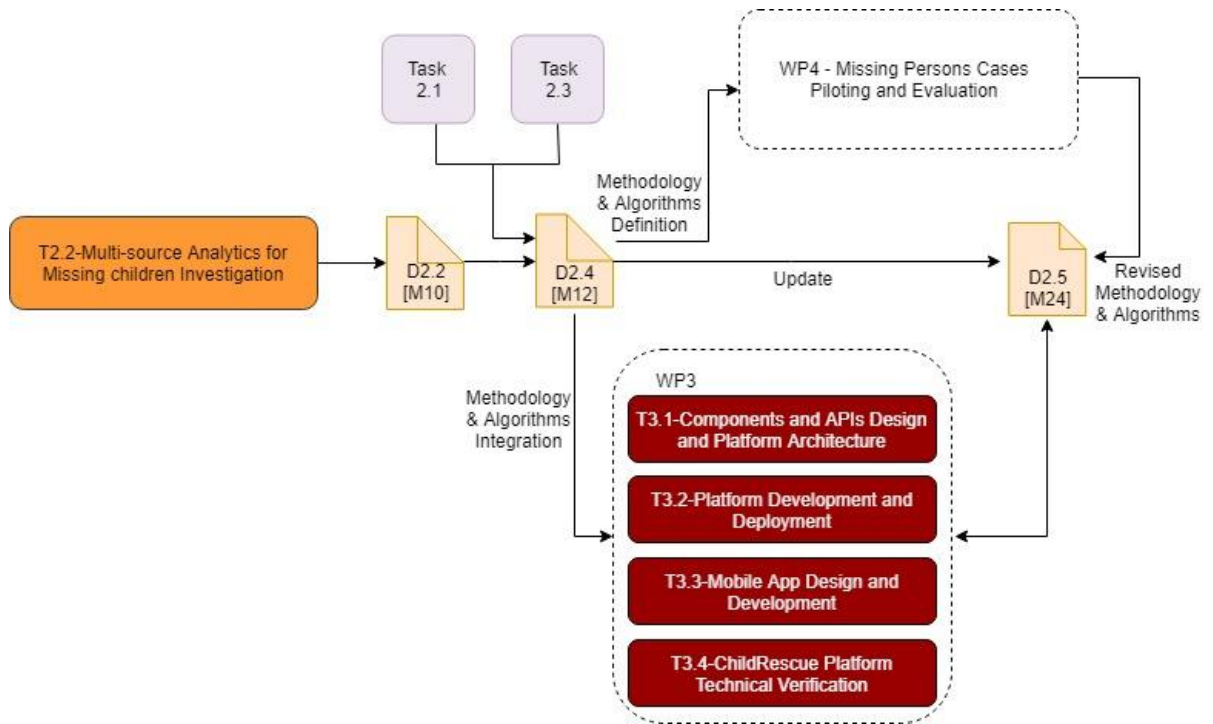


Figure 1-1 Relation to other WPs/tasks

2 Landscape Analysis

In the emerging era of Big Data, the analysis of information deriving from multiple sources is motivated by the increasing volume and availability of different types of data and the desire to create data-driven computer applications that can help human beings make complex decisions. The constant sensing and capturing, storing and sharing of personal and other information – knowingly or not – through mobile phones, social media networks or the so-called Internet of Things, creates massive quantities of data produced by and about people, things, and their interactions [1]. As a consequence, the notion of data analytics is getting an increasing amount of attention, in both the academic and business community.

Data analytics is an arbitrary collection of computational methods and techniques that are employed to examine data sets and harvest knowledge and meaningful insights from them.

Data analytics is commonly viewed from three major perspectives: descriptive, predictive and prescriptive [2]. In short, descriptive analytics, as the name implies, refers to methods that attempt to describe raw data and extract some form of useful information interpretable by humans. In a way, its purpose is to describe the past, and as such, it is closely related to Data Mining [3]. Predictive analytics, on the other hand, aims to forecast the future and make predictions based on discovered patterns in the given dataset. It originates from AI (Artificial Intelligence) theories and aims to unravel future events or trends. The last category, prescriptive analytics, not only predicts the future but can also recommend particular actions and solutions based on the predictions, in order to optimise decision making.

From an applications point of view, data analytics can be divided into six major categories, each one representing a relevant set of tools and techniques (Figure 2-1).

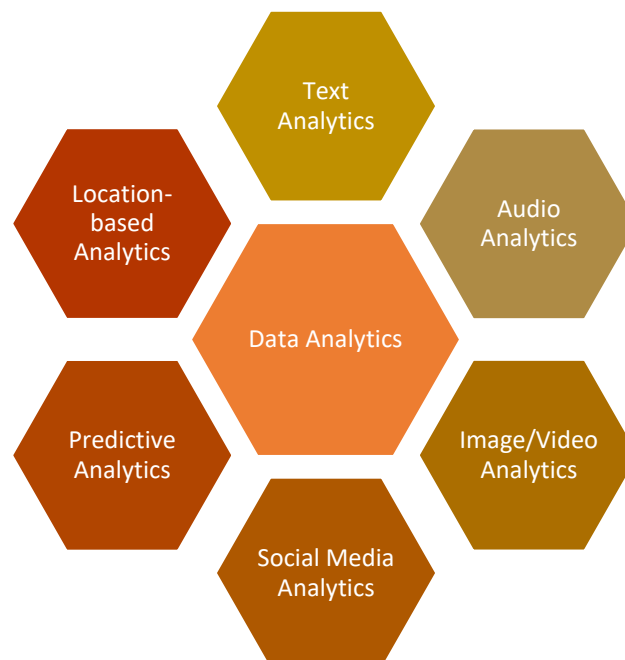


Figure 2-1 Main categories of Data Analytics techniques

In their recent findings, Gandomi & Haider proposed the following five categories [4]:

- **Text analytics** (or text mining) aims to extract information from documents and textual data. Social media feeds, e-mails, blogs, product reviews, news feeds, are some of the most common examples of textual datasets.
- **Audio analytics** is employed usually to process unstructured audio data, such as human speech. Currently, customer call centres and healthcare are their primary application areas.
- **Image & Video analytics** involves a variety of techniques to monitor, analyse, and extract meaningful information from images and video streams.
- **Social Media analytics** explores social networking channels and platforms and process user-generated content (e.g. text, images, videos), as well as the relationships and interactions among network entities.
- **Predictive analytics** consists of a variety of methods originating, as already mentioned, from the AI field, that seek to uncover patterns, capture feature correlations and predict future trends or actions.

But, considering the current trends in technology, we opt to add a sixth one, which encompasses a large set of today's tools and applications:

- **Location-based analytics** that focus on geospatial data and aim to improve location-based services.

It is worth noting that these categories may integrate one another or overlap with each other. For instance, the text and image analytics can play an important part in social media analytics, or location-based methods can be employed to assist in predictive analytics and vice-versa. In other words, a data analytics project may include some or all of the aforementioned categories in order to achieve the desired goals and objectives.

Today, data analytics technologies are widely utilised in commercial industries to improve business performance mostly through targeted advertisement, as well as governmental organisations, to improve on security and control. They can be applied in almost every sector: from agriculture, energy, or economics to healthcare, sports, or transportation. Depending on the particular application, the data that is analysed can consist of either historical records or new information coming from real-time processes.

One area that has been somewhat limited in its acceptance and use of these powerful new techniques is the public safety field, particularly in crime analysis and operations. Surprisingly, this comes in contradiction to what someone would expect, since police analysts, detectives, agents and other operational personnel, base their investigations on many of the principles shared by data mining and knowledge discovery. For example, the behavioural analysis of a violent crime, its characterisation, modelling and related predictions are very close to the computational methods associated with data mining and predictive analytics [5]. Nonetheless, after the events of 9/11 and even more during the last decade, predictive analytics and pattern recognition in crime-fighting have seen a strong surge of interest [6]. Considering several related studies, the data and methods utilised by what is today called predictive policing, can be divided into three broad categories [7][8]:

- Analysis of human behaviour profiles (to identify groups or individuals at risk of offending or become victims in the future)
- Analysis of time and space (to forecast places and times with an increased risk of crime)
- Analysis of social networks (to find patterns in network relationships and activities)

Although the cases of missing children or unaccompanied migrant minors, which drive the cause of ChildRescue, are not necessarily related to a crime, the sources and the types of data involved, as well as the methods employed, can be similar, if not identical, according to the research results of Task 2.1 (see D2.1). The analysis of human behaviour and relationships, along with spatiotemporal information, play a pivotal part in the investigation process for both a criminal activity and a missing child case.

Consequently, the main focus of our research study will be directed towards the most significant advancements in human behavioural profiling based on computational learning. In addition, we will explore a wide spectrum of algorithms for spatiotemporal data processing, as well as the most significant methods and techniques for social media analytics. As already mentioned, methods that deal with human behavioural models and patterns may overlap in several cases or be complementary to each another. This is also the case for the three selected areas of research interest. In closing, challenges and perspectives will be discussed for each of these domains, and the main take-aways to be considered by the ChildRescue methodology will be outlined.

2.1 Computational Learning in Human Profiling

The computational learning theory is a branch of Artificial Intelligence that studies and explores the ability of machines (i.e. computers) to learn from data. The idea of applying learning to computing devices may go as far back as the Turing machine, but the seminal paper that is usually referenced as the origin of the theory is accounted to L. G. Valiant [9] back in 1984. In his scientific work, Valiant proposed an attractive general model to study the computational, statistical and other aspects of learning. From then on, an explosive growth of the field commenced with numerous methods, theories and algorithms devoted to computational learning [10]. Today, the applied theories of computational learning have become quite popular by the name "Machine Learning" (for the purposes of this deliverable both notions will mean the same thing) and are employed in a wide range of computing tasks, from email filtering and fraud detection to speech and image processing [11].

Computational learning is also encountered in the analysis of human behaviour and activity. Recent advancements in computer vision have enabled video processing for behaviour recognition, e.g. for surveillance purposes [12], while ambient intelligence and Active and Assistive Living (AAL) applications can automatically classify human physical activities based on wearable (or other) sensor measurements [13]. Human-computer interaction [14], web usage and social networks [15], recommendation systems [16], are some other research domains that human behavioural patterns are extracted from and exploited so that modern applications can offer better and more personalised services.

Modelling and predicting behaviour through AI techniques is still an on-going task for the research community. Several behaviour-based approaches emphasise not on representing or reasoning about intentions, goals, or motivations, but instead rely on how predictions and patterns directly flow from data [17]. They argue it is impossible to actually find out why some behaviours occur and only mere speculations can be made about the reasons that drove these behaviours. This is in contrast to other AI trends, where the objective is to model cognitive belief states, intentions and internal structures, supporting the idea that an intelligent system can successfully simulate a cognitive creature [18]. In this informal battle, many researchers raise concerns regarding the generalisation ability of the behaviour-based approaches and the problems deriving from it [19]. Truth is, they are partially valid,

since the success of search engines such as Google and many other popular applications in behaviour recognition show that much can be done by using “just” correlation patterns.

In this context, the notion of *privacy* has surfaced time and again. However, while the word has remained the same, its meaning never stopped evolving. In the age of big data, and even more so in the future of the Internet of Things, this notion is poised to become all the more important. This phenomenon is taken to another level with “*profiling*”, the use of a person’s data to guess about aspects of his or her personality, generating insights about traits or habits that one may not even know are existing [20].

While extracting data and information from specific individuals is related to identification and control, in the context of data analytics the concept of *profiling* makes it possible to go beyond the personal level and track, monitor or measure various groups of individuals at an aggregated level. So, someone might reasonably wonder, “what is the difference between these two levels?”

The crucial difference is that the *individual level* deals with personal data of a specific individual, and this information is actually observed and recorded, i.e. it is factual knowledge. On the other hand, at the *profiling level*, the knowledge is not usually available. Instead the profile is applied to an individual so as to infer additional facts, preferences or intentions. Therefore, the profile consists of data-driven models that represent correlations, patterns or rules, that apply to a subset of the individuals; for example, missing children who have a family status X and are in the age group Y are more likely to be of missing category Z, if their daily engagement with social media is over W hours.

Profiling provides the means to infer knowledge about an individual that is not actually observed or recorded [21].

The aggregated level example is what Hildebrandt calls *non-distributive* profiles [22], in a sense that the profile properties will not hold true for all individuals that belong to that profile (in contrast to distributive profiles, where a property belongs to each and every member of the group).

In the literature study that follows, we investigate and review papers of the aggregated level, showcasing the fact that the profiling process may include various sources of information and have a huge impact in predicting personality traits and uncovering hidden behavioural patterns.

2.1.1 Research Literature Study

The analysis of human behaviour and activity is a quite broad topic with roots in many different sciences. A simple search for these terms (“human behaviour and activity analysis”) in Scopus¹ search engine yields more than 64,000 results, broken down into a large number of categories such as Psychology, Sociology, Criminology, Medicine, Biochemistry, Neuroscience and Computer Science. We, therefore, need to narrow down our scope, so we query Scopus about “Computational learning and profiles”. The results are then reduced to 1,177 documents, out of which 614 are related to computer science.

In order for the literature study to be efficient, a subset of these approaches and methodologies has been selected so as to present valuable insights and the underlying advancements and challenges of the field in respect to the ChildRescue project. More specifically, the following considerations were taken while reviewing the available publications and selecting the papers and reports to be analysed in this deliverable:

¹ <https://www.scopus.com>

- *Research significance*, by citation count, which is indicative of the article's research value.
- *Latest trends*, for the specific domain, in order to balance the fact that older publications have naturally more citations. Therefore, a separate search is made to include highly cited articles of the last few years (from year 2013 and afterwards).
- *Computational Learning* algorithms should be employed and evaluated.
- *Applicability*, towards the ChildRescue project. Papers with high relation and applied solutions to the ChildRescue project are favoured among others of similar citation number.

In Annex II.1 – Computational Learning in Human Profiling Literature, sixteen (16) selected papers have been analysed in depth and compared.

2.1.2 Key points extracted from the Literature Analysis

In our literature review we examined and compared a number of methods and algorithms for learning models out of human behaviour data. This automatic process is what we call *profiling*, and the models derived from it are the *profiles*. Profiles can consist of rules or correlations about the relations between features, but they can also divide the group of individuals into a number of subgroups, the members of which share certain properties, or they can be complex functions that compute a value or class label based on some features.

Profiling can be applied on any task that involves human interaction and activity. In our study we encountered various application domains, such as healthcare [23], education [24], economy and marketing [25][26], gaming [27], cyber security [28][29], mobile communications [29][30][31] and, of course, crime analysis and missing persons investigations [32][33][34].

In the majority of these cases, the methodology followed, involves the same general, but sequential, steps:

- First, some type of clustering is performed to shape groups of similar characteristics out of raw data. This process results in an initial profiling model, which in some research studies can be deemed sufficient (e.g. in [24], [27] and [32]).
- The next, usually optional, step introduces a form of dimensionality reduction, either by using a known algorithm or some heuristic method based on domain expertise, in order to decrease the dimension space and keep only the useful and informative input features. This process optimises the profiling model.
- In the last step, the task of classification (or regression) is executed using well-known computational learning algorithms. This process leads to predictions based on the profiling model deduced from previous steps. In cases where a model already exists, as in most personality prediction tasks (e.g. [29][30][31]), the classification or regression process can take place as a first and only step.

Other approaches that deal with textual data, something very common in social media, employ Natural Language Processing techniques, as an additional, albeit necessary, step to the aforementioned routine [35][36][37]. On the contrary, in cases where a sequence of actions or events is used to describe a behaviour, the procedure involved is, in many ways, different. For instance, dealing with web navigation patterns or system intrusion command sequences, as in [28] or [38] respectively, the preferred method is to use stochastic approaches, such as the hidden Markov model.

A sum-up of all algorithms examined in this literature review is presented in the following table:

Table 2-1 List of computational learning algorithms for human profiling

Objective	Algorithms
Dimensionality reduction	<ul style="list-style-type: none"> • Isomap, • Principal Component Analysis (PCA), • Singular-value decomposition (SVD)
Clustering	<ul style="list-style-type: none"> • K-means, • Expectation-Minimisation (EM), • Agglomerative algorithm, • K-medoids, • Archetypal analysis
Classification	<ul style="list-style-type: none"> • Support Vector Machines (SVM), • Logistic regression, • Naïve Bayes, • Decision trees, e.g. C4.5, • K-nearest neighbours (k-NN), • Artificial Neural Networks (ANNs), • Ensemble Bagging, • MultiBoostAB, • AdaBoostM1, • Random Forest
Regression	<ul style="list-style-type: none"> • Linear regression, • Poisson linear regression
Deep Learning (time-series)	<ul style="list-style-type: none"> • Recurrent Neural Networks (RNNs)
Anomaly detection	<ul style="list-style-type: none"> • Hidden Markov models (HMM)
Rule induction	<ul style="list-style-type: none"> • BruteDL

At this point, it is worth mentioning that we deliberately omitted research publications concerning activity recognition and computer vision techniques on human motion patterns and face/body expressions (for examples see [39]), since these methods, most probably, will not be adopted by this project as they raise the most serious privacy concerns.

From a data source viewpoint, it is evident that retrieving social and activity characteristics from humans is not a simple task. The complexity and unreliability of human behaviour, as well as the ethical aspects surrounding it, make the effort even more demanding. Data usually include digital records collected by an organisation, which can be enriched by survey questionnaires and psychometric tests, as well as by external sources. Our literature analysis reveals that these external data sources can be found in wearable sensors and bio-signals, mobile phones, internet applications and social media, among others.

In the special case of missing persons, which relates to the ChildRescue project, a rather low number of research papers have been published that employ computational learning methods for profiling. One explanation for this could be the confidential and sensitive nature of the required data sets,

which cannot be available to anyone. Blackmore et al. in their study [33], present a digital record of a missing individual case that consists of several variables taken from police files (Table 2-2 **Error! Reference source not found.**). With the use of data mining techniques, they proceed to train and test a classifier. The results reported in this work, nevertheless, indicate there are some inconsistencies, probably due to some of the variables containing human judgments and estimations (e.g. "Is the missing person known to be socially deviant or rebellious" or "What does the reporting person suspect has happened") that, according to the authors, may eventually affect data quality. A note we must keep.

Table 2-2 Variables based on information in police files describing missing persons [33]

Variables	
<ul style="list-style-type: none"> • Does missing person have any dependents • Residential status • Time of day when last seen • Day of week when last seen • Season of year when last seen • Last seen in public • Is this episode out of character for the missing person • What does the reporting person suspect has happened • Any known risk factors for foul play • Is missing person known to be socially deviant or rebellious • Is there a past history of running away 	<ul style="list-style-type: none"> • Is there a past history of suicide attempt or ideation • Any known mental health problems • Any known drug and alcohol issues • Any known short-term stressors • Any known long-term stressors • Method of suicide • Was the perpetrator known or a stranger to the victim • Was the missing person alive, deceased or hospitalised when located

Regarding external data sources, social networking media have become a major research and business activity field due to the huge amount of public data that gets generated each day, as well as the availability of tools to retrieve and analyse them. It is also a research area undergoing rapid development and evolution, because of the commercial pressure and the potential for using social media data for computational (profiling) research. For instance, Kosinski in [35] processes about 58,000 user profiles from Facebook, exploiting user Likes on music, movies, product pages, etc., and the results of several psychometric tests based on the popular Five-Factor model [40], to predict personality attributes, such as political views, religion or sexual orientation.

