Ref. Ares(2020)2288367 - 29/04/2020

Grant Agreement No.: 823783
Call: H2020-FETPROACT-2018-2020
Topic: H2020-FETPROACT-2018-01
Type of action: RIA

WENET
INTERNET OF US

D2.1 INITIAL INDIVIDUAL LEARNING METHODS

Revision: v.1.0

Work package	WP2
Task	2.1, 2.2
Due date	30/04/2020
Submission date	30/04/2020
Deliverable lead	Idiap Research Institute
Version	1.0
Authors	Lakmal Meegahapola (Idiap Research Institute, Switzerland) William Droz (Idiap Research Institute, Switzerland) Daniel Gatica-Perez (Idiap Research Institute, Switzerland) Qiang Shen (University of Trento, Italy) Andrea Passerini (University of Trento, Italy) Fausto Giunchiglia (University of Trento, Italy)
Reviewers	Andrea Passerini (University of Trento, Italy)

Abstract	The overall objective of WP2 (Diversity-Aware Learning of Individual Behaviour) is to design and implement, from mobile sensor and app data, new algorithms to achieve diversity-aware individual routine learning, and

	diversity-aware user category learning. In other words, the learning methods in WP2 provide the situational context of users of the diversity-aware, mobile WeNet platform. The main partners contributing to WP2 are IDIAP and UNITN.
Keywords	Diversity-aware, Machine Learning, Mobile Sensing, Routines, Behavior

Document Revision History

Version	Date	Description of change	List of contributor(s)
V1	30/04/2020	1st version of the submission	Lakmal Meegahapola (Idiap Research Institute, Switzerland), William Droz (Idiap Research Institute, Switzerland), Daniel Gatica-Perez (Idiap Research Institute, Switzerland), Qiang Shen (University of Trento, Italy), Andrea Passerini (University of Trento, Italy), Fausto Giunchiglia (University of Trento, Italy)

DISCLAIMER

The information, documentation and figures available in this deliverable are written by the "WeNet - The Internet of US" (WeNet) project's consortium under EC grant agreement 823783 and do not necessarily reflect the views of the European Commission.

The European Commission is not liable for any use that may be made of the information contained herein.

COPYRIGHT NOTICE

© 2019 - 2022 WeNet Consortium

Project co-funded by the European Commission in the H2020 Programme		
Nature of the deliverable:	R	
Dissemination Level		
PU	Public, fully open, e.g. web	✓
CL	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to WeNet project and Commission Services	

** R: Document, report (excluding the periodic and final reports)*

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

OTHER: Software, technical diagram, etc.

1 EXECUTIVE SUMMARY

The overall objective of WP2 (Diversity-Aware Learning of Individual Behaviour) is to design and implement, from mobile sensor and app data, new algorithms to achieve diversity-aware individual routine learning, and diversity-aware user category learning. In other words, the learning methods in WP2 provide the situational context of users of the diversity-aware, mobile WeNet platform. The main partners contributing to WP2 are IDIAP and UNITN.

As stated in the proposal, WP2 has three tasks:

T2.1. Diversity-aware routine learning *[Lead: IDIAP; Participant: UNITN; M1-M36]*. Development of methods to learn routines (regularities in time, space, and activities) according to diversity principles.

T2.2 Diversity-aware learning and missing data *[Lead: IDIAP; Participant: UNITN; M1-M40]*. Development of methods to design tradeoffs between utility and diversity in data (e.g. due to privacy and sharing practices).

T2.3 Diversity-aware user category learning *[Lead: UNITN; Participant: CSIC, OUC, IDIAP; M1-M40]*. Development of methods to discover user categories (groups of people) above individual attributes.

In this deliverable, we describe the work done to develop and test the set of initial individual learning methods for the project. In summary, the work described in this document spans six outcomes:

(1) A literature review on smartphone sensing to support the well-being of youth, framing research in social science and ubiquitous computing through the WeNet goals of diversity, social interaction, and ethics (related to Tasks T2.1 and T2.2, conducted by IDIAP).

(2) Basic routine models for integration in the WeNet platform based on spatial-temporal Bag-of-Word and Topic Models (related to Task T2.1, conducted by IDIAP).

(3) A method for protection of diversity attributes in mobile food journals, which uses auto-encoders and multi-task neural networks to perform application-related inferences while preserving personal attributes (related to Task T2.2, conducted by IDIAP).

(4) A method to cope with the problem of untruthfulness of human annotators during real-time reporting of activities using mobile applications, based on a skeptical learning approach (related to Task T2.2, conducted by UNITN).

(5) A method for user context recognition using an ontology-based model of daily life (temporal, spatial and social), which extracts the user's context patterns by learning association rules of different dimension labels (Tasks T2.1 and T2.2, conducted by UNITN).

(6) Design, implementation, and initial data analysis of the Mexico pre-pilot experiment, about eating and physical activity patterns of university students, collected with the iLog mobile application (conducted by IPICYT, IDIAP, and UNITN, as joint work between WP2 and WP7).

The deliverable systematically presents each of the outcomes described above in separate sections, and concludes with some final remarks.

© 2019-2022 WENET

Co-funded by the Horizon 2020
Framework Programme of the European Union

CONTENTS

© 2019-2022 WENET

Co-funded by the Horizon 2020
Framework Programme of the European Union

LIST OF FIGURES

Co-funded by the Horizon 2020
Framework Programme of the European Union

List of Tables

© 2019-2022 WENET

Co-funded by the Horizon 2020
Framework Programme of the European Union

2 LITERATURE REVIEW ON SMARTPHONE SENSING TO SUPPORT YOUTH WELL-BEING, FRAMING UBICOMP RESEARCH THROUGH THE WENET GOALS OF DIVERSITY AND ETHICS

Wearable sensing in health monitoring emerged as a trending topic two decades ago with researchers focusing on using wearable sensors for sensing behavioral patterns, health conditions and lifestyles [48, 85, 89]. But the use of that research in real world settings was rare due to many reasons such as high cost involved in creating such wearable devices, the mindset of people regarding wearable devices, or the inability to distribute devices to wide populations. Hence, many of the research efforts were done in controlled lab settings and yet, these research shed light to the potential for mobile sensing for well-being. With the widespread use of mobile phones and low cost sensors in the 2000s, several of the issues went out of the equation as more and more people, specifically *youth*, embraced mobile phones. Even though with limited sensors, there was an emerging literature regarding the potential and utility of mobile phone sensing for large scale applications. A pioneering study on this regard is Reality Mining [42] that demonstrated the utility of phone sensing to collect contextual data passively (gps, bluetooth traces, app usage, charging events), in the wild (out of the lab setting, with 100+ people), for an extended period of time (1 year). Works such as UbiFit Garden [33], MyExperience [49] further demonstrated the capabilities of mobile phones in processing data obtained with external and internal sensors combined with self-reports for behavioral analysis. The emergence of smartphones with more sensing capabilities and interactions compared