It is clear that the domain of social media analytics is quickly rising and encompasses more and more applications, with profiling being just one of them [41]. For this reason, and for its important role in the ChildRescue project, social media analytics will be investigated and discussed separately in one of the sections to follow.

2.1.3 Challenges and Perspectives

Computational learning research on profile assessment has been based on a philosophy that emphasises prediction over explanation [42]. It has thus focused almost exclusively on the convergence of predictive analytics models with established personality measures. In contrast, little research has been done to use computational learning and digital records of behaviour to further understand a personality. So, there is appreciable potential in using current machine learning technology to develop improved tools for better understanding what the models are actually

measuring, from a psychological point of view [43]. In other words, sociologists and psychologists, apart from a profile prediction, would like to know why and under what conditions something will occur, so that they can act pre-emptively.

Another challenge in profiling is the temporal factor of a behaviour. Human behaviours change through time, depending on external stimulations and certain events. Most research studies seem to ignore this part and prefer to form models of static knowledge in a particular time-frame. More recent studies though, explore the power of deep learning and seem to tackle this issue with success. Work in [25] is a good example.

The major challenge of profiling, however, is data privacy, and that is because the primary task of profiling is the development of models from aggregated personal data, which does raise several privacy concerns. In response to this, from the early days of data mining in the late 1990s, a part of the academic community focused their work on privacy issues leading to the notion of Privacy-Preserving Data Mining [44]. In this framework, and following the advancements in database technologies, many researchers started to study the technical feasibility of realising the data mining methods using perturbed records of individuals (i.e. randomly modified values with sensitive pieces of information)[45].

Nevertheless, after the enforcement of the EU regulation 2016/679 (known as GDPR²) in May 2018, most social networks and internet platforms have ceased to provide access (through APIs) to personal public data. A period of adjustments and optimisation from all sides seems to have begun, with unknown results. We do hope the desired balance between data security and protection, and scientific progress, will be reached soon.

In the ChildRescue conceptual approach, the profile assessment deriving from multiple data sources is regarded as one of the cornerstones of the project. It is, therefore, crucial to apply the most appropriate profiling methodology, depending on the data available, harnessing the power of predictive analytics and machine learning in the most efficient way. Data anonymisation techniques will be applied as well, so that the process of profiling is in line with the data privacy protection rules and regulations.

Some of the aforementioned challenges will, most probably, be encountered in the context of the ChildRescue project. It is, therefore, important for the ChildRescue overall methodology to be able to perceive these challenges in time and confront them efficiently. In particular, the explanatory aspect of rule-inferring computational techniques should be considered when selecting the appropriate algorithms. Additionally, the human temporal behaviour will play a significant role in routing estimation and POI suggestion, and as such it will be covered in detail in section 3.3 of this document, while privacy issues will be addressed in the framework of Task 2.3 and documented in D2.3.

2.2 Exploiting Spatiotemporal Data coming from Multiple Sources

As already shown, research about large scale human behaviour patterns is often based on user-generated data either within wireless communication networks such as mobile phone networks, or on social media networks like Facebook or Twitter. In combination with Geographic Information Systems

² <https://eugdpr.org/>

(GIS) these inherently spatiotemporal data enable novel capacities to analyse and visualise large-scale human dynamics in a more integrated manner, even close to real-time.

The first actual use of spatial analysis ever recorded was back in 1854 when a British physician, John Snow, began mapping cholera outbreak locations and eventually noticed that the majority of cholera cases were commonly found along the water line [46]. A century later, from early 1960s to 1980s the advancements in mapping technology, computers and data storage led to what we know today as GIS. Actually, it was Roger Tomlinson, an English geographer, who first used the term "Geographic Information System" in his publication in 1968 [47]. He has been later acknowledged as the father of GIS.

Another account of early spatiotemporal analysis is found during the first decades of the 20th century and is closely related to the migration mobility problem. Many sociologists and demographers put effort on the interpretation of people movement in space using mathematical formulas and statistical analysis to predict the distance or direction of the (internal or not) migration streams, either in rural or urban environments [48].

It is true that human mobility patterns reflect many aspects of life, from the global spread of infectious diseases to urban planning, traffic forecasting, tourism and daily commute patterns. Today, more than ever before, the large availability of human location tracking datasets through location-aware devices and services due to the advancements in social media, mobile telecommunication networks and the large-scale deployment of GPS technologies, has generated growing scientific interest in their possible exploitation and interpretation. As this information is usually timestamped, it is also possible to detect occurrences of activities in temporal relation to each other or to specific daytimes. Examples of these datasets include: mobile phone calls, credit card transactions, bank notes dispersal, check-ins in internet applications and geotagged user-generated content in social networks, among several others. Both personally identifiable information of individuals, but also massive anonymous data are geolocated. User GPS logs or social media posts are examples of the first category, whilst mobile network traffic and anonymous smart-card transactions fall under the second.

Interdisciplinary approaches, generally deriving from knowledge discovery, are utilised towards the deciphering of raw mobility data. Algorithms and techniques from the fields of data mining, machine learning and statistical analysis are employed and in return lead to the discovery of human movement or transportation patterns and event detection [49]. Especially for human trajectories, it is found that they show a high degree of temporal and spatial regularity, with each individual being characterised by a time-independent characteristic travel distance and a significant probability to return to a few highly frequented locations [50]. Mobility patterns were also found to represent human behaviour in catastrophic events, e.g. the 2010 Haiti earthquake [51].

Our research literature study aims to shed light on the domain of spatiotemporal data analysis (also encountered as *human mobility patterns analysis*) and review particular publications related to behaviour modelling tasks of tracking persons. Predicting their trajectory patterns or locating points of interest (POIs) will be our primary focus as well.

2.2.1 Research Literature Study

The target domain of our literature study is the spatiotemporal data analysis. A Scopus search for these terms ("spatiotemporal data analysis") yields more than 13,500 results, out of which the 3,891 belong to computer science. For the shake of a beneficial literature review towards the ChildRescue

project, we need to narrow down the publications set, adhering to some criteria that can lead to better and more valuable insights, as well as challenges and perspectives related to the project's cause. More specifically, the following considerations were taken while reviewing the available research studies and selecting the papers to be analysed in this deliverable:

- *Research significance*, by citation count, which is indicative of the article's research value.
- *Latest trends*, for the specific domain, in order to balance the fact that older publications have naturally more citations. Therefore, a separate search is made to include highly cited articles of the few years (from year 2013 and afterwards).
- *Human patterns* should be in the core analysis of each work. We therefore excluded approaches that tracked animals, natural phenomena or objects.
- *Applicability*, towards the ChildRescue project. Papers with high relation and applied solutions to the ChildRescue project are favoured among others of similar citation number.

In Annex II.2 – Exploiting Spatiotemporal Data from Multiple Sources Literature, seventeen (17) selected papers have been analysed in depth and compared.

2.2.2 Key points extracted from the Literature Analysis

According to Toch et al.'s taxonomy [49], human mobility pattern analysis can be categorised into three broad categories:

- User modelling, where the object of analysis is a single individual.
- Place modelling, where the object is a geographic area (visited by different individuals).
- Trajectory modelling, where the object is a set of spatial-temporal points created by the same individual.

User modelling applications analyse the mobility patterns of a single individual for extended periods of time. In such applications, the model can predict where a particular user will be at different times of the day or recommend POIs to visit (e.g.[52][53][54]). *Place modelling* applications analyse the characteristics of a geographic location or a set of locations, in order to profile it and classify the type of place according to the mobility patterns of people coming in and out of it (e.g. [55][56]). *Trajectory modelling* applications require a set of spatial-temporal points that reflect a trajectory, defined as a movement pattern through a set of locations or a set of objects and time (e.g. [57][58]). In contrast to user modelling, in trajectory modelling, the identities of the moving objects are not necessarily a factor in the analysis. All three categories hold a different value on human mobility analysis, but in several cases, the borderline between them is not really clear.

In respect to applications, most of the reviewed studies aim to create some form of recommendation for the user, as part of a location-based service. For example, in [52][54][59] & [60] the next - predicted - point of interest is suggested to the user, based on his or her mobility profile, while in [61] the analysis of GPS trajectories from 107 users infers the right sequence of POIs, which could be tourist attractions, and recommends it to the user. If a man is lost in the wilderness, his route trajectory can be predicted based on the profile of the missing person (age, gender, professions, intention, etc., which translate into direction, distance, and dispersion of travel) and the man can be safely found, or so claim Mohibullah & Julie [57] and Lin & Goodrich [62] in their related work. Furthermore, an event or a place of social commotion can be detected using a probabilistic location inference on mobile phone patterns, according to the results of [63].

A very interesting idea, coming from Cho et al.[53], is to combine social networking connections with place check-ins, following a hypothesis that we tend to go where our friends go, when we are not at

work (or school). It is shown that social relationships can explain about 10% to 30% of all human movement, while periodic behaviour (e.g. going to work and back home every day) explains 50% to 70%. The rest is more or less unpredictable (i.e. no regular patterns).

Another emerging area of human mobility research is urban planning, where spatial and temporal patterns are studied and processed so as to explain urban traffic and adjust traffic lights or plan new transportation routes. Such information can be derived from transportation cards transactions [64] or location-based social media content [65]. Speaking of transportation, even the medium of transport can be predicted for a given mobility pattern, using a large set of recorded human GPS trajectories and urban infrastructure details [58].

From an algorithmic point of view, which defines our main interest, we encountered a clear distinction from the early review of the literature. On one hand, there are methods that make use of probabilistic (stochastic) modelling and analysis, and on the other hand, we have methods that employ machine learning algorithms and techniques.

The common ground on all these methods is the necessary step of data pre-processing which allows for an efficient and producible analysis. Usually a method of trajectory or area segmentation is required at this stage. Then, spatiotemporal data modelling can be accomplished either using a probabilistic or statistical approach (e.g. Markov model), or a machine learning technique (e.g. K-means clustering). In some cases, there can be a combination of both, each applied on different aspects of the same problem, like in [66]. In more recent works, deep learning algorithms have been employed to deal (mostly) with the temporal aspect of mobility pattern analysis, using Recurrent Neural Networks (RNNs) that have the ability to recall past states [54][59].

A sum-up of all methods and algorithms examined in this literature review is presented in the following table:

Table 2-3 Methods and Algorithms for spatiotemporal data analysis

Objective	Algorithms
Probabilistic (Stochastic) mobility modelling	<ul style="list-style-type: none"> • Markov decision process and Order-K Markov model, • Probability density functions, • Bayesian model, • Custom models
Trajectory Clustering	<ul style="list-style-type: none"> • K-means, • Voronoi diagrams (partitions), • Non-Negative Matrix Factorisation, • Expectation-Minimisation (EM) • Density-based (e.g. OPTICS)
Classification	<ul style="list-style-type: none"> • Support Vector Machines (SVM), • Naïve Bayes, • Decision trees, • Artificial Neural Networks (ANNs), • Random Forest
Text based analysis	<ul style="list-style-type: none"> • TF-IDF, • Latent Dirichlet Allocation (LDA),

	<ul style="list-style-type: none"> • Dirichlet Multinomial Regression (DMR)
Reinforcement Learning	<ul style="list-style-type: none"> • Inverse reinforcement learning
Deep Learning (time-series)	<ul style="list-style-type: none"> • Recurrent Neural Networks (RNN, ST-RNN, LSTM)
Expert systems	<ul style="list-style-type: none"> • Fuzzy inference system
Recommenders	<ul style="list-style-type: none"> • Collaborative Filtering

Experimentation data originated from a great variety of sources, which constitutes another distinction of the relevant literature publications. Many researchers take advantage of mobile phone data, such as call detail records and GPS log files in order to predict significant locations [56], recommend interesting locations or POIs for next visit [54][59][60], detect social events [63] or travel sequences [61][68], even simulate and predict human mobility behaviour when lost in the wilderness [57][62].

A more recent trend in the research community targets social media data from location-based networks, like Foursquare [65][66], while in some cases the combination of social networks with mobile phone data seems to produce better results [53]. On the spatial aspect of the analysis, the use of open data, such as transportation infrastructure, road networks or land use, enrich the information available and improve predictive capabilities as shown in [58][55][67]. Of course, any other data source that can explicitly or implicitly offer some type of human mobility patterns (e.g. transportation card transactions [64]) can prove very useful for a successful data analysis.

The table below summarises the data sources encountered during our literature investigation, along with the respective datasets, that could potentially be used in the ChildRescue framework.

Table 2-4 Most widely used data sources in spatiotemporal data analysis

Data Sources	Relevant Datasets
Mobile Communications	<ul style="list-style-type: none"> • Mobile phone calls and text messages (location points) • General mobile phone usage (user trajectories)
GPS	<ul style="list-style-type: none"> • Log files
Social Networks	<ul style="list-style-type: none"> • User check-ins (e.g. from Foursquare, Facebook) • Geo-tagged images and videos (e.g. from Flickr, Facebook) • Location reference in a post or activity (e.g. from Facebook, Twitter) • Other geo-referenced content (e.g. Yelp crowd-sourced location reviews) • Network connections
Open data	<ul style="list-style-type: none"> • Transportation infrastructure (e.g. from OpenStreetMap) • List of popular POIs in a particular area (e.g. from OpenStreetMap, national open data) • Transportation time schedule (city open data) • Land use (e.g. MassGIS) • Road networks (city open data) • Weather data

Other data	<ul style="list-style-type: none">• Transportation card transactions (e.g. Oyster card)• Surveys (e.g. Dutch National Travel survey [69])
------------	--

2.2.3 Challenges and Perspectives

We reviewed a number of different use cases regarding the exploitation of spatial and temporal data in order to infer location-based practical information. In some of these cases, additional data sources were employed, either from social networks (such as user connections or POI reviews) or originating from open data services (such as road networks or weather data).

Statistical analysis has been the most common approach for analysing spatial data, where a large number of algorithms exist including various optimisation techniques. A major drawback of this approach is the assumption of statistical independence among the spatially distributed data. Furthermore, statistical analysis cannot model nonlinear rules efficiently and cannot work well with incomplete or erroneous data, or with categorical values. These disadvantages were partially solved with the advent of spatial data mining [70]. However, traditional data mining techniques and algorithms perform poorly on spatiotemporal tasks, and require significant modifications to exploit the rich spatial and temporal correlations and patterns embedded in the datasets. The unique characteristic of spatiotemporal datasets is that they carry time, distance and topological information which require geometric, as well as temporal, computations. In most cases spatial and temporal relationships are not explicitly defined, and thus, they should be extracted from data, which ultimately causes a processing overhead [71].

Another challenge, related to using open data, is the fact that it is difficult to match the available open data to the datasets one plans to analyse, in respect to the desired location or time-frame. For example, while there are plenty of data for a metropolitan city in US for the year 2015, when you seek a relevant dataset for a rural area in Belgium or Greece for 2017, you hardly find any. Therefore, careful consideration of the available open data sources should be made before selecting the appropriate ones.

Lastly, as in every aspect of this project, privacy issues play and will continue to play a prominent role when challenges are discussed. Preserving anonymity and protecting personal information should be the first priority. Thankfully, privacy is an essential requirement for the provision of electronic and knowledge-based services in modern e-business, e-government, or e-health environments, and thus, there are several methods and algorithms for protecting location privacy, among others, without losing much of their efficiency and performance [72].

For ChildRescue, the spatiotemporal data analysis will be an essential part of the active missing children investigation process, for its ability to recommend possible POIs the child might visit, and predict his or her trajectory near the area the child was last seen. However, we expect the available spatiotemporal data to be scarce and dispersed, so it is important for the analysts to have selected and modified the appropriate algorithms in such a way so that some probability indications can be obtained. Security precautions will also need to be considered to ensure the protection of data privacy.

2.3 Social Media Analytics

Social media can be defined as a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content [73]. Social media, nowadays, constitute an integral part of our everyday lives and affect us in several ways. The communication and interaction barriers have been diminished, while the ability for users to generate content, interact with friends and expose personal thoughts, visiting of places and snapshots of their everyday lives, have made social networks extremely popular, especially among teenagers. At the same time, social media have proven to be a valuable pool of information and trends, widely exploited in the business, bioscience and social science areas, e.g. for measuring customer satisfaction, forecasting stock market fluctuations, predicting social impact, or providing insights into community behaviour [74].

The opportunities associated with the analysis of all these data have helped different organisations generate significant interest in Social Media Analytics, which is often referred to as the informatics tools, techniques, frameworks and applications that collect, monitor, analyse and visualise critical data from social media, so as to facilitate and extract useful patterns and knowledge [75].

Social media analytics is actually an extension of methods and technologies used in web analytics. Over the past two decades, web analytics, that emerged with the coming of the internet and World Wide Web, has been an active field of research, building on the data mining and statistical analysis foundations of data analytics and on the information retrieval and text mining. The advancements in internet technologies and mobile communications led rapidly to the rise of social media platforms and services, with social networking being the most popular online activity. In fact, internet users seem to spend more than 20% of their online time on such platforms, according to recent statistics [74], generating an enormous amount of informative content, interacting with it and being affected by it. Thus, it was only natural for the web analytics research community to divert its interest towards the, most promising, field of social media analytics. This has led to the publication of numerous research works, as well as the creation of novel data services, tools and analytics platforms. The analysed social media may include blogs, message boards, multimedia platforms, and customer reviews platforms, among others. However, it is the web search and microblogging social media (e.g. Twitter, Facebook, Instagram, or Foursquare) that have attracted most of the interest.

Social media data is clearly the largest, richest and most dynamic evidence base of human behaviour, bringing new opportunities to understand individuals, groups and society. In fact, social media analytics are not only about informing, but also transforming existing practices in politics, marketing, investing, entertainment and news media. Because of this, the scope of social media analytics varies greatly. The largest part of research study is devoted to text mining and processing techniques, like sentiment analysis, topic or influence modelling, impact monitoring, opinion mining or news and trends analysis [76]. Other approaches focus on community network connections, personality traits prediction, venue recommendations or microeconomics and marketing analysis. Even approaches that attempt to track down the spread of infectious diseases through social media have been proposed [77].

In this context, predictive analytics is the primary tool, able to utilise user content (text and images), personal profile information and user preferences, and the social network itself (network connections and their attributes), sometimes combined with open data and information from other sources, in order to provide meaningful insights and content-based analysis [78]. In the lines to follow we will

investigate the power of predictive analytics when employed to process the rich and prosperous social media data.

2.3.1 Research Literature Study

As a relatively new term, but with a flourishing popularity, the “social media analytics” query in Scopus brings back about 2,241 documents. It is noteworthy to say that the first publications start to appear in 2005 and 2006, and from then on there is a large increase after 2012.

The wide spectrum of methods and applications is also observed. From news and opinion mining to stock market prices and election predictions. It is therefore necessary for us to focus on reviewing publications related to the purposes of ChildRescue, and in particular, under the prism of predictive analytics, we opt to investigate the subdomains of personality prediction, venue recommendation and sensitivity analysis.

For the first two subdomains, we have already had a glimpse in the previous sections. Here, we are going to investigate both domains from the social analytics perspective, covering the most informative, and useful to our cause, cases.

During our review, a few considerations were taken so as to carefully select the papers to be analysed, which are the following:

- *Applicability*, towards the ChildRescue project. Papers with high relation and applied solutions to the ChildRescue project are favoured among others of similar citation number.
- *Research significance*, by citation count, which is indicative of the article’s research value.
- *Latest trends*, for the specific domain, in order to balance the fact that older publications have naturally more citations. Therefore, a separate search is made to include highly cited articles of the few years (from year 2013 and afterwards).

In Annex II.3 – Social Media Analytics Literature, twenty-one (21) selected papers have been analysed in depth and compared.

2.3.2 Key points extracted from the Literature Analysis

The methodologies reviewed in the field of Social Media Analytics can be grouped into three main application fields: Personality prediction, Recommender Systems, and Sentiment Analysis.

Personality Prediction is a subcategory of predictive analytics, concerned with the classification of distinct personality traits, which are serving as the class categories. This classification can be performed on users of social media by sole use of openly available profile information. The purpose of this process is to determine hidden personality traits or attributes, perhaps unknown to a person’s close family or friends. In the majority of the here presented articles, these categories are provided by the Big Five model [40]. Although the goal is the same, the feature set utilised by these methods varies greatly. In the most typical approach, personality prediction is conducted using social media posts, i.e. short public messages, that usually describe some activity or express some feeling, upon which a method of linguistic analysis is applied [79]. As an improved alternative, some researchers suggest the usage of meta-attributes that characterise a post, such as the number of words, length of text, emoticons, or exclamation marks [80], while others promote the exploitation of details that characterise a user profile, such as personal preferences, activities and networking structure [81]. In more recent publications, the combination of two or more social media platforms is proposed, which seems to improve the error rates of the classifiers [82][83].

A drawback of the big-five approach is the fact that the users, the profiles of which are used as dataset, have to answer a questionnaire survey in order for their personality to be categorised and for the algorithms to be trained. On the other hand, the exceptions to the big-five-centric approach, focus on particular emotions and intentions. For instance, De Choudhury et al. tackle the challenge of depression detection [84], while a more recent study investigates reddit forums for traces of suicidal ideation [85].

Recommender Systems seek to predict a user's preferences so as to suggest an item or service that the user will appreciate (and perhaps purchase). So, the unprecedented opportunity offered by the social media, for gaining insights into the needs and desires –current and future- of millions of potential customers, could not be left unexploited. A large number of the so-called recommender systems was developed over the last years, combining available user information (demographics, family status, purchases, personal preferences) and recommending items that were calculated as the most appealing for the user. An advancement in recommendation systems was the aggregation of social network parameters in order to refine the resulting recommendations with influence of the social circle. The prevailing techniques in recommender systems are the Content-based Filtering (CBF) and the Collaborative Filtering (CF), as well as combinations of the two.

In respect to location or POIs recommendation based on social media data, a variety of approaches can be found in literature. Some suggested systems exploit not only direct satisfaction indicators, such as ratings, but also user location history and check-ins to calculate user preferences and perform user profiling [86][87]. In other cases, collaborative filtering was adapted for increased estimation accuracy and minimised complexity. This was achieved by limiting the user space, for example only to the friends of the user based on behavioural studies claiming that friends are more likely to share tastes and preferences, but also drag one another to an event/venue [88]. Another approach was the selection of "local experts" and inferring a venue score based on their opinions [89]. Sentiment analysis has also been utilised for extracting user preference scores and assisting in location recommendation [90]. In more recent studies, the aggregation and collaboration of multiple recommenders is examined in multi-dimensional contextual information [90][91][92].

The most popular of the three techniques in the studied literature was collaborative filtering, with the reason probably lying on the selected field of interest -location recommendation-, which is slightly differentiated from tangible items recommendation.

Sentiment Analysis is a subject of study, both as an applied technique used in other applications of Social Media Analytics (for example in predictive analytics), but also as an application itself. As defined in [93], sentiment analysis is the computational study of people's opinions, attitudes and emotions toward an entity. As such, sentiment analysis has been employed in many research studies that analyse social media data, especially for brand building and marketing monitoring applications. Sentiment analysis techniques are divided into two main categories, which can also be intersected in hybrid models: the machine learning approach (un-/semi-/supervised) and the lexicon-based approach [94]. Generally, both techniques achieve high performances.

The usual applications of sentiment analysis involve some form of text processing and linguistic analysis. Twitter is one of the most commonly studied social platform mainly for two reasons: the feasibility to easily access user content through the Twitter API and the character limitation of tweets. The number of related research works, prove this claim [95][96][97][98]. Facebook however,

although not used as corpus often, is also a microblogging platform, and as such, its posts and metadata can be analysed in a way similar to tweets [99].

Sentiment analysis can also be expanded on images, a field relatively less covered. For instance, in [100], the visual information of an image is exploited for sentiment classification purposes. When the image sentiment analysis is enhanced with contextual information, such as tags and comments, it is demonstrated that the positive effect of textual information in the classification of images is limited up to a threshold, after which the calculations are subject to overfitting [101].

According to Kalampokis et al. [102], a common methodology for predictive analytics when dealing with social media data in general, involves the following three steps/subprocesses:

1. Data conditioning, i.e. collection and filtering of data, determining time-window and location/area, identifying profile characteristics.
2. Feature selection, i.e. choosing appropriate prediction variables usually based on some feature selection or feature transformation algorithm.
3. Predictive analytics, i.e. data modelling, pattern extraction and performance evaluation.

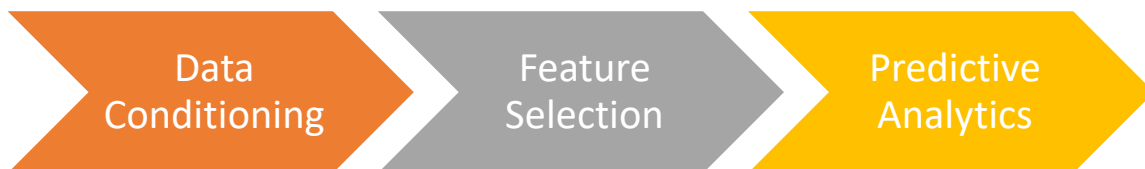


Figure 2-2 General methodology for predictive analytics in social media

The first phase, data conditioning, includes not only the collection of appropriate data, but also the selection of social media attributes that will lead to meaningful insights. Depending on the task, a selection should be made among volume variables (e.g. the number of posts, frequency), sentiment-related variables (measuring the sentiment load of the data), variables depending on profile characteristics (e.g. number of friends, preferences, activities), as well as location and time-window specifics. The second phase, called feature selection, deploys an algorithmic strategy that excludes variables of low information, or creates new variables by transforming the data. The last phase of predictive analytics consists of: 1) the creation of the predictive model, such as linear regression, Markov models etc., 2) the incorporation of external predictors, like mobile phone data or GPS logs, and 3) the evaluation of the developed model using a generally approved metric.

The methods and algorithms encountered during the study of the social media analytics literature were differentiating depending on the given task. Table 2-5 summarises all of the examined approaches.

Table 2-5 Methods and Algorithms for social media analytics

Task	Objective	Algorithms
Personality prediction	Linguistic Analysis (NLP)	<ul style="list-style-type: none"> • Linguistic Inquiry and Word Count (LIWC), • Latent Dirichlet Allocation (LDA)
	Dimensionality reduction	<ul style="list-style-type: none"> • Principal Component Analysis (PCA), • F-statistic subsampling

	Clustering	<ul style="list-style-type: none"> • K-means
	Regression	<ul style="list-style-type: none"> • Ordinary Least Squares, • Linear regression, • Decision Trees, • Random Forest, • Support Vector Machines (with RBF kernel)
	Classification	<ul style="list-style-type: none"> • Naïve Bayes, • Logistic regression, • Support Vector Machines, • Multi-Layer Perceptron (MLP)
Recommenders	Location Recommendation	<ul style="list-style-type: none"> • Model-based Collaboration Filtering, • Memory-based Collaboration Filtering, • Custom Collaboration Filtering approaches
	Clustering	<ul style="list-style-type: none"> • Non-negative Matrix Factorisation
Sensitivity Analysis	Linguistic Analysis (NLP)	<ul style="list-style-type: none"> • Lexicon based clustering, • Lexicon based classification
	Deep Learning (images)	<ul style="list-style-type: none"> • Convolution Neural Networks
	Deep Learning (word-embeddings)	<ul style="list-style-type: none"> • Convolution Neural Networks
	Clustering	<ul style="list-style-type: none"> • K-means, • Non-negative Matrix Factorisation
	Classification	<ul style="list-style-type: none"> • Decision Trees (C4.5), • Support Vector Machines, • Naïve Bayes, • Maximum Entropy, • Random Forest

Although methods and algorithms may differ, the social media data types that are available for research are, more or less, the same. Text messages, user profile details, check-ins, reviews, activities and network connections are only some of them. In the following table, the reader can view all types of data sources utilised by methods and algorithms in the literature.

Table 2-6 Most widely used data types in social media analytics

Data types	Relevant Social Media
Text messages	<ul style="list-style-type: none"> • Twitter, Facebook, Reddit, Digg, MySpace

User check-ins	<ul style="list-style-type: none"> • Facebook, Foursquare, Gowalla, Brightkite³
Profile details	<ul style="list-style-type: none"> • All, but Facebook has the most information
Customer Reviews	<ul style="list-style-type: none"> • Foursquare venues
Activities (Likes, shares, etc.)	<ul style="list-style-type: none"> • Facebook, Twitter, Instagram
Images	<ul style="list-style-type: none"> • Instagram, Flickr
Audio-Video	<ul style="list-style-type: none"> • YouTube

2.3.3 Challenges and Perspectives

It is by now clear that social media analytics can be a powerful tool to the scientific toolkit. For some, the introduction of social media data analytics has had impacts in the study of human behaviour similar to the invention of the microscope or the telescope in the fields of biology and astronomy: it has produced a qualitative shift in the scale, scope and depth of possible analysis. Such a dramatic leap, however, raises several concerns to the researchers that derive from the multi-faceted and complex nature of human communications and socio-cultural interactions[14] [103].

One such issue is the prevalence of single social platform studies (e.g. Twitter, or Facebook), which overlook the wider social ecology. Even in cases where multiple platforms are examined, their data are processed separately, for comparison reasons. Therefore, multi-platform analyses should be sought and multi-methods should be examined, that will be able to capture the overall user social interaction and diffusion. In addition, whenever possible, the analysis of social media data should be paired with surveys, interviews, ethnographies, and other methods so that biases and short-comings of each method can be used to balance one another and arrive at richer answers.

Another major difficulty is the undesirable computational overhead set by the large volume of social media data. Therefore, the systematic filtering of irrelevant and noisy social media data prior to the analysis, is of high importance for the accuracy of the resulting predictions and the elimination of unnecessary complexity.

Furthermore, the raw data encountered in social media is usually of poor processing quality. Fake accounts, type errors, or on purpose lies and false data, create a headache for the analyst, adding an extra overhead to the overall procedure of extracting useful information. Clear examples are the attempts to predict presidential elections, as explicitly shown in [104], where the methods tested perform slightly better than clear chance. According to the same paper, for such an election prediction system to be reliable, it must have a strong algorithmic background, consider the idiosyncrasy of the area and the possible manipulations from spammers and thirdly, identify why the system produces the results it does. These conclusions could be generalised for predictions using social media analytics in other fields, as well.

As already noted in previous sections, during the last few years, there are raising concerns about social media public data harnessing. Researchers worldwide are facing the fact of the increasing

³ Gowalla and Brightkite used to be location-based social platforms that both closed in 2012. Various datasets extracted from these two social networks are still utilised for research experiments.

access restriction set by the companies (be it for user privacy or business protection) towards social media data. The new EU regulation (GDPR - effective since the 25th of May 2018 worldwide), in particular, has already made the most popular social networks restrict the retrieval of public data through their APIs, which were freely available before.

This brings forth a major issue for the ChildRescue project, since social media analytics aim to provide significant added value to the current status of missing children investigation. In what way the project will handle the limitations in accessing social media information, has become an open challenge, which ChildRescue further discuss and address in section 3.4 of the presented deliverable.

2.4 Discussion and key-takeaways

The ability of analysing data coming from multiple sources, in order to model human behaviour and make applicable predictions, has been investigated through a literature overview.

The results show that, on one hand there is great multi-disciplinary research interest in this field, but on the other hand, the potential benefits are hindered by the overall human complexity and unpredictability. Traditional data origins, like forms, surveys, or digital records, as well as modern sources of information, such as open data, mobile phone activity, GPS logs, and social networks and applications, were encountered during this investigation, exposing the vast availability of human behavioural data.

Most of the methods and algorithms that were studied and analysed in depth, derive from the field of computational learning and data mining, whereas probabilistic techniques were applied when the research was related to spatiotemporal datasets.

Some important key-takeaways taken from the literature analysis are the following:

- Multiple data sources produce better results than a single source.
- The selection of the most appropriate features in a dataset increases performance.
- The quality of data plays a significant part on data analytics in general.

The traditional methods of collecting and processing behavioural data and classifying the data-subjects used to rely on sociological methods, manual analysis and interpretation. Nowadays, the process is highly automated and dependent on computer technology and many organisations routinely use data mining technology and techniques to analyse the large amounts of data available. However, adapting to an era of data-driven decision making is not always a simple task. Some companies have invested heavily in technology but have not yet changed their organisations so they can make the most of these investments. Others are struggling to develop the talent, business processes, and organisational muscle to capture real value from analytics. But the most common mistake is the insatiable and aimless gathering of unstructured and uninformative data just for the sake of data collection, which results in low quality datasets that require a lot of manual labour in order to be qualified for data analytics. Thus, in practice, the integration and analysis of multiple data sources coming from different organisations becomes even more challenging.

For sure, ChildRescue has a difficult task lying ahead: The modelling of missing children and unaccompanied migrant minors' behavioural patterns, taking the most out of available data sources. Using this profiling model, estimations and predictions regarding the child's whereabouts can be made. If successful, a missing child can return home, to its parents, in a faster, safer and more reliable way. The sections to follow are devoted to this exact mission.

3 Methodological Approach for Multi-source Analytics in ChildRescue

ChildRescue is a project with social impact that aims to benefit children and their families, relying on new technologies and modern data processing techniques. Therefore, one of the project's core objective is the efficient, and scientifically sound, analysis of data. Data coming from various sources and usually containing sensitive pieces of information. In this section we are going to investigate what are, or can be, the various types of data sources for ChildRescue, how they can be combined and on what type of tasks the selected algorithms can be applied, so as to produce useful information and help decision-making for all our pilot cases.

Data analytics in ChildRescue could be viewed as two separate processes with discreet features. The first process involves datasets that are already existing at the beginning of an investigation, and takes place during the PROFILING phase, while it is supported by the ARCHIVING procedures. The second process analyses, in a more dynamic fashion, incoming information from the social sensors (i.e. citizens sending informative messages or pieces of evidence through a mobile app) and can be considered as a continuous effort, involving mostly spatiotemporal data. It lasts for the duration of the ACTION phase, assisting also in the COLLABORATION operations.

In order to conceptualise, develop and adapt the multi-source data processing methodology, the following 6-step approach was adopted that spans the foundations building, the pilot cases analysis and the specifications design axes, as follows:

- Setting up the possible sources and tasks: During this preparatory step, the scope of the data acquisition and processing framework and thus, the scope of the deliverable at hand was discussed and agreed.
- Studying the state-of-the art: Following the definition of the overall scope, a state-of-the-art analysis was carried out in order to screen the landscape along 3 dimensions, namely: Human Profiling, Spatiotemporal Data analysis and Social Media Analytics. The research perspective was explored in order to extensively review and classify existing approaches / methodologies and to form a thorough listing and comparative analysis of the most popular algorithms in all the aforementioned high-level areas. Based on the state-of-the-art analysis conducted, useful conclusions on each of these areas were drawn while open issues and gaps were effectively identified.
- Iteratively identifying the data sources available in the pilot cases and discussing with the pilot partners to get a better understanding of the data essence.
- Preparing the data templates in order for the ChildRescue pilot cases to start recording in a consistent way the new data, and deliver historical data and any other operational data in a uniform manner.
- Designing a high-level ChildRescue data model incorporating information from the state-of-the-art analysis and the ChildRescue specific pilot cases and data requirements in order to create a common understanding in the consortium.
- Elaborating on the data analysis along with the data manipulation and processing perspectives in order to bring forward the preliminary ChildRescue considerations, requirements and constraints for the next steps of the project implementation.

Data sources

The data sources available from pilot partners include the archived documents of past cases for missing children, unaccompanied migrant minors' records and tracing requests. Digital files, such as filled-in forms and photos, as well as hand-written documents, comprise the usual set of a case file. Some of these cases contain a link to a social media account.

Data Templates

The purpose of these templates is to present a set of required data fields (data schema) that could be employed by the recommended methods and algorithms in order to produce useful insights and predictions. We considered two types of templates: The Profiling template and the Events template (in order to collect *Profile data* and *Events data*, respectively).

The development of the Profiling template was determined by two main factors: The available formats of data structuring and archiving already utilised by pilot partners, and the results of the research landscape analysis of Task 2.1 and Task 2.2 of WP2. In other words, we compiled an initial list of 96 features with the data fields the pilots already use and the fields we additionally want.

The next step was to ask the pilot partners to look into their archives and specify which of these data fields always contain some form of information for all cases (mandatory), in some cases (optional) or in none (unused). Finally, we resulted in a more flexible list, limiting the field number to 40 features, according to the replies of the partners and the expected contribution of each feature to the data analytics. The outcome of this process is presented in Table 3-1.

Table 3-1 List of data fields included in the Profiling Template

Name of field	Type of reply	Importance	Category	Possible Values/List (examples)
Case ID	Alphanumeric value	Mandatory	Case data	e.g. IA23891023D
Area/Location child was last seen	Text value or coordinates	Mandatory	Case data	e.g. "Around Limassol castle, Limassol"
Date and time the child was found	Date time	Mandatory	Case data	e.g. 15/08/2017 11:30
Date and time of disappearance	Date time	Mandatory	Case data	e.g. 16/07/2017 23:00
Conditions of disappearance	<i>Select 1 from list</i>	High	Case data	On way to school; Returning home from activity; Was out with friends; visiting a friend; on trip to another city;
Possible reasons of disappearance	<i>Select multiple from list</i>	High	Case data	argument with family; victim of school bullying; argument with boy/girlfriend; visit friend in another hosting facility; in search for relatives; recently moved to new hosting facility;
State of child when found	<i>Select multiple from list</i>	High	Case data	abused; normal; shocked; dead; wounded; etc.
Carrying mobile phone	Yes / No	High	Case data	Yes / No
Carrying money or credit card	Yes/No	High	Case data	Yes / No
Has area knowledge	Yes / No	High	Case data	Yes / No
Rescue teams utilised	Yes / No	Low	Case data	Yes / No
Volunteers utilised	Yes / No	Low	Case data	Yes / No
Transit country/-ies	Separate	Low	Case data	e.g. Turkey; Greece; Albania

reached or intended to be reached	values using semicolon [;]			
Date of arrival at hosting facility	Date time	Low	Case data	e.g. 16/07/2017 23:00
Type of disappearance (Category)	<i>Select 1 from list</i>	High	Case data	runaway; parental abduction; criminal abduction; lost/injured/otherwise missing; missing U/A minor; tracing request; unclear;
Multiple-times case	Number	High	Case data	2
Family members	Number	Medium	Case data	5
Probable destinations (location/city/country)	Separate values using semicolon [;]	Medium	Case data	friend's house in Brussels; Music event in Apollon Limassol stadium;
Clothing with scent (for dogs)	Yes / No	Medium	Case data	Yes / No
Home / Facility Address (area only)	Text value or coordinates	High	Demographics	e.g. Ritsona Refugee Camp, Vathy, Euboea, Greece
Education level & current participation in educative activities	<i>Select 1 from list</i>	Low	Demographics	first grade; second grade; third grade; none; unknown;
Languages spoken (number of)	Number	Low	Demographics	e.g. 1
Nationality	<i>Select 1 from list</i>	Low	Demographics	e.g. Syrian
Place of birth / Country of origin	<i>Select 1 from list</i>	Low	Demographics	e.g. Syria
Age (or Birthday)	Number (Date)	High	Demographics	e.g. 15 (or 12/9/2003)
Gender	<i>Select 1 from list</i>	High	Demographics	Male/Female
Health issues (current)	<i>Select multiple from list</i>	High	Medical Profile	allergies; substance abuse; diabetes; asthma; heart disease; pregnant; etc.
Medical treatment required?	Yes/No	High	Medical Profile	Yes / No
Social media accounts	Link(s) separated with semicolon or no link	High	Personality/Social Profile	https://www.facebook.com/georgeorge12121
Protection concerns/Vulnerabilities	<i>Select multiple from list</i>	High	Personality/Social Profile	child headed household; disabled; medical case; street child; girl mother; living with vulnerable person; abuse situation; trafficking/exploitation risk; early marriage; other;
Specific personality characteristics/psychological disorders	<i>Select multiple from list</i>	High	Personality/Social Profile	antisocial, suicidal, autistic, depressive, schizophrenic, other mental or emotional disorders
Family situation	<i>Select 1 from list</i>	High	Personality/Social Profile	(living with both biological parents, living with single parent (divorced, separated, widowed, other), living under relative's care, living in foster care, living in hosting facility/other institution, separated child, unaccompanied child)
Parents' (Tracing enquirer) profile	<i>Select 1 from list</i>	High	Personality/Social Profile	Excellent; Good; Sufficient; Not good; Really Bad;

evaluation				
School grades, Absences	<i>Select 1 from list</i>	Low	Personality/Social Profile	A' student with zero absences; A' student with normal absences; A' student with excess num of absences; B' student with zero absences; B' student with normal absences; B' student with excess num of absences; C' student with zero absences; C' student with normal absences; C' student with excess num of absences;
Interests/Hobbies	<i>Select multiple from list</i>	Medium	Personality/Social Profile	music; football; basketball; dancing; painting; singing; etc.
Relationship status	<i>Select 1 from list</i>	Medium	Personality/Social Profile	single; in a relationship; married; its complicated; other;
Religion	<i>Select 1 from list</i>	Medium	Personality/Social Profile	Orthodox Church; Catholic Church; Protestantism; Other Christian; Islam (Sunni); Islam (Shiite); Other Islam; Hinduism; Buddhism; Nonreligious; Other;
Weight	Number	Low	Physical data	e.g. 42
Height	Number	Low	Physical data	e.g. 140
Distinguishing features	<i>Select multiple from list</i>	Medium	Physical data	tattoos; scars; skin marks; missing teeth; front teeth gap; body piercing; other;

As one can observe, the fields required are divided by type, importance and general category. The type of reply is related to the technical implementation of the respective field and the expected way the data will be acquired by the ChildRescue platform. Where a *List* is denoted, its contents will be populated by the aggregated values existing in the compiled past cases, along with the suggestions of experts. The importance is an arbitrary estimated value for the respective data field's contribution to data analytics. The categorisation column groups the fields of similar characteristics into 5 classes.

For the Events Template, things were more straightforward. We created a template that records the sequence of events (spatiotemporal input) and incoming proofs of evidence, starting from the event of the disappearance (place and time the child was last seen) until his or her finding.

Table 3-2 List of data fields for the Events Template

Name of Field	Example value
Step ID	E.g. 1,2,3...
Events in chronological order / Tracing steps (Date time)	E.g. 12/07/16 18:30
Location child was seen	E.g. Panepistimio metro station, Athens, Greece
Transportation means the child used to reach location	E.g. On Foot, Subway, Bus, unknown, etc.
Evidence True?	E.g. TRUE/FALSE
Reasons for location selection	E.g. Relative's home, there is a subway station, etc.

Child Physical status	E.g. Clothing have changed, Had a hair-cut, etc.
Extra info	E.g. Child was accompanied by middle-aged woman

Data Model

The data coming from multiple sources need to end up into a form of data model that completely describes what ChildRescue is about. From the development and compilation of the data templates, it soon became evident that the core elements of a high-level model should be three: The Case Profile, the Child Profile, and the Events Log (Figure 3-1).

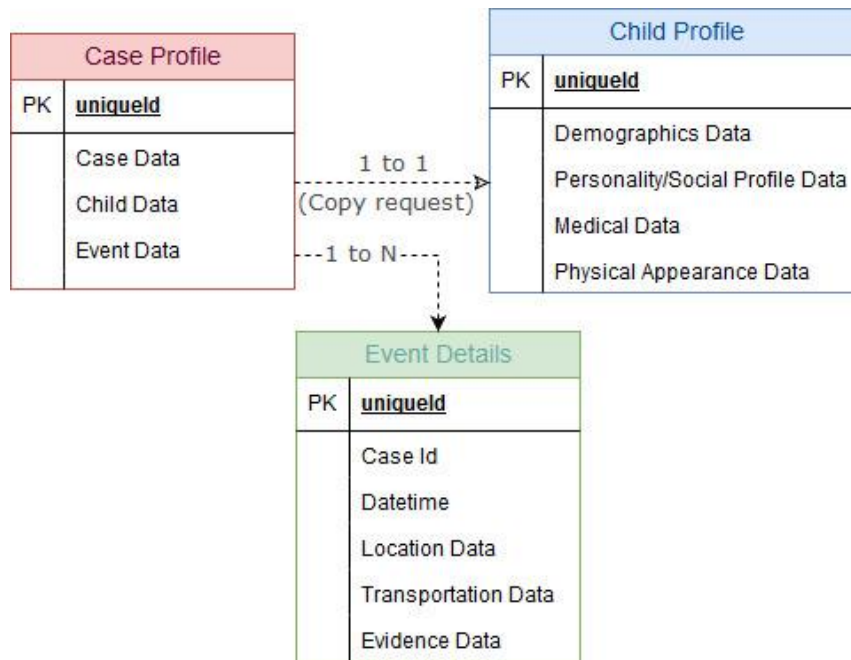


Figure 3-1 The proposed preliminary Data Model for ChildRescue

The main entity of our model is the Case Profile. The Case Profile includes case data, related to the current case, child data, that are transferred to the case from the current (i.e. updated) profile of the child, and a list of Events. Each Event includes, at least, a date and a time of the event, location data, transportation data and details on the related piece of evidence.

The Child Profile includes demographics, personality traits, medical and physical data, and maintains the up-to-date version of the child (where, in comparison, the case data keeps the child version of the time the case took place). The Child Profile transfers a copy of itself if and when requested by the Case Profile (i.e. when a new case is opened for an existing, in the database, child. Otherwise a new child profile should be created in advance). In addition, when a case involves more than one child, say x children, then x Case Profiles should be created so that each child is tracked separately.

Following this practice may lead to the same information being kept multiple times, but this is a small sacrifice to pay, if we aim to cover each case thoroughly and correctly. After all, due to the very nature of the project, it is the first priority of ChildRescue to propose and follow a rigorous and sound methodology of data analysis which produces robust, accurate and secure results. Other factors, such as the computational speed or cost of resources, although important, come at a second place.

Data Analysis

The last step of the proposed methodology is concerned with the data manipulation and analysis, in order to bring forward the preliminary ChildRescue considerations, requirements and constraints for the next steps of the project implementation. This procedure involves several stages, from data ingestion and transformation to the actual data analysis and the visualisation of the results.

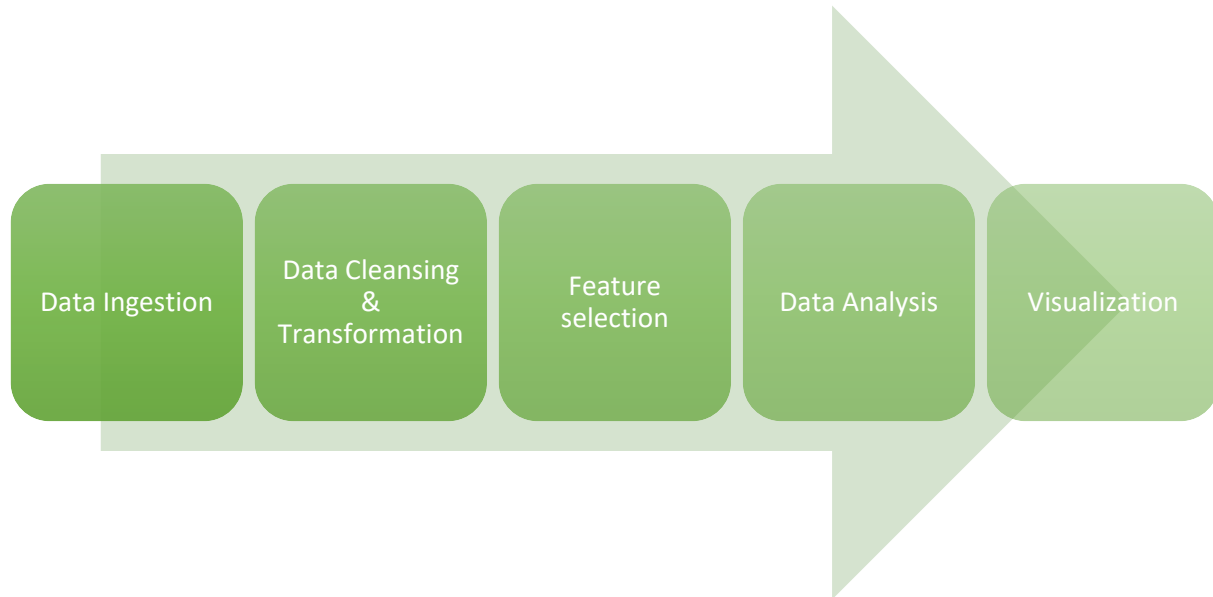


Figure 3-2 A simple data analytics flow

Judging from the data available and the knowledge we can infer from them, we conclude that there can be three types of data analytics that can be performed within the context of ChildRescue. These are depicted in the following table, with each one being further detailed in the lines to follow.

Table 3-3 Relation of ChildRescue investigation phases and Analytics types

Analytics type	Investigation Phase	Data sources
Predictions based on Behavioural and Activity Profile Data	PROFILING, ARCHIVING	Past cases profile data, Current child profile, Social media
Evidence Analysis and Evaluation	ACTION, COLLABORATION	Evidence data, User profile data, Past cases events data
Real time Route/Destination Estimation	ACTION, COLLABORATION, PROFILING	Past cases data, Evidence data, Open data (transportation, events, etc.), Linked data

3.1 Predictions based on Behavioural and Activity Profile Data

As described in section 2.1 of the present document, profiling is all about predicting future behaviours of an unobserved individual based on aggregated knowledge from a large group of observed individuals. According to ChildRescue workflow (see deliverable D1.3), during the PROFILING phase, information coming from multiple sources is collected and refined so as to create a more complete profile. This can be applied on both the missing children and the unaccompanied migrant minors' cases.

Using the concept of our proposed data model, our definition of profile requires all the elements that comprise the specific case, as well as the child's details. Once we have the individual's profile data, we can then compare it with the aggregated set of past cases and discover hidden patterns and correlations.

For any personality attributes that are missing or not yet acquired, we can take advantage of social media analytics and retrieve them, in the same fashion as described in [107].

More specifically, the objectives of this procedure should be:

- 1) To assess the profile model of the case-child (descriptive analysis),
- 2) To extract missing information using social media analytics,
- 3) To employ behavioural predictive analytics based on past cases.

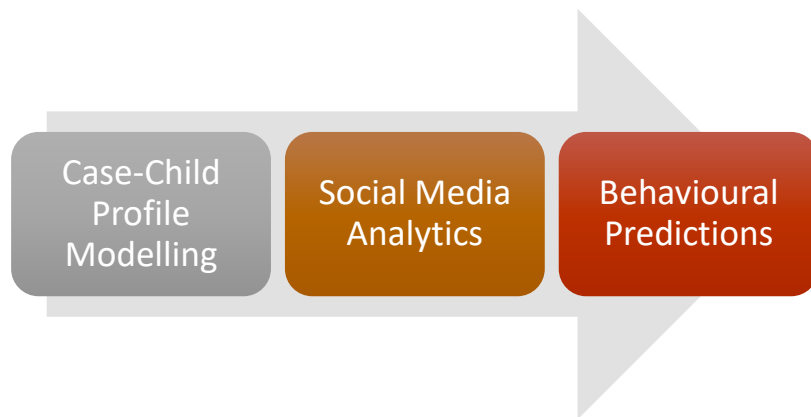


Figure 3-3 Behavioural Prediction process

The behavioural prediction concept could be further broken down into more specific tasks, examples of which can be:

- 1) The classification of a new profile to a Case category,
- 2) The clustering of behavioural characteristics, either on child or case level, or
- 3) Personality classification, based on the big-five model.

If the social media account of the missing child is not confidential (for anonymity reasons), we can also enrich the profile with his or her social account preferences, activities and networking circle. However, an obstacle to this procedure would be the fact that we would need to acquire the same pieces of information for the archived cases, as well.

3.1.1 Possible sources

The data required for the profile assessment and pattern extraction can be derived from the previous investigation cases and, optionally, from social media provided data. For this purpose, each ChildRescue pilot partner was asked to compile a number of past cases from their archives, in order to modify their data according to the Data Template presented in the beginning of this section. The complete results of this strenuous process will be documented in detail in deliverable D2.4.

In Annex III: Past cases list of reference”, a sample of 122 past cases, compiled by all pilot partners, is presented.

The requirements for every case of this sample were:

- to hold as much information as possible,
- to be relatively recent, and

- for the sample itself, to be statistically representative of the Case categories.

The list contains 122 past cases, out of which 8 are still open. All cases refer to actual events that took place in the last few years (since 2009 and onwards) with the majority being in the last 4 years (since early 2015). This dataset will be employed as input to train the appropriate algorithms in order to create a profiling model. The resulted model will then be used to assess any new profile. The optional use of social networks data, as already mentioned, has to do mostly with the prediction of the child's missing attributes, like religion, relationship status, etc.

The individual's social data and networking connections would be also of great value to the profiling assessment, in a sense that many hidden traits and habits could be located and utilised, that are otherwise very difficult to obtain (see for example [84] and [85]).

The table below presents the data sources required by this part of the methodology.

Table 3-4 Data sources to be used for profiling

Type	Data sources
Mandatory	Past cases data, current Child Profile
Optional	Social media data (aggregated)
Optional/Privacy issues	Social media data (individual profile)

3.1.2 Methods and algorithms

In this section we present the methods related to profiling predictive analytics. The three objectives mentioned, namely profile modelling, social media analytics, and predictive analytics, usually utilise data mining techniques and computational learning algorithms.

The following table includes all algorithms and methods encountered in literature analysis, that in one way or the other, can effectively deal with the ChildRescue profile data and play an integral part in predictive analytics. The specific tasks are yet to be determined since the actual data availability, structure, and quality are not finalised at this preliminary stage.

Table 3-5 Summary of algorithms for profile modelling and predictions

Objective	Related Family of algorithms	Popular algorithms
Profile modelling	Clustering, Classification, Correlation analysis	K-means SVM, k-NN, decision tree, Random Forest Ordinary Least Squares regression
Social media personality prediction	Classification/Regression, Linguistic Analysis	SVM, k-NN, Random Forest NLP techniques
Behavioural Analysis	Classification/Regression	SVM, k-NN, Decision tree, Random Forest, Naïve Bayes
Sensitivity Analysis	Linguistic Analysis, Classification	NLP techniques Decision trees (C4.5), Naïve Bayes, SVM

3.2 Evidence Analysis and Evaluation

In a broad scope, evidence can be defined as anything presented to support an assertion. In ChildRescue, pieces of information are accumulated and assembled to support a missing child's investigation process during the ACTION phase (see deliverable D1.3). In other words, it is the evidence collection and analysis that construct, piece by piece, the tracing path towards the missing child.

The incoming pieces of evidence may include errors, or can be completely false, either accidentally or on purpose. One piece may hold truth in respect to location, but have the time wrong, while another can be 100% correct, but it is sent by an anonymous, and thus, less trustworthy, user.

It is, therefore, a crucial task for ChildRescue to be able to rapidly and reliably analyse the data coming from exterior sources (i.e. citizens, also known as "social sensors" in this project) in an intelligent and efficient manner.

The main goal of the task is to assess the quality and reliability of the incoming information, so that the extracted data offer valuable assistance in the investigation for a missing child and do not mislead or otherwise, hinder the whole process.

More specifically, the objectives of the evidence analysis should be:

- 1) To assess the quality of evidence
- 2) To assess the credibility of the sender
- 3) To weigh the value and reliability of one particular piece of evidence against another or the credibility of a sender against another (contradicting information)
- 4) To validate a piece of evidence through some form of collective intelligence (i.e. Crowdsourcing)

In this framework, ChildRescue should adopt a clear methodology of evaluating various forms and types of evidence by analysing not only the content of a message but also the source. In other words, in ChildRescue evidence evaluation implies the analysis of both the provider and the information provided. In our opinion, this evaluation process should involve three steps:

- 1) User evaluation through his/her historical behaviour
- 2) Content evaluation, through cross-checking with similar (in time and space) information
- 3) Evidence validation, through Crowdsourcing, by sending relevant info to the appropriate and eligible receivers (usually in order to verify or dismiss it)

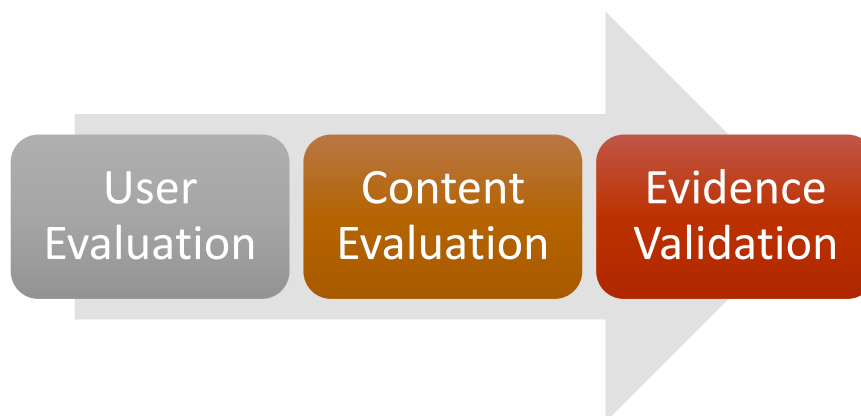


Figure 3-4 Evidence Evaluation process

In the following subsections, we are going to describe thoroughly these steps, by specifying the sources and the means that can achieve as much accurate evaluation of evidence as possible. The theories of Crowdsourcing and collective intelligence will also be presented as they will play a key part in the validation process.

3.2.1 Possible sources

In the case of evidence, ChildRescue primary contributors will be the active “Social Sensors”, i.e. all the users, registered or not, who have downloaded the ChildRescue mobile application and wish to contribute to the investigation process by sending proof of evidence regarding the case.

The incoming information can be of several types and formats. Text messages, images, videos, voice recordings and geolocation data are the most common types supported by modern applications. In ChildRescue we will consider all of these types as possible evidence, albeit not all of them can be used by algorithms.

The data that will be used as input for the evaluation process will contain, apart from the submitted evidence, its accompanying metadata (timestamp, device details, geolocation, etc.). The results of the predictive algorithms (i.e. the calculated POIs) and current case information, along with the user’s profile history, will be combined with the evidence in order to provide a reliable and accurate evaluation.

Table 3-6 Data sources to be used for evidence evaluation

Type	Data sources
Mandatory	Text message, user geolocation, evidence timestamp and location, user profile history
Optional	Image, Video, voice recordings, User profile details, User device details, Past cases related info

3.2.2 Methods and algorithms

In this section we present the methodological approach towards evidence analysis. The three steps recognised previously, are further analysed and matched to known algorithmic tasks. The algorithmic families along with widely used algorithms that will be considered for the implementation of the evidence evaluation are presented in Table 3-7.

User evaluation through past behaviour

The expected user evidence will be received mainly through the ChildRescue mobile app. The app users are graded in different groups with different access to case data, different permissions and functionalities. All of the above-mentioned users will be able to submit evidence. It is obvious however, that this evidence could (and should) not be equally ranked, as data coming from a trustworthy, authenticated source should generally be prioritised.

User history is another important factor when considering user credibility, as a user who has previously provided valid leads is considered more trustworthy. For this evaluation to happen, it is necessary for the system to be able to assess the provided evidence after closure of the case (with human expert intervention). Apart from user history, the profiles are characterised from other attributes also, such as demographics, place of residence, etc.

When it comes to new users without a history of actions it becomes more complex. In order for their evidence to be evaluated, we could utilise clustering algorithms, which will enable us to “predict” the user evaluation factor before she/he has even provided evidence. The clustering could be performed on user profiles and the “prediction” could be based on the previous performance of similar users.

However, because demographics do not usually suffice for extracting safe results, we should also consider applying a fixed cold-start credibility value for every newcomer. After having enough history data, we then can proceed with the ranking of this user. An issue for further consideration on this approach, is the starting point: should we consider new users as having the highest credibility and maintain or lower it according to their provided evidence, or should we set the default credibility for a newcomer to zero and let him/her build piece by piece the required trust with the system? In our opinion, the European culture strongly favours the former approach and this is what we recommend: every user should be by default trustworthy.

Content evaluation, through cross-checking with similar (in time and space) information

Content evaluation consists of finding contradictory elements inside the uploaded evidence, or in relation to other case information, which has already been evaluated and is considered valid. These elements could concern both time and space and indicate a possibly incorrect and misleading piece of evidence. For example: a user states that she/he witnessed the child half an hour ago and pinpoints the location of last seen. However, the geolocation of the user, which is also included in the submitted information, shows that she/he is in another city, and her/his statement of witnessing could not be physically feasible. In cases such as the above, the submitted evidence will not be filtered or excluded. They will however receive lower ranking and be appropriately tagged.

Contradicting evidence could be perceived as data differentiating from the expected results (outliers). An outlier is a data pattern not conforming to the expected behaviour and anomaly detection is the locating of such spikes in a data set [108]. Thus, an analogy could be drawn between contradicting evidence and spatiotemporal anomaly detection, a rather well-studied field in the literature. In our case, the expected data values could be calculated by factors such as the location the missing person was last seen, the expected location, etc. Nevertheless, such algorithmic approach requires a sufficient amount of data to be able to discern anomalies. For instance, if there are only two patterns contradicting each other, then each one sees the other as outlier and are both right (or wrong).

Evidence validation, through Crowdsourcing, by sending relevant info to the appropriate receivers (usually in order to verify or dismiss it)

Following the positive aspects of the collective intelligence, which will be briefly presented below, a Crowdsourcing process will take place in order to validate the evidence at hand. The power of the crowd will be exploited in two ways. On the one side, the active participation of the ChildRescue users will be needed. The examined evidence (part or the whole of it) will be disseminated for verification to the legitimate receivers. The extent of dissemination, location and content will be defined accordingly. It could vary from a simple alert to all mobile app users nearby the evidence location, for activation of the local community, to the uncensored evidence being sent to authorised volunteers and S&R members, who will consecutively go to the indicated place for self-inspection and verification or dismissal of the information.

The second step of the validation could be an automated process and thus, active validation from the users would not be required. Indirect confirmation of the evidence through a clustering algorithm can take place in such case. Text analysis and location proximity among submitted evidence could be the

major similarity factors used for the clustering. If a considerable quantity of similar evidence is accumulated, then it should acquire a higher validity factor.

Table 3-7 Algorithms related to Evaluation Steps

Objective	Related Family of algorithms	Popular algorithms
User evaluation (profiling)	Clustering, Classification	K-means SVM, k-NN, Random Forest
Anomaly detection	Regression/Classification, Stochastic analysis, Linguistic Analysis	SVM, k-NN, Random Forest Markov models NLP techniques
Crowd sourced Evidence validation	Clustering of incidents	K-means, Non-negative Matrix Factorisation

Collective Intelligence

Collective intelligence is the accumulated and shared intelligence resulting from the collaboration of a loosely organised group towards a purpose. An expression of collective intelligence is through Crowdsourcing platforms, like Wikipedia or iStockphoto⁴, where a call for individuals' contribution (regardless of their expertise) leads to a remarkable result. According to Surowiecki [109], a successful solution often emerges from a large basis of people. It is also stated that the aggregated ideas contributed by a large group can exceed in intelligence those of the smartest individuals. Except for content uploading, another example of a popular collective intelligence harnessing method is by explicitly asking the users to rate or vote for an item [110].

In [111], which also addresses the problem of information correctness assessment in social-sensing applications, we also read about the collective response to a social purpose, varying from the reporting of potholes to the participation in rescue missions. The collective behaviour response to disasters has also been identified and studied previously [112]. Considering the instantaneity of information diffusion through the internet and mobile phones, combined with the active response of citizens to numerous emergencies or disasters, as for example in the Virginia Tech Shootings [113], we can see the emerging potential for ChildRescue.

Crowdsourcing in ChildRescue will not be limited to just asking for and gathering potential evidence from users. It will also be utilised in the evidence evaluation process, were the collective intelligence will be extracted as was described previously in the methodology (both directly and indirectly).

3.3 Real time Route/Destination Estimation

The task of route and destination estimation was encountered in literature by the name "trajectory prediction" and "POI recommendation", respectively. Actually, this task is the most complex of the three tasks assigned to ChildRescue analytics and the reason is threefold: Firstly, we have to deal with multiple heterogeneous sources of data; secondly, these data are both spatial and temporal in

⁴ <https://www.istockphoto.com/>

nature; and thirdly, ChildRescue does not rely on using mobile phone or GPS log data since a child that has gone missing usually does not carry a cell phone, and even then, the call data or GPS logs are not accessible.

In particular, the objectives of this complicated process should be:

- 1) To use profiling information and extract an initial set of related POIs for the child.
- 2) To predict next POIs or locations the child will visit based on dynamically assessed information from verified evidence and transportation data.
- 3) To simulate and predict possible movement trajectories (routes) of the child.

Because of the apparent low probability to locate a moving individual within a city's limits or within a country using the aforementioned data sets, the actual purpose of this task is not so much to locate the child, but to estimate a centre (POI) and a radius of an action circle where the available Social Sensors, volunteers and/or rescue teams will receive appropriate notifications to act.

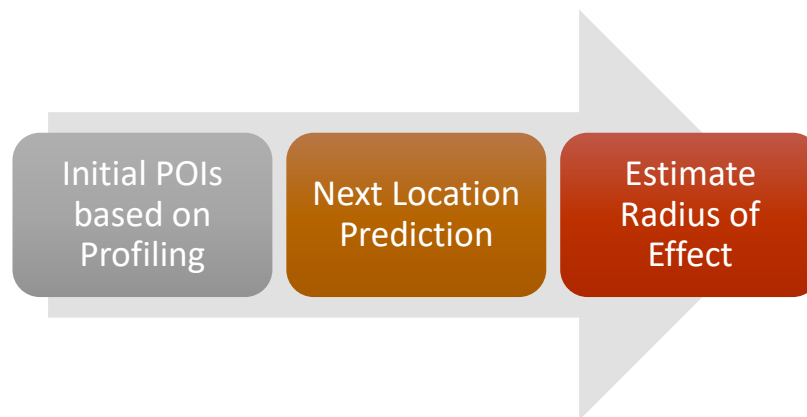


Figure 3-5 Route estimation process

On the bright side, machine learning techniques, as well as recommendation systems, possess enough power to tackle these issues successfully and produce useful indicators on the child's general whereabouts, as long as the data deriving from these multiple data sources is of sufficient quality. In other words, this is the task where the data sources come to play the most significant role.

3.3.1 Possible sources

The available data sources can be divided into two groups: static ones, like the past cases with their respective spatiotemporal events, open data, or social network preferences, and dynamic ones, which mostly have to do with the incoming geolocated evidence in real time. The former is analysed during the PROFILING phase with the sole purpose to create a number of initial predictions for POIs or possible destinations. During the ACTION phase, the spatiotemporal information collected, can be utilised to recommend next possible locations and form route trajectories based on stochastic methods.

The data sources that will be used as input for the recommended methods are summarised below.

Table 3-8 Data sources that will be used as input for the recommended methods

Type	Data sources
Mandatory	Past cases events data, place the child was last seen, evidence suggested locations

Optional	Open data (Weather, social events, transportation routes & infrastructure), social media personal preferences, social media status posts
----------	--

The events recorded in a previous case, as already explained, should follow the past cases Event Template, which includes, among other information, the date and time of the event, the location, and any used transport. Apparently, for the training of any algorithm we consider only the verified (to be true) events.

3.3.2 Methods and algorithms

The objectives of this task can be viewed as common undertakings in the research community that deals with spatiotemporal data. POI recommendations and trajectories predictions have been a subject of study for numerous publications. However, in the ChildRescue case, the surrounding conditions are not that favourable.

Recommending a point of interest or predicting a route, is quite challenging when the primary dataset you base your analysis on, does not possess this kind of information. In that case, the influence of external data sources becomes crucial and making the most of state-of-the-art methods becomes a necessity.

Under these circumstances, the proposed algorithms attempt to cover all possible objectives. The location recommendations, either initial or later, should employ various techniques and select the most efficient. Even ensembles can be of use. For the trajectory prediction, the utilisation of probabilistic models, combined with clustering methods, seems to be the only route.

Ultimately, depending on the resulted routes or POIs, machine learning algorithms should estimate the centre and radius of influence that will infuse geofencing techniques with input attributes.

The proposed sets of algorithms, divided by objective, are summarised in the next table.

Table 3-9 Methods and algorithms for route/destination estimation

Objective	Related Family of algorithms	Popular algorithms
Initial and next POI recommendations	Recommenders, User mobility modelling, Clustering, Classification, Deep Learning	Collaborative Filtering Markov models and decision process Non-negative Matric factorisation Decision tree, Random Forest, ANNs, Naive Bayes, SVM, Fuzzy expert systems RNNs (e.g. LSTM)
Route (trajectory) prediction	Probabilistic modelling, Clustering	Markov models Density-based clustering algorithms (e.g. OPTICS, DBSCAN), Voronoi partitions
Radius estimation	Regression	Decision tree, Random Forest, ANNs, SVM

Geofencing

The widespread use of smartphones equipped with several modern sensors allows for the detection of the user's current physical activity or the user's presence in a designated area. The latter is often referred to as Geofencing [114]. Geofencing technology is a location-based service, mostly encountered in mobile phone applications, that allows the sending of notifications to users who enter or exit a specified geographical area. As such, this service has become very popular among today's mobile marketing strategies and smart urban environments [115].

In order to specify a geographical area for this mobile sensing service, one has to designate a centre and a radius. The resulted circle functions, more or less, like a fence, triggering an event when a mobile user (with the appropriate application) enters or exits the area. This is the reason it is required by our methodology to be able to infer the circle's attributes.

Geofencing technology, or similar techniques, will be adopted by the ChildRescue platform in order to send location-based notifications to registered users, citizens or volunteers. Since it is a well-founded technology, more details on its technical aspects and how this is going to be implemented in the framework of the communication functions of ChildRescue will be examined in the related WP3 Tasks.

3.4 Discussion & Limitations

In this section, we proposed a multi-source analytics methodology based on the knowledge extracted from the state-of-the-art review, properly adapted to facilitate the special data properties and characteristics of ChildRescue. A step-by-step procedure was established first, to explicitly describe all the required actions that can lead to a successful data analytics strategy. The data analytics are, then, broken down into three separate approaches, each one dealing with multiple sources of data.

All three of them derive from the examined literature domains, but with focus on predictive analytics based on human behavioural profiling, evidence evaluation and validation, and the geospatial data analysis required to estimate points of interest and possible routes. These three fields of analytics compose the core of the mechanism that we call ChildRescue multi-source data analytics. A common data source in all three methods is the past cases archive of the pilot partners, and this is why a specific data template design was clearly defined from an early stage, so that we can collect these datasets in a structured and uniform fashion.

All the methods and algorithms presented in sections 3.1, 3.2 and 3.3 will be tested and comparatively assessed on the basis of the data to be provided by the pilot partners. The most suitable ones will serve as the algorithmic basis in the relevant components of the ChildRescue platform.

The major challenges expected to emerge during the ChildRescue project, in respect to data collection and analysis, can be summarised as follows:

- Heterogeneity of data sources available at the different pilot partners (data types, naming conventions, different languages, etc.)
- Currently, most information regarding the psychological profiles is maintained in large text files. In some cases, the full content is not even digitised.
- Past cases may have missing details in several aspects of the data profile.
- The spatiotemporal data analysis, as estimated given the availability of data, will be hardly able to predict real trajectories and routes efficiently.
- Data analysis privacy issues, since ChildRescue is expected to handle personal, and sometimes, sensitive, data.

- Crucial GDPR issues with social media. At the time of this writing, the EU regulation concerning data privacy has been applied to all data owners/providers in a global scale. The most popular social media platforms (e.g. Facebook) have decided to be very strict about enforcing the privacy regulations and access to their data is now very limited.

For each of these limitations, a contingency plan was considered. The table below presents the limitations, characterised by their respective probability of occurrence, along with the recommended contingency plan.

Table 3-10 Summary of data collection and analysis limitations

Challenge / Limit	Probability	Contingency Plan
Heterogeneity of data sources available at the different pilot-partners legacy systems (data types, naming conventions, language, etc.)	High	Generic data model representation of the different data sources;
Most historical information is unstructured in text format (e.g. word documents)	High	Manual transformations; descriptive analytics.
Complex spatiotemporal data processes that may lead to low performance trajectory prediction	High	Focus the analysis on estimating approximate values for centre and radius of a circle, instead of estimating the exact route followed by a missing child.
Privacy issues	Medium	Use of anonymisation techniques on data. Use of algorithms and methods able to cope with anonymised data.
Cases with missing information	Medium	Social media analysis approach to retrieve missing values based on profiling; Methods for handling missing values.
Limited social media data	Low	Examine many different social platforms and combine data.

The ChildRescue architectural design and implementation should reflect on these limitations, as well as the suggested workarounds, and ensure that the respective components can successfully cope with any, data-related, challenge.

4 Conclusions & Next Steps

The focus of this deliverable has been to lay the background for the multi-source data analytics methodology that will be applied in ChildRescue. Initially, we setup the main scenery by identifying the main domains of interest: Computational Learning in Human Profiling, Exploiting Geospatial Data coming from multiple sources, and Social Media Analytics. An extensive state-of-play analysis across these domains produced methods and technologies for processing behavioural profiles, mobility patterns, and social networking data. Several academic papers were reviewed and compared, presenting the most significant aspects of the three fields.

The following step was to define the foundations for a concrete multi-source analytics methodology. A robust and uniform way to collect data from past cases maintained by the pilot partners was initially designed, which helped us identify the actual datasets available. A preliminary data model was derived from these data, with the intention to conceptually assist in the implementation of the ChildRescue platform. The data model is expected to continuously evolve and grow and be subjected to the project's advancements and the feedback of the end-users. The tools that can manipulate and analyse the data available, were then proposed. They were divided into three main ChildRescue axes: the predictive capabilities of profiling data, the evidence evaluation and validation, and the estimation and forecast of routes and points of interests one might want to follow and visit. Using all of these tools successfully, in the framework of ChildRescue, will certainly lead to a more efficient investigation process of a missing child or the tracing of an unaccompanied migrant minor.

The data model and related methodology will play an important role in the tasks of "WP3-ChildRescue Platform Architecture Definition and Implementation". Especially in Task 3.1, during the design of the platform's architecture, the recommended methods and algorithms show that several avenues regarding some well-documented good practices and design patterns can be considered and, eventually, adopted.

The results reported in this deliverable, together with the results of D2.1 and D2.3, will be aggregated to form a complete methodology in deliverable "D2.4 - Profiling, Analytics and Privacy Methodological Foundations, Release I" in M12. Any modifications up to that point, such as an updated version of the data template, will be documented as well.

The final, incorporated, methodology, along with the initial input and provided feedback or updates from pilots' experiences will be presented in deliverable "D2.5 - Profiling, Analytics and Privacy Methodological Foundations, Release II" on M24.

Annex I: References

- [1] Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5), 662-679.
- [2] Deka, G. C. (2016). Big data predictive and prescriptive analytics. In *Big Data: Concepts, Methodologies, Tools, and Applications* (pp. 30-55). IGI Global.
- [3] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
- [4] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137-144.
- [5] McCue, C. (2014). *Data mining and predictive analysis: Intelligence gathering and crime analysis*. Butterworth-Heinemann.
- [6] Jonas, J., & Harper, J. (2006). *Effective counterterrorism and the limited role of predictive data mining*. Washington DC: Cato Institute.
- [7] Perry, W. L. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.
- [8] Pramanik, M. I., Lau, R. Y., Yue, W. T., Ye, Y., & Li, C. (2017). Big data analytics for security and criminal investigations. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(4), e1208.
- [9] Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11), 1134-1142.
- [10] Angluin, D. (1992, July). Computational learning theory: survey and selected bibliography. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing* (pp. 351-369). ACM.
- [11] Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.
- [12] Afsar, P., Cortez, P., & Santos, H. (2015). Automatic visual detection of human behavior: a review from 2000 to 2014. *Expert Systems with Applications*, 42(20), 6935-6956.
- [13] Atallah, L., & Yang, G. Z. (2009). The use of pervasive sensing for behaviour profiling—a survey. *Pervasive and Mobile Computing*, 5(5), 447-464.
- [14] Pantic, M., Pentland, A., Nijholt, A., & Huang, T. S. (2007). Human computing and machine understanding of human behavior: a survey. In *Artificial Intelligence for Human Computing* (pp. 47-71). Springer, Berlin, Heidelberg
- [15] Jin, L., Chen, Y., Wang, T., Hui, P., & Vasilakos, A. V. (2013). Understanding user behavior in online social networks: A survey. *IEEE Communications Magazine*, 51(9), 144-150.
- [16] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering*, (6), 734-749.
- [17] Brooks, R. A. (1991). Intelligence without representation. *Artificial intelligence*, 47(1-3), 139-159.
- [18] Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, 10(2), 141-160.
- [19] Søraker, J. H., & Brey, P. (2007). Ambient intelligence and problems with inferring desires from behaviour. *International review of information ethics*, 8(1), 7-12.
- [20] Hildebrandt, M., & Gutwirth, S. (2008). *Profiling the European citizen*. Dordrecht: Springer.

- [21] Anrig, B., Browne, W., & Gasson, M. (2008). The role of algorithms in profiling. In *Profiling the European Citizen* (pp. 65-87). Springer, Dordrecht.
- [22] Hildebrandt, M. (2008). Defining profiling: a new type of knowledge?. In *Profiling the European citizen* (pp. 17-45). Springer, Dordrecht.
- [23] Srividya, M., Mohanavalli, S., & Bhalaji, N. (2018). Behavioral Modeling for Mental Health using Machine Learning Algorithms. *Journal of medical systems*, 42(5), 88.
- [24] Harley, J. M., Trevors, G. J., & Azevedo, R. (2013). Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. *JEDM| Journal of Educational Data Mining*, 5(1), 104-146.
- [25] Han, X., Wang, L., & Huang, H. (2017). Deep Investment Behavior Profiling by Recurrent Neural Network in P2P Lending.
- [26] Chen, Y., Pavlov, D., & Canny, J. F. (2009, June). Large-scale behavioral targeting. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 209-218). ACM.
- [27] Rattinger, A., Wallner, G., Drachen, A., Pirker, J., & Sifa, R. (2016, September). Integrating and inspecting combined behavioral profiling and social network models in destiny. In *International Conference on Entertainment Computing* (pp. 77-89). Springer, Cham.
- [28] Yeung, D. Y., & Ding, Y. (2003). Host-based intrusion detection using dynamic and static behavioral models. *Pattern recognition*, 36(1), 229-243.
- [29] Fawcett, T., & Provost, F. J. (1996, December). Combining Data Mining and Machine Learning for Effective User Profiling. In *KDD* (pp. 8-13).
- [30] de Montjoye, Y. A., Quoidbach, J., Robic, F., & Pentland, A. S. (2013, April). Predicting personality using novel mobile phone-based metrics. In *International conference on social computing, behavioral-cultural modeling, and prediction* (pp. 48-55). Springer, Berlin, Heidelberg.
- [31] Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2011, June). Who's who with big-five: Analyzing and classifying personality traits with smartphones. In *Wearable Computers (ISWC), 2011 15th Annual International Symposium on* (pp. 29-36). IEEE.
- [32] Nath, S. V. (2006, December). Crime pattern detection using data mining. In *Web intelligence and intelligent agent technology workshops, 2006. wi-iat 2006 workshops. 2006 ieee/wic/acm international conference on* (pp. 41-44). IEEE.
- [33] Blackmore, K., Bossomaier, T., Foy, S., & Thomson, D. (2005). Data mining of missing persons data. In *Classification and Clustering for Knowledge Discovery* (pp. 305-314). Springer, Berlin, Heidelberg.
- [34] Blackmore, K. L., & Bossomaier, T. R. J. (2003). Soft computing methodologies for mining missing person data. *INTERNATIONAL JOURNAL OF KNOWLEDGE BASED INTELLIGENT ENGINEERING SYSTEMS*, 7(3), 132-138.
- [35] Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 201218772.
- [36] Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., & Stillwell, D. (2012, June). Personality and patterns of Facebook usage. In *Proceedings of the 4th annual ACM web science conference* (pp. 24-32). ACM.
- [37] Oberlander, J., & Nowson, S. (2006, July). Whose thumb is it anyway?: classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions* (pp. 627-634). Association for Computational Linguistics.
- [38] Borges, J., & Levene, M. (1999, August). Data mining of user navigation patterns. In *International Workshop on Web Usage Analysis and User Profiling* (pp. 92-112). Springer,

- Berlin, Heidelberg.
- [39] Turaga, P., Chellappa, R., Subrahmanian, V. S., & Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video technology*, 18(11), 1473.
- [40] Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48, 26-34.
- [41] Vasanthakumar, G. U., Sunithamma, K., Shenoy, P. D., & Venugopal, K. R. An Overview on User Profiling in Online Social Networks.
- [42] Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100-1122.
- [43] Bleidorn, W., & Hopwood, C. J. (2018). Using Machine Learning to Advance Personality Assessment and Theory. *Personality and Social Psychology Review*, 1088868318772990.
- [44] Agrawal, R., & Srikant, R. (2000). *Privacy-preserving data mining* (Vol. 29, No. 2, pp. 439-450). ACM.
- [45] Aggarwal, C. C., & Philip, S. Y. (2008). A general survey of privacy-preserving data mining models and algorithms. In *Privacy-preserving data mining* (pp. 11-52). Springer, Boston, MA.
- [46] Brody, H., Rip, M. R., Vinten-Johansen, P., Paneth, N., & Rachman, S. (2000). Map-making and myth-making in Broad Street: the London cholera epidemic, 1854. *The Lancet*, 356(9223), 64-68.
- [47] RF, Tomlinson. (1969). A geographic information system for regional planning. *Journal of Geography (Chigaku Zasshi)*, 78(1), 45-48.
- [48] Lee, E. S. (1966). A theory of migration. *Demography*, 3(1), 47-57.
- [49] Toch, E., Lerner, B., Ben-Zion, E., & Ben-Gal, I. (2018). Analyzing large-scale human mobility data: a survey of machine learning methods and applications. *Knowledge and Information Systems*, 1-23. Shen, L., & Stopher, P. R. (2014). Review of GPS travel survey and GPS data-processing methods. *Transport Reviews*, 34(3), 316-334.
- [50] Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *nature*, 453(7196), 779.
- [51] Bengtsson, L., Lu, X., Thorson, A., Garfield, R., & Von Schreeb, J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. *PLoS medicine*, 8(8), e1001083.
- [52] Massimo, D., & Ricci, F. (2018, September). Harnessing a generalised user behaviour model for next-POI recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems* (pp. 402-406). ACM.
- [53] Cho, E., Myers, S. A., & Leskovec, J. (2011, August). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1082-1090). ACM.
- [54] Chen, N. C., Xie, W., Welsch, R. E., Larson, K., & Xie, J. (2017, June). Comprehensive Predictions of Tourists' Next Visit Location Based on Call Detail Records Using Machine Learning and Deep Learning Methods. In *Big Data (BigData Congress), 2017 IEEE International Congress on* (pp. 1-6). IEEE.
- [55] Yuan, J., Zheng, Y., & Xie, X. (2012, August). Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 186-194). ACM.
- [56] Ashbrook, D., & Starner, T. (2003). Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous computing*, 7(5), 275-286.

- [57] Mohibullah, W., & Julie, S. J. (2013, October). Developing an agent model of a missing person in the wilderness. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on* (pp. 4462-4469). IEEE.
- [58] Biljecki, F., Ledoux, H., & Van Oosterom, P. (2013). Transportation mode-based segmentation and classification of movement trajectories. *International Journal of Geographical Information Science, 27*(2), 385-407.
- [59] Liu, Q., Wu, S., Wang, L., & Tan, T. (2016, February). Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts. In *AAAI* (pp. 194-200).
- [60] Horozov, T., Narasimhan, N., & Vasudevan, V. (2006, January). Using location for personalized POI recommendations in mobile environments. In *Applications and the internet, 2006. SAINT 2006. International symposium on* (pp. 6-pp). IEEE.
- [61] Zheng, Y., Zhang, L., Xie, X., & Ma, W. Y. (2009, April). Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th international conference on World wide web* (pp. 791-800). ACM.
- [62] Lin, L., & Goodrich, M. A. (2010). A Bayesian approach to modeling lost person behaviors based on terrain features in wilderness search and rescue. *Computational and Mathematical Organization Theory, 16*(3), 300-323
- [63] Traag, V., Browet, A., Calabrese, F., & Morlot, F. (2011). Social event detection in massive mobile phone data using probabilistic location inference.
- [64] Hasan, S., Schneider, C. M., Ukkusuri, S. V., & González, M. C. (2013). Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics, 151*(1-2), 304-318.
- [65] Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A tale of many cities: universal patterns in human urban mobility. *PLoS one, 7*(5), e37027.
- [66] Preotjiuc-Pietro, D., & Cohn, T. (2013, May). Mining user behaviours: a study of check-in patterns in location based social networks. In *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 306-315). ACM.
- [67] Calabrese, F., Di Lorenzo, G., & Ratti, C. (2010). Human mobility prediction based on individual and collective geographical preferences.
- [68] Laube, P., van Kreveld, M., & Imfeld, S. (2005). Finding REMO—detecting relative motion patterns in geospatial lifelines. In *Developments in spatial data handling* (pp. 201-215). Springer, Berlin, Heidelberg.
- [69] Bohte, W., & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies, 17*(3), 285-297.
- [70] Koperski, K., Adhikary, J., & Han, J. (1996, June). Spatial data mining: progress and challenges survey paper. In *Proc. ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada* (pp. 1-10).
- [71] Rao, K. V., Govardhan, A., & Rao, K. C. (2012). Spatiotemporal data mining: Issues, tasks and applications. *International Journal of Computer Science and Engineering Survey, 3*(1), 39.
- [72] Gedik, B., & Liu, L. (2005, June). Location privacy in mobile systems: A personalized anonymization model. In *Distributed Computing Systems, 2005. ICDCS 2005. Proceedings. 25th IEEE International Conference on* (pp. 620-629). IEEE.
- [73] Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons, 53*(1), 59-68.
- [74] Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM, 57*(6), 74-81.
- [75] Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010). Social media analytics and intelligence.

- IEEE Intelligent Systems*, 25(6), 13-16.
- [76] Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, 30(1), 89-116.
- [77] Salathe, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., ... & Vespignani, A. (2012). Digital epidemiology. *PLoS computational biology*, 8(7), e1002616.
- [78] Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. *In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01* (pp. 492-499). IEEE Computer Society.
- [79] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS one*, 8(9), e73791.
- [80] Lima, A. C. E., & De Castro, L. N. (2014). A multi-label, semi-supervised classification approach applied to personality prediction in social media. *Neural Networks*, 58, 122-130.
- [81] Golbeck, J., Robles, C., & Turner, K. (2011, May). Predicting personality with social media. *In CHI'11 extended abstracts on human factors in computing systems* (pp. 253-262). ACM.
- [82] Skowron, M., Tkalčič, M., Ferwerda, B., & Schedl, M. (2016, April). Fusing social media cues: personality prediction from twitter and instagram. *In Proceedings of the 25th international conference companion on world wide web* (pp. 107-108). International World Wide Web Conferences Steering Committee.
- [83] Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., ... & De Cock, M. (2016). Computational personality recognition in social media. *User modeling and user-adapted interaction*, 26(2-3), 109-142.
- [84] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *ICWSM*, 13, 1-10.
- [85] De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016, May). Discovering shifts to suicidal ideation from mental health content in social media. *In Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2098-2110). ACM.
- [86] Berjani, B., & Strufe, T. (2011, April). A recommendation system for spots in location-based online social networks. *In Proceedings of the 4th Workshop on Social Network Systems* (p. 4). ACM.
- [87] Gao, H., Tang, J., Hu, X., & Liu, H. (2015, January). Content-Aware Point of Interest Recommendation on Location-Based Social Networks. *In AAAI* (pp. 1721-1727).
- [88] Ye, M., Yin, P., & Lee, W. C. (2010, November). Location recommendation for location-based social networks. *In Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems* (pp. 458-461). ACM.
- [89] Bao, J., Zheng, Y., & Mokbel, M. F. (2012, November). Location-based and preference-aware recommendation using sparse geo-social networking data. *In Proceedings of the 20th international conference on advances in geographic information systems* (pp. 199-208). ACM.
- [90] Yang, D., Zhang, D., Yu, Z., & Wang, Z. (2013, May). A sentiment-enhanced personalized location recommendation system. *In Proceedings of the 24th ACM Conference on Hypertext and Social Media* (pp. 119-128). ACM.
- [91] Lu, Z., Wang, H., Mamoulis, N., Tu, W., & Cheung, D. W. (2017). Personalized location recommendation by aggregating multiple recommenders in diversity. *GeoInformatica*, 21(3), 459-484.
- [92] Yao, L., Sheng, Q. Z., Wang, X., Zhang, W. E., & Qin, Y. (2018). Collaborative Location Recommendation by Integrating Multi-dimensional Contextual Information. *ACM*

- Transactions on Internet Technology (TOIT)*, 18(3), 32.
- [93] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093-1113.
- [94] Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89, 14-46.
- [95] Paltoglou, G., & Thelwall, M. (2012). Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4), 66.
- [96] dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 69-78).
- [97] Bouazizi, M., & Ohtsuki, T. (2017). A pattern-based approach for multi-class sentiment analysis in twitter. *IEEE Access*, 5, 20617-20639.
- [98] Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10, No. 2010, pp. 1320-1326).
- [99] Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in human behavior*, 31, 527-541.
- [100] You, Q., Luo, J., Jin, H., & Yang, J. (2015, January). Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks. In *AAAI* (pp. 381-388).
- [101] Wang, Y., Wang, S., Tang, J., Liu, H., & Li, B. (2015, July). Unsupervised Sentiment Analysis for Social Media Images. In *IJCAI* (pp. 2378-2379).
- [102] Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*, 23(5), 544-559.
- [103] Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *ICWSM*, 14, 505-514.
- [104] Liu, J., Li, J., Li, W., & Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, 134-142.
- [105] Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70(6), 543.
- [106] Metaxas, P. T., Mustafaraj, E., & Gayo-Avello, D. (2011, October). How (not) to predict elections. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (pp. 165-171). IEEE.
- [107] Mislove, A., Viswanath, B., Gummadi, K. P., & Druschel, P. (2010, February). You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 251-260). ACM.
- [108] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3), 15
- [109] Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business. *Economies, Societies and Nations*, 296.
- [110] Alag, S. (2009). *Collective intelligence in action* (pp. 274-306). Greenwich, CT: Manning.
- [111] Wang, D., Abdelzaher, T., Kaplan, L., & Aggarwal, C. C. (2013, July). Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications. In *Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on* (pp. 530-539). IEEE.
- [112] Kreps, G. A. (1984). Sociological inquiry and disaster research. *Annual review of*

- sociology*, 10(1), 309-330.
- [113] Vieweg, S., Palen, L., Liu, S. B., Hughes, A. L., & Sutton, J. N. (2008). *Collective intelligence in disaster: Examination of the phenomenon in the aftermath of the 2007 Virginia Tech shooting*. Boulder, CO: University of Colorado.
- [114] Cardone, G., Cirri, A., Corradi, A., Foschini, L., Ianniello, R., & Montanari, R. (2014). Crowdsensing in urban areas for city-scale mass gathering management: Geofencing and activity recognition. *IEEE Sensors Journal*, 14(12), 4185-4195.
- [115] Garg, A., Choudhary, S., Bajaj, P., Agrawal, S., Kedia, A., & Agrawal, S. (2017, November). Smart Geo-fencing with Location Sensitive Product Affinity. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (p. 39). ACM.

Annex II: Research Literature

II.1 Computational Learning in Human Profiling

A comparison list of the studied literature is presented in the following table. Each row describes one scientific paper and the related methodology using the following fields:

Title [Reference]: Article title and reference to related citation.

Year: The year the article was published.

Category: Depending on the focus of the article, this field contains the category of the algorithms presented in the article. The categorisation is made in respect to the tasks involved, such as classification, regression, clustering, statistical analysis, etc.

Algorithm / Method: The algorithm/method suggested by the article, along with proposed adaptations for enhancement of results or mitigation of computation complexity.

Experimentation dataset: The source and form of the data used to test and evaluate the proposed methodology, if any.

Primary focus: The general research field of the paper and any special focus.

Results: Brief summarisation of results, conclusions and key points.

Title [reference]	Year	Category	Algorithm/ Method	Experimentation dataset	Primary focus	Results
Behavioral Modeling for Mental Health using Machine Learning Algorithms[23]	2018	<ul style="list-style-type: none"> • Clustering • Classification 	<ul style="list-style-type: none"> • K-means, Agglomerative, K-medoids. • Logistic regression, Naïve Bayes, Support Vector Machines (SVM), Decision tree, K-nearest neighbours, ensemble bagging and random forest 	656 individuals, 20 features, 3 class labels	Identify state of mental health by creating mental health profiles	<ul style="list-style-type: none"> • Clusters individuals based on responses in questionnaire into 3 classes (mentally distressed, neutral and happy). • SVM, k-NN, ensembles and random forest outperform other classifiers. • The inclusion of physiological parameters is recommended.
Deep Investment Behavior Profiling by Recurrent Neural Network in P2P Lending[25]	2017	<ul style="list-style-type: none"> • Deep learning 	<ul style="list-style-type: none"> • Recurrent Neural Networks (RNNs) with GRU and LSTM 	Dataset of 7455 investors of Prosper platform (16 features of investor and invest details)	User Profiling and time-series prediction	<ul style="list-style-type: none"> • Profiling users' investment preferences and forecasting trends. • Considers it a time-series analysis problem. • Compares with k-NN and Bayesian structural time-series.
Integrating and inspecting combined behavioral profiling and social network models in destiny[27]	2016	<ul style="list-style-type: none"> • Clustering 	<ul style="list-style-type: none"> • Archetypal analysis 	10,000 players of Destiny game (performance and networking data)	Constructing behavioural profiles based on on-line gaming data	<ul style="list-style-type: none"> • Clusters players into five profiles based on their gaming performance, as well as their social networking inside the game. • Interesting patterns were extracted.
Clustering and Profiling Students According to their Interactions with an Intelligent Tutoring System Fostering Self-Regulated Learning[24]	2013	<ul style="list-style-type: none"> • Clustering 	<ul style="list-style-type: none"> • Expectation-Minimization (EM) 	Data from 106 college students (12 variables) containing their interactions (log file) with an e-learning system	Design adaptive systems relying on behavioural profiles	<ul style="list-style-type: none"> • Analysis of student profiles derived from clustering methods. • Additional data sources are suggested, such as gaze behaviour and emotions (from analysis of video recordings).
Private traits and	2013	<ul style="list-style-type: none"> • Dimensionality 	<ul style="list-style-type: none"> • Singular-value 	Around 58.500	Psycho-	<ul style="list-style-type: none"> • Illustrates the potential of predictive

attributes are predictable from digital records of human behavior [35]		<p>Reduction</p> <ul style="list-style-type: none"> • Regression • Classification 	<p>decomposition (SVD)</p> <ul style="list-style-type: none"> • Linear regression (for predicting numeric values, e.g. age) • Logistic regression (for predicting binary variables, e.g. gender) 	Facebook profiles from "myPersonality" app	demographic profile extraction from Facebook Likes	<p>analytics in today's digital society.</p> <ul style="list-style-type: none"> • Predicts individual traits and attributes, such as religion, sexuality, political views, relationship status, alcohol consumption, age, personality scores, etc. • Application in marketing and recommender systems.
Predicting personality using novel mobile phone-based metrics [30][75]	2013	<ul style="list-style-type: none"> • Classification 	<ul style="list-style-type: none"> • Support Vector Machines (SVM) 	69 individuals with their mobile data records	Behavioural modelling from mobile-phone metrics	<ul style="list-style-type: none"> • Infers user personalities (five-factor model) based on basic information provided by mobile phone usage. • Uses mobile data indicators.
Personality and patterns of Facebook usage[36]	2012	<ul style="list-style-type: none"> • Classification • Natural Language Processing (NLP) • Dimensionality reduction 	<ul style="list-style-type: none"> • Support Vector Machines (SVM), Simple Minimal Optimization (SMO), MultiBoostAB, AdaBoostM1 	250 user instances with activity and demographic data and ~10,000 status updates	Personality modelling from Facebook data	<ul style="list-style-type: none"> • Achieves high precision in predicting user personality based on big-five model. • A set of 725 features was used broken down into five groups: demographics, activities, status updates, networking data, word classification schemes.
Who's who with big-five: Analyzing and classifying personality traits with smartphones [31]	2011	<ul style="list-style-type: none"> • Classification 	<ul style="list-style-type: none"> • C4.5, Support Vector Machines (SVM) with RBF kernel 	83 participants with mobile usage data collected over 8 months (Call logs, sms, Bluetooth scans, app logs)	Classifying personality traits based on smartphone usage	<ul style="list-style-type: none"> • To determine personality, the authors use the big-five personality framework. • Variety of data sources • Sensorial data from mobile phones may increase performance
Large-scale behavioral targeting [26]	2009	<ul style="list-style-type: none"> • Dimensionality reduction • Regression 	<ul style="list-style-type: none"> • Inverted index • Poisson linear regression 	Yahoo user base (training data: 5-week period, 500 mil. Examples)	Behavioural targeting for ads	<ul style="list-style-type: none"> • Predicts click-through rate (CTR) from user behavioural data. • Large-scale implementation with Hadoop MapReduce framework.

Crime pattern detection using data mining [32]	2006	<ul style="list-style-type: none"> • Clustering 	<ul style="list-style-type: none"> • K-means 	Hundreds of confidential crime cases	Grouping crimes with similar attributes in a geographical region	<ul style="list-style-type: none"> • Identifies crime patterns using clustering techniques. • Assigns results to a geo spatial plot (map).
Whose thumb is it anyway?: classifying author personality from weblog text [37]	2006	<ul style="list-style-type: none"> • Classification • Dimensionality reduction • Natural Language Processing (NLP) 	<ul style="list-style-type: none"> • Support Vector Machines (SVM), • Naïve Bayes classifier • N-grams 	71 authors and their blog posts (raw text)	Classification of weblog authors personality	<ul style="list-style-type: none"> • Predicting author personality based on big-five model. • Comparison of binary and multi-class classification.
Data mining of missing persons data [33]	2005	<ul style="list-style-type: none"> • Decision trees • Classification • Rule induction 	<ul style="list-style-type: none"> • C4.5 tree 	Careful selection of 357 missing persons cases	Data mining on missing person profiles	<ul style="list-style-type: none"> • The implemented algorithm produced inconsistent results. • Data used consisted of human, non-structured, judgments and estimations(suspicions) that may have deteriorated data quality.
Host-based intrusion detection using dynamic and static behavioral models [28][77]	2003	<ul style="list-style-type: none"> • Anomaly detection 	<ul style="list-style-type: none"> • Dynamic model: Hidden Markov models (HMM) with max. likelihood • Static model: Frequency distributions with min. cross entropy 	Unix system calls and shell commands datasets	Behavioural modelling for user profiling	<ul style="list-style-type: none"> • Builds user profiles from Unix system usage data. • Performs intrusion detection based on a dynamic HMM model which outperforms the static model.
Soft computing methodologies for mining missing person data [34]	2003	<ul style="list-style-type: none"> • Classification • Dimensionality reduction 	<ul style="list-style-type: none"> • Artificial Neural Networks (ANNs) • Isomap, PCA 	Selection of 326 missing persons cases	Data mining on missing person profiles	<ul style="list-style-type: none"> • Classify as runaway, suicide, or foulplay. • ANNs achieved high accuracy (93%) that outperforms rule-based systems.
Data mining of user navigation patterns [38]	1999	<ul style="list-style-type: none"> • Stochastic modelling • Dimensionality 	<ul style="list-style-type: none"> • Markov chains • N-gram model, Depth-First Search 	2-month user navigation log data from websites	Predict users' web navigation	<ul style="list-style-type: none"> • Mining log data for user navigation patterns. • Uses probability grammar modelling

		reduction	graph			to model navigation.
Combining Data Mining and Machine Learning for Effective User Profiling. [29]	1996	<ul style="list-style-type: none">• Rule induction	<ul style="list-style-type: none">• BruteDL	610 accounts comprising ~350,000 calls over 4 months	Methods for detecting fraudulent behaviour from cellular phone usage	<ul style="list-style-type: none">• Presents methods for detecting fraudulent usage of mobile phones based on profiling customer behaviour and evidence combination.

II.2 Exploiting Spatiotemporal Data coming from multiple sources

A comparison list of the studied literature is presented in the following table. Each row describes one scientific paper and the related methodology using the following fields:

Title [Reference]: Article title and reference to related citation.

Year: The year the article was published.

Category: Depending on the focus of the article, this field contains the category of the algorithms presented in the article. The categorisation is made in respect to the tasks involved, such as classification, regression, clustering, statistical analysis, etc.

Algorithm / Method: The algorithm/method suggested by the article, along with proposed adaptations for enhancement of results or mitigation of computation complexity.

Experimentation dataset: The source and form of the data used to test and evaluate the proposed methodology, if any.

Primary focus: The general research field of the paper and any special focus.

Results: Brief summarisation of results, conclusions and key points.

Title [reference]	Year	Category	Algorithm/ Method	Experimentation dataset	Primary focus	Results
Harnessing a generalised user behaviour model for next-POI recommendation [52]	2018	<ul style="list-style-type: none"> • Recommender • User behaviour modelling • Clustering • Reinforcement Learning 	<ul style="list-style-type: none"> • Markov decision process • Non-Negative Matrix Factorization • Inverse Reinforcement Learning 	575 geo-localized trajectories of users' POI-visits,	Recommendations of POIs	<ul style="list-style-type: none"> • Recommendations based on user behaviour (set of trajectories formed by sequences of POIs). • The trajectory generation includes also weather data.
Comprehensive Predictions of Tourists' Next Visit Location Based on Call Detail Records Using Machine Learning and Deep Learning Methods [54]	2017	<ul style="list-style-type: none"> • Classification • Deep Learning 	<ul style="list-style-type: none"> • Decision Tree, Random Forest, Artificial Neural Networks, Naïve Bayes, SVM (with RBF kernel) • Recurrent Neural Network (LSTM) 	16,568,179 data points in January 2015 (Call Data Records)	Location prediction based on mobile phone calls	<ul style="list-style-type: none"> • Predict tourists' next stops using existing stops, other related information and POIs. • Recurrent network outperforms (by 7%) the other methods.
Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts [59]	2016	<ul style="list-style-type: none"> • Deep Learning 	<ul style="list-style-type: none"> • Recurrent Neural Network (ST-RNN) 	Two real world datasets: Gowalla location-based social network data and Global Terrorism Database	Prediction of next location from contextual information	<ul style="list-style-type: none"> • Models time intervals in a recurrent architecture. • Incorporates distance-specific transition matrices for modelling geographical distances. • Experiments in real-world datasets show that ST-RNN outperforms other methods.
Spatiotemporal patterns of urban human mobility [64]	2013	<ul style="list-style-type: none"> • Probabilistic Modelling 	<ul style="list-style-type: none"> • Simplistic models 	1,000 users Oyster card transactions	Urban mobility patterns detection and prediction	<ul style="list-style-type: none"> • The article aims to detect spatial and temporal patterns of human mobility in a city using the data from smart subway fare card transactions. • Authors suggest that the heterogeneity of individuals and time dependence in trip selection should also be incorporated in the model.

Mining user behaviours: a study of check-in patterns in location based social networks [66]	2013	<ul style="list-style-type: none"> • Clustering • Movement prediction 	<ul style="list-style-type: none"> • K-means • Order-K Markov mode • Use of transition probability matrix between venue categories 	10,000 Foursquare users, each with 1-month trails	Urban mobility patterns detection and prediction	<ul style="list-style-type: none"> • The method first clusters users based on their behavior and then predicts their mobility taking into account temporal periodicities. • Splits venues into 9 categories (e.g. residence, nightlife, shop, work, etc.). • Interesting associations were discovered (e.g. very strong weekly patterns, increase of activity as the week progresses). • Incorporating explicitly the periodicity of user behavior shows to give the best result.
Developing an agent model of a missing person in the wilderness [57][103]	2013	<ul style="list-style-type: none"> • Stochastic modelling 	<ul style="list-style-type: none"> • Probabilistic Markov model for movement states • Topography transition table for attractive force modelling 	GPS logs of 10 users	Modelling human movement in unfamiliar environment	<ul style="list-style-type: none"> • Complex human behavioural states and interactions with environment are described through an agent-based model (UAV). • Agent model is controlled by a set of strategies and can switch between goals. • Performs better than diffusion model which was less accurate and focused.
Transportation mode-based segmentation and classification of movement trajectories [58]	2013	<ul style="list-style-type: none"> • Classification 	<ul style="list-style-type: none"> • Trajectory segmentation • Fuzzy expert system 	Data from Dutch National Travel survey (benchmark dataset)	Transportation mode estimation based on movement trajectories	<ul style="list-style-type: none"> • Selection of indicators resulted in nine values: 3 speed related (nearly maximum speed, mean speed, mean moving speed), 5 average proximities from infrastructures (railway, tram lines, roads, bus lines, metro lines) and the location of the trajectory in respect to water surfaces. • Uses OpenStreetMap for geographical data and transportation infrastructure. • Data are organised into geometry (polylines) and attributes (tags).
A tale of many cities: universal patterns in human urban mobility [65]	2012	<ul style="list-style-type: none"> • Probabilistic Modelling 	<ul style="list-style-type: none"> • Probability Density Functions (for displacement probability distribution) • Kolmogorov-Smirnov test for distribution synthesis 	35,289,629 Foursquare <i>check-ins</i> from 925,030 unique users over 4,960,496 venues	Modelling urban mobility	<ul style="list-style-type: none"> • The authors consider consecutive check-ins within same city as one movement. • Distance is not the deciding factor in human mobility, but place density is important. • Each place is described through a rank metric, which presents the #places offered

			<ul style="list-style-type: none"> • Maximum Likelihood Estimation for parameter fine-tuning 			<p>between origin and destination.</p> <ul style="list-style-type: none"> • The rank-based model is used to calculate human mobility in the city. • Distribution of human displacements approximated with power law.
Discovering regions of different functions in a city using human mobility and POIs [55]	2012	<ul style="list-style-type: none"> • Clustering • Topic Modelling • Regression • Semantic Annotation 	<ul style="list-style-type: none"> • K-means clustering for region functionality aggregation • TF-IDF on POIs data • Latent Dirichlet Allocation (LDA) on mobility data • Dirichlet Multinomial Regression (DMR) • Kernel Density Estimation model for functionality intensity (number of visits) 	<p>Two Beijing POI datasets for years 2010 and 2011</p> <p>Two GPS trajectory datasets from 12,000 taxicabs</p> <p>Beijing road networks</p>	Mobility pattern detection in big city regions	<ul style="list-style-type: none"> • Uses Regions of different Functions (DRoF) and points of interest. • Employs topic-based inference model which regards region as document, function as topic, categories of POIs as metadata and human mobility patterns as words. • Uses raster-based model to represent road network. • Compares DRoF against TF-IDF-based and LDA-based methods.
Friendship and mobility: user movement in location-based social networks [53]	2011	<ul style="list-style-type: none"> • Probabilistic Modelling 	<ul style="list-style-type: none"> • Probability distributions • Expectation-Minimization (EM) 	<ol style="list-style-type: none"> 1) 6.4 mil. check-in data from Gowalla social platform 2) 4.5 mil. from Brightkite social platform 3) 450 mil. phone calls dataset 	Modelling human mobility patterns and temporal dynamics	<ul style="list-style-type: none"> • Modelling human mobility that combines short range movement with travel. • Multiple sources, from social networks to phone call logs, are analysed. • Network connections (friendships) are also considered. • Gives an order of magnitude better performance than other stochastic models. • Social networks influence long distance travel more than short distance. • Periodicity also investigated.
Social event detection in massive mobile phone data using probabilistic location inference [63]	2011	<ul style="list-style-type: none"> • Probabilistic Modelling 	<ul style="list-style-type: none"> • Voronoi partition (for location inference) • Bayesian location inference model 	~ 900 million calls and text messages (anonymized caller and callee, location estimation from towers)	Social events detection and attendance prediction	<ul style="list-style-type: none"> • Aims to detect unusual massive gatherings (concerts, sports finals, emergencies, protests etc.) from mobile network data. • Focus on non-routine behaviour. • Non-routine behaviour of people correlates with behaviour of other people too. • The framework detected successfully the

						events that took place in the examined areas.
Human mobility prediction based on individual and collective geographical preference [67]	2010	<ul style="list-style-type: none"> • Probabilistic Modelling 	<ul style="list-style-type: none"> • Custom model 	<ol style="list-style-type: none"> 1) Mobile phones locations: traces from 2000 users 2) Land use data from MassGIS 3) POI data from Yelp 	Predict location of a person over time	<ul style="list-style-type: none"> • Method combines data from multiple sources, using open data and mobile phones data • Model is based on the geographical features of the area where the person moves, in terms of land use, points of interests and collectivity's habits • Can be used as recommender
A Bayesian approach to modeling lost person behaviors based on terrain features in wilderness search and rescue [62]	2010	<ul style="list-style-type: none"> • Probabilistic Modelling 	<ul style="list-style-type: none"> • Bayesian model for probability distribution map • Order-1 Markov mode 	Synthetic GPS data	Predict route of missing persons	<ul style="list-style-type: none"> • Behaviour modelling is based on three terrain features: topography, vegetation and local slope. • Past operations data, expert opinions and past statistical data can be incorporated into the model. • A temporal state transition matrix allows the generation of predictions for any given time interval.
Mining interesting locations and travel sequences from GPS trajectories [61]	2009	<ul style="list-style-type: none"> • Clustering • Ranking (of location interest) • Graph theory 	<ul style="list-style-type: none"> • OPTICS: density-based clustering algorithm • nDCG, MAP 	GPS trajectories of 107 users over 1 year	Discover interesting locations and travel sequences	<ul style="list-style-type: none"> • HITS-based inference model • Works best as a recommender for tourist attractions and best sequence to visit them • Clear advantages over <i>rank-by-count</i> and <i>rank-by-frequency</i> methods
Using location for personalized POI recommendations in mobile environments [60]	2006	<ul style="list-style-type: none"> • Recommender 	<ul style="list-style-type: none"> • Collaborative filtering (based on location) 	12,000 restaurant POIs	Recommend POIs	<ul style="list-style-type: none"> • Recommendations based on user ratings for restaurant POIs that are in the same vicinity • Ways to deal with the "Cold start" problem (when there are no votes yet)
Finding REMO— detecting relative motion patterns in geospatial lifelines [68]	2005	<ul style="list-style-type: none"> • Clustering 	<ul style="list-style-type: none"> • Voronoi partition • Geometric algorithms for track & encounter patterns 	Lifeline data (id, location, time) and Motion attributes (speed, change of speed, azimuth)	Cluster detection of motion patterns in space-time	<ul style="list-style-type: none"> • The method relies solely on point observations in a Euclidean space • Introduces a wide variety of motion patterns • REMO patterns are spatiotemporal REMO patterns can be used for any object that can be depicted as point and leaves a

						track behind
Using GPS to learn significant locations and predict movement across multiple users [56]	2003	<ul style="list-style-type: none">• Clustering• Predictive Modelling• Probabilistic Theory	<ul style="list-style-type: none">• K-means for location and sublocation definition from individual places• Markov model	GPS logs from one user for a 4-month period	Meaningful locations modelling and extraction	<ul style="list-style-type: none">• Single-user and multi-user applications• Significant places are traced from GPS time-gaps (due to lack of signal from inside buildings)

II.3 Social Media Analytics

A comparison list of the studied literature is presented in the following table. Each row describes one scientific paper and the related methodology using the following fields:

Reference

Title [Reference]: Article title and reference to related citation.

Year: The year the article was published.

Category: Depending on the focus of the article, this field contains the category of the algorithms presented in the article. The categorisation is made in respect to the tasks involved, such as classification, regression, clustering, statistical analysis, etc.

Algorithm / Method: The algorithm/method suggested by the article, along with proposed adaptations for enhancement of results or mitigation of computation complexity.

Experimentation dataset: The source and form of the data used to test and evaluate the proposed methodology, if any.

Primary focus: The general field of the paper and any special focus.

Results: Brief summarisation of results, conclusions and key points.

Title [reference]	Year	Category	Algorithm/ Method	Experimentation dataset	Primary focus	Results
Discovering shifts to suicidal ideation from mental health content in social media [85]	2016	<ul style="list-style-type: none"> • Linguistic Analysis • Classification 	<ul style="list-style-type: none"> • Variable extraction • Logistic regression 	Discussion data from 880 users in Reddit (approx. 13,000 posts & 100,000 comments)	Intention discovery	<ul style="list-style-type: none"> • Mostly statistical approach • The approach considers 5 models: linguistic structure, interpolator awareness, interaction, content, & full. • Method can be easily adapted to other social media and for other vulnerable groups of people
Fusing social media cues: personality prediction from twitter and Instagram [82]	2016	<ul style="list-style-type: none"> • Linguistic Analysis • Feature extraction • Regression 	<ul style="list-style-type: none"> • Linguistic Inquiry and Word Count (LIWC), ANEW lexicon • F-statistic subsampling for feature extraction • Random forest • Big Five model 	Instagram and Twitter (images, texts -tweets and image captions-, #followers & #followees)	Personality prediction based on various social media data	<ul style="list-style-type: none"> • Best results when using input from both social networks • Method can be easily adapted to other social media • The overall best regressor was the one using the complete feature set (linguistic, image, meta)
Computational personality recognition in social media [83]	2016	<ul style="list-style-type: none"> • Linguistic Analysis • Regression 	<ul style="list-style-type: none"> • Linguistic Inquiry and Word Count (LIWC) • Univariate regression (SVM with radial kernel and decision tree), • Multivariate regression algorithms (SVM, 	<ul style="list-style-type: none"> • Facebook, Twitter, (demographics, texts, activities) • YouTube vlogs (audio-video features) 	Personality prediction based on various social media platforms and data	<ul style="list-style-type: none"> • Age and gender showed high correlation with all personality traits • Decision tree models generally outperformed SVM • Overall best

			decision tree) • Big Five model			performance: MTSC and ERCC with decision tree base learner
A multi-label, semi-supervised classification approach applied to personality prediction in social media [80]	2014	• Semi-supervised Classification	• Naïve Bayes, Support Vector Machine, MLP • Big Five Model	Tweets with grammatical meta-attributes	Personality prediction based only on tweets meta-attributes	• Extroversion, agreeableness and neuroticism are accurately predicted • Openness and conscientiousness were more difficult to predict with the suggested meta-attributes • Less dependent of language in comparison to grammar-based approaches
Predicting depression via social media [84]	2013	• Linguistic Analysis • Feature extraction • Regression	• Statistical approach • PCA • SVM (with RBF kernel)	Tweets and attributes (188 features vector)	Depression prediction based on various social media data	• Depression prediction ahead of onset yields accuracy ~70% • Model that uses only linguistic features performs better
Personality, gender, and age in the language of social media: The open-vocabulary approach [79]	2013	• Linguistic Analysis • Feature extraction • Correlation analysis	• Linguistic Inquiry and Word Count (LIWC) • Latent Dirichlet Allocation (LDA) • Ordinary Least Squares regression	Analysed 700 mil. words of 75,000 Facebook users	Open vocabulary method for personality prediction	• Open-vocabulary technique • Largest study to date • Predictions on gender, age and personality • A word cloud visualization is presented
Predicting personality with social media [81]	2011	• Linguistic Analysis • Feature weighting • Regression	• Linguistic Inquiry and Word Count (LIWC) • Multiple linear regression to predict feature weights for each personality trait (M5' and Gaussian	Facebook profiles (personal info, preferences and activities, posts, network characteristics)	Personality prediction based on various social media data	• Achieved personality prediction within 11% of actual values • Short texts, like fb posts, can be insufficient for linguistic analysis, so individual texts were

			processes) • Big Five Model			unified before analysis
Collaborative Location Recommendation by Integrating Multi-dimensional Contextual Information [92]	2018	• Recommender	<ul style="list-style-type: none"> • Model-based Collaborative Filtering (Tensor Factorization) • Non negative Matrix Factorization • Stochastic gradient descent (SGD) 	<ol style="list-style-type: none"> 1)Brightkite user check-ins 2) Gowalla user check-ins 	Location (POI) recommendation	<ul style="list-style-type: none"> • Incorporates friends' influence • Outperforms other CF models
Personalized location recommendation by aggregating multiple recommenders in diversity [91]	2017	• Recommender Ensemble	• Various Collaborative Filtering (CF) methods	<ol style="list-style-type: none"> 1) Foursquare user check-ins 2) Gowalla user check-ins 	Location recommendation with recommender ensemble	<ul style="list-style-type: none"> • Novel framework (LURWA) of recommender ensemble using various weighting strategies • Linear aggregation • Outperforms typical CF models
Content-Aware Point of Interest Recommendation on Location-Based Social Networks [87]	2015	<ul style="list-style-type: none"> • Recommender • Sentiment Analysis 	• Custom CF method (CAPRF)	Foursquare user check-ins, user tweets and POIs properties	Location recommendations based on social networks	<ul style="list-style-type: none"> • Combines POI properties, user interests, and sentiment indications to recommend a location • Sparse dataset that produces very low precision results
A sentiment-enhanced personalized location recommendation system [90]	2013	<ul style="list-style-type: none"> • Recommender • Sentiment Analysis 	<ul style="list-style-type: none"> • Collaborative Filtering (CF) • Location-based Social Matrix Factorization • Dictionary-based unsupervised sentiment analysis 	<ol style="list-style-type: none"> 1)Foursquare user check-ins and network connections and 2) Foursquare venue categories & tips 	Personalised location recommendation using user check-ins and contextual information	<ul style="list-style-type: none"> • Social influence and venue similarity are considered. • This hybrid model outperforms other models that only use check-ins or consider tips in a random way • Models that incorporate social influence impact

						perform better
Location-based and preference-aware recommendation using sparse geo-social networking data	2016	<ul style="list-style-type: none"> • Recommender 	<ul style="list-style-type: none"> • Collaborative Filtering (CF) 	Number of visits to venue	Location recommendation based on user preferences and social opinions	<ul style="list-style-type: none"> • Opinions from local experts with whom the user shares interests are included in the calculation • User preferences extracted from location history This approach outperformed major recommendation methods MPC, LCF, PCF
A recommendation system for spots in location-based online social networks [86]	2011	<ul style="list-style-type: none"> • Recommender 	<ul style="list-style-type: none"> • Memory-based Collaborative Filtering (CF) • Regularised Matrix Factorization • Specific region restriction 	Gowalla check-ins	Personalised location recommendation	<ul style="list-style-type: none"> • Provides a solution to the lack of straight forward ratings • The error metrics were below the baseline • Geographic and temporal aspects are considered in the recommendation
Location recommendation for location-based social networks [88]	2010	<ul style="list-style-type: none"> • Recommender 	<ul style="list-style-type: none"> • Collective Matrix Factorization • Friend-based CF (FCF) • Geo-Measured FCF 	Foursquare user profiles (User-location pairs & network)	Location recommendation based on social and geographical characteristics	<ul style="list-style-type: none"> • Considering friend similarity limits the user space and calculations • Distance between friends is also a factor in the geo-measured FCF method • Both proposed techniques offer lower computational complexity
A pattern-based approach	2017	<ul style="list-style-type: none"> • Linguistic Analysis • ML Classification 	<ul style="list-style-type: none"> • Natural Language Processing (NLP) 	40,000 Tweets (manually labelled)	Multi-class sentiment analysis	<ul style="list-style-type: none"> • Classify into 7 emotion-based classes instead of

for multi-class sentiment analysis in twitter [97]			techniques <ul style="list-style-type: none"> • Random Forest 			usual 3 <ul style="list-style-type: none"> • Performance ~60% • Software (java) called SENTA was created after this approach
Unsupervised Sentiment Analysis for Social Media Images [101]	2015	<ul style="list-style-type: none"> • Unsupervised Classification 	<ul style="list-style-type: none"> • Non-negative matrix factorization • SentiBank and EL frameworks with K-means instead of SVM/logistic regression 	Flickr, Instagram images with text captions	Image sentiment analysis, enhanced by contextual information	<ul style="list-style-type: none"> • Proposes an Unsupervised Sentiment Analysis framework (USEA) • Incorporates visual and textual information, in order to provide more accurate predictions on image sentiment content • Too much textual information dominates the learning process and results in overfitting •
Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks [100]	2015	<ul style="list-style-type: none"> • Deep learning 	<ul style="list-style-type: none"> • Deep Convolution Neural Networks • Progressive CNN for noise reduction • Transfer learning 	Flickr images, sentiment labelled Twitter images	Large scale image sentiment analysis	<ul style="list-style-type: none"> • Both progressive training and transfer learning based on a small, confidently labelled subset, increased performance • Leveraged large training and testing datasets for developing a more robust model
Deep convolutional neural networks for sentiment analysis of short texts [96]	2014	<ul style="list-style-type: none"> • Deep learning 	<ul style="list-style-type: none"> • Word-embeddings (e.g. word2vec) • Deep Convolution Neural Networks 	Movie reviews and twitter messages (SSTb and STS corpora)	Sentiment analysis with deep convolutional neural networks	<ul style="list-style-type: none"> • Performs analysis using character-level, word-level and sentence-level representations
Sentiment analysis in Facebook and its	2014	<ul style="list-style-type: none"> • Linguistic Analysis and Lexicon-based classification 	<ul style="list-style-type: none"> • Natural Language Processing (NLP) techniques 	Facebook posts	Text sentiment classification with both lexicon-based	<ul style="list-style-type: none"> • The approach detects emotional changes by comparing the "current"

application to e-learning [99]		<ul style="list-style-type: none"> • Feature selection • ML Classification 	<ul style="list-style-type: none"> • Correlation-based feature selection • Decision trees (C4.5), Naïve Bayes, SVM 		and ML techniques	<p>sentiment of a user with the “usual” one.</p> <ul style="list-style-type: none"> • ML algorithms combined with lexicon-based techniques, achieve higher performance • Decision trees had slightly better performance among ML algorithms
Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media [95]	2012	<ul style="list-style-type: none"> • Linguistic Analysis and Lexicon-based classification • ML Classification 	<ul style="list-style-type: none"> • Natural Language Processing (NLP) techniques • Naïve Bayes, Maximum Entropy, SVM 	Twitter, Digg, Myspace comments	Subjectivity and polarity classification in less domain-specific, informal texts	<ul style="list-style-type: none"> • In polarity classification, the proposed lexicon-based classifier outperformed ML techniques • In subjectivity classification the proposed lexicon-based classifier outperformed ML techniques, except for the twitter dataset • Emotional word detection and neighbourhood scanning for negators/intensifiers
Twitter as a corpus for sentiment analysis and opinion mining [98]	2010	<ul style="list-style-type: none"> • ML Classification 	<ul style="list-style-type: none"> • Multinomial Naïve Bayes • Also experimented with SVM and CRF 	Tweets for corpus collection POS-tags	Sentiment classification of microblogging posts	<ul style="list-style-type: none"> • The proposed method for negative/positive/neutral sentiment corpus collection from Twitter posts can be fully automated and has no volume limitation • Bigrams performed better than uni- and trigrams

Annex III: Past cases list of reference

#	Pilot Name	Case ID	Date Created	Date Closed	Notes
1	Child Focus	CM: 12345	31-6-2009 11:45	27-08-2009	
2	Child Focus	CM: 580	30-12-2009 11:15	03-01-2010	
3	Child Focus	CM: 1271	18-04-2010 16:36	24-05-2010	
4	Child Focus	CM: 4173	22-05-2011 18:36	25-05-2011	
5	Child Focus	CM: 5070	6-10-2011 20:18	16-10-2011	
6	Child Focus	CM: 5318	15-11-2011 13:43	17-11-2011	
7	Child Focus	CM: 6044	7-03-2012 12:18	07-06-2012	
8	Child Focus	CM: 6498	20-05-2012 19:00	03-06-2012	
9	Child Focus	CM: 7054	11-08-2012 16:21	30-10-2012	
10	Child Focus	CM: 7790	12-12-2012 12:03	17-12-2012	
11	Child Focus	CM: 8797	13-06-2013 23:06	25-06-2013	
12	Child Focus	CM: 11637	20-11-2014 22:05	09-12-2014	
13	Child Focus	CM: 13376	8-11-2015 19:17	09-11-2015	
14	Child Focus	CM: 13823	15-02-2016 10:06	28-02-2016	
15	Child Focus	CM: 14196	26-04-2016 19:18	09-05-2016	
16	Child Focus	CM: 14286	16-05-2016 19:27	17-05-2016	
17	Child Focus	CM: 15868	18-03-2017 14:13	22-03-2017	
18	Child Focus	CM: 16214	26-05-2017 23:38	10-01-2018	
19	Child Focus	CM: 16563	25-07-2017 0:42	25-07-2017	
20	Child Focus	CM: 16931	30-09-2017 19:51	01-11-2017	
21	Child Focus	CM: 123456	6-04-2009 10:56	19-04-2009	
22	Child Focus	CM: 2676	24-10-2010 10:40	26-10-2010	
23	Child Focus	CM: 4183	23-05-2011 19:03	26-05-2011	
24	Child Focus	CM: 4395	23-06-2011 12:53	07-07-2011	
25	Child Focus	CM: 4643	31-07-2011 0:50	04-08-2011	
26	Child Focus	CM: 5009	25-09-2011 18:27	28-09-2011	

27	Child Focus	CM: 6805	5-07-2012 12:11	09-07-2012	
28	Child Focus	CM: 7501	16-10-2012 18:49	23-10-2012	
29	Child Focus	CM: 7643	14-11-2012 7:32	18-11-2012	
30	Child Focus	SM: 8196	21-02-2013 8:17	16-09-2013	
31	Child Focus	CM: 8719	30-05-2013 21:37	06-06-2013	
32	Child Focus	CM: 9359	14-09-2013 18:33	15-09-2013	
33	Child Focus	CM: 9828	16-12-2013 13:35	13-01-2014	
34	Child Focus	CM: 10511	3-05-2014 12:16	04-05-2014	
35	Child Focus	CM: 11207	7-09-2014 10:36	11-09-2014	
36	Child Focus	CM: 11286	16-09-2014 15:17	08-10-2014	
37	Child Focus	CM: 11479	20-10-2014 14:36	21-10-2014	
38	Child Focus	CM: 11886	17-01-2015 10:39	27-01-2015	
39	Child Focus	CM: 15181	25-10-2016 3:16	02-11-2016	
40	Child Focus	CM: 17019	16-10-2017 15:51	18-10-2017	
41	Smile of the Child	SOCB1	25-02-2018	02-03-2018	Alarming disappearance 1
42	Smile of the Child	SOCB2	17-11-2016	17-11-2016	Alarming disappearance 2
43	Smile of the Child	SOCB3	25-07-2018	27-07-2018	Alarming disappearance 3
44	Smile of the Child	SOCB4	12-11-2017	12-11-2017	Alarming disappearance 4
45	Smile of the Child	SOCB5	09-08-2016	09-08-2016	Alarming disappearance 5
46	Smile of the Child	SOCB6	10-06-2017	10-06-2017	Alarming disappearance 6
47	Smile of the Child	SOCB7	24-04-2018	26-04-2018	Alarming disappearance 7
48	Smile of the Child	SOCB8	15-12-2017	16-12-2017	Alarming disappearance 8
49	Smile of the Child	SOCB9	23-01-2017	23-01-2017	Alarming disappearance 9
50	Smile of the Child	SOCB10	02-07-2018	03-07-2018	Alarming disappearance 10
51	Smile of the Child	SOCC1	18-09-2016	23-09-2016	Parental abduction 1
52	Smile of the Child	SOCC2	10-11-2015	11-12-2015	Parental abduction 2
53	Smile of the Child	SOCC3	14-01-2016	11-10-2016	Parental abduction 3
54	Smile of the Child	SOCC4	30-06-2016	01-07-2016	Parental abduction 4
55	Smile of the Child	SOCC5	08-05-2017	09-10-2017	Parental abduction 5
56	Smile of the Child	SOCD1	24-11-2016	30-11-2016	Missing unaccompanied migrant

					minors 1
57	Smile of the Child	SOCD2	13-01-2016	OPEN	Missing unaccompanied migrant minors 2
58	Smile of the Child	SOCD3	25-04-2017	25-04-2017	Missing unaccompanied migrant minors 3
59	Smile of the Child	SOCD4	26-07-2016	OPEN	Missing unaccompanied migrant minors 4
60	Smile of the Child	SOCA1	29-02-2016	01-03-2016	Runaways of teenagers 1
61	Smile of the Child	SOCA2	22-10-2015	22-10-2015	Runaways of teenagers 2
62	Smile of the Child	SOCA3	19-09-2015	10-11-2015	Runaways of teenagers 3
63	Smile of the Child	SOCA4	04-06-2015	06-06-2015	Runaways of teenagers 4
64	Smile of the Child	SOCA5	12-11-2016	12-11-2016	Runaways of teenagers 5
65	Smile of the Child	SOCA6	01-09-2016	03-09-2016	Runaways of teenagers 6
66	Smile of the Child	SOCA7	20-12-2016	07-01-2017	Runaways of teenagers 7
67	Smile of the Child	SOCA8	27-04-2015	16-05-2015	Runaways of teenagers 8
68	Smile of the Child	SOCA9	09-02-2018	10-02-2018	Runaways of teenagers 9
69	Smile of the Child	SOCA10	06-02-2017	07-02-2017	Runaways of teenagers 10
70	Smile of the Child	SOCA11	18-01-2017	20-01-2017	Runaways of teenagers 11
71	Smile of the Child	SOCA12	12-02-2018	13-02-2018	Runaways of teenagers 12
72	Smile of the Child	SOCA13	21-04-2018	23-04-2018	Runaways of teenagers 13
73	Smile of the Child	SOCA14	30-06-2017	04-07-2017	Runaways of teenagers 14
74	Smile of the Child	SOCA15	13-05-2018	14-05-2018	Runaways of teenagers 15
75	Smile of the Child	SOCA16	31-05-2016	01-06-2018	Runaways of teenagers 16
76	Smile of the Child	SOCA17	30-05-2018	20-06-2018	Runaways of teenagers 17
77	Smile of the Child	SOCA18	06-07-2018	07-07-2018	Runaways of teenagers 18
78	Smile of the Child	SOCA19	24-06-2018	26-06-2018	Runaways of teenagers 19
79	Smile of the Child	SOCA20	31-07-2018	15-08-2018	Runaways of teenagers 20
80	Smile of the Child	SOCA21	17-06-2017	17-06-2017	Runaways of teenagers 21
81	Smile of the Child	SOCA22	16-03-2018	16-03-2018	Runaways of teenagers 22
82	Smile of the Child	SOCA23	01-04-2015	08-04-2015	Runaways of teenagers 23
83	Hellenic RedCross Tracing Division	GRC-001854	07-12-2016	23-01-2017	

84	Hellenic RedCross Tracing Division	GRC-001854	07-12-2016	23-01-2017	
85	Hellenic RedCross Tracing Division	GRC-001175	25-11-2015	26-11-2015	
86	Hellenic RedCross Tracing Division	GRC-001175	25-11-2015	26-11-2015	
87	Hellenic RedCross Tracing Division	GRC-001175	25-11-2015	26-11-2015	
88	Hellenic RedCross Tracing Division	GRC-002221	30-06-2017	24-07-2015	
89	Hellenic RedCross Tracing Division	GRC-002121	05-05-2017	12-06-2017	(still open-lost after located)
90	Hellenic RedCross Tracing Division	GRC-002121	05-05-2017	12-06-2017	(still open)
91	Hellenic RedCross Tracing Division	GRC-001372	07-03-2016	OPEN	(still open)
92	Hellenic RedCross Tracing Division	GRC -001404	30-10-2015	21-03-2016	
93	Hellenic RedCross Tracing Division	GRC-002216	22-06-2017	04-07-2017	
94	Hellenic RedCross Tracing Division	GRC-001974	25-01-2017	20-04-2017	
95	Hellenic RedCross Tracing Division	GRC-001700	07-01-2016	25-02-2017	
96	Hellenic RedCross Tracing Division	GRC-002384	25-10-2017	04-06-2018	
97	Hellenic RedCross Tracing Division	GRC-002294	24-08-2017	OPEN	(still open)
98	Hellenic RedCross Tracing Division	GRC-001380	12-03-2016	OPEN	(still open)
99	Hellenic RedCross Tracing Division	GRC-002438	06-11-2017	OPEN	(still open)
100	Hellenic RedCross Tracing Division	GRC-001378	10-03-2016	30-03-2016	
101	Hellenic RedCross Tracing Division	GRC-001424	10-03-2016	OPEN	(still open)
102	Hellenic RedCross Tracing Division	GRC-001056	02-10-2015	OPEN	(still open)
103	HRC Kalavryta Reception Center for UMC	2	28-07-2017	15-09-2017	
104	HRC Kalavryta Reception Center for UMC	4	28-07-2017	14-01-2018	
105	HRC Kalavryta Reception Center for UMC	5	28-07-2017	14-01-2018	
106	HRC Kalavryta Reception Center for UMC	6	28-07-2017	15-09-2017	
107	HRC Kalavryta Reception Center for UMC	7	31-07-2017	17-03-2018	
108	HRC Kalavryta Reception Center for UMC	8	31-07-2017	14-01-2018	
109	HRC Kalavryta Reception Center for UMC	12	31-07-2017	07-02-2018	
110	HRC Volos Reception Center for UMC	1383	28-08-2018	13-09-2018	
111	HRC Volos Reception Center for UMC	1379	27-08-2018	13-09-2018	
112	HRC Volos Reception Center for UMC	1378	23-08-2018	03-09-2018	
113	HRC Volos Reception Center for UMC	1380	28-08-2018	30-08-2018	

114	HRC Volos Reception Center for UMC	1381	28-08-2018	29-08-2018	
115	HRC Volos Reception Center for UMC	1330	30-06-2017	17-06-2018	
116	HRC Volos Reception Center for UMC	1366	30-03-2018	15-06-2018	
117	HRC Safe Zone program in Ritsona	358484	26-06-2018	01-08-2018	
118	HRC Safe Zone program in Ritsona	97665	04-05-2018	07-05-2018	
119	HRC Safe Zone program in Ritsona	81159	04-05-2018	07-05-2018	
120	HRC Safe Zone program in Ritsona	101067	02-03-2018	18-03-2018	
121	HRC Safe Zone program in Ritsona	66046	08-06-2017	09-08-2017	
122	HRC Safe Zone program in Ritsona	59481	12-05-2017	09-08-2017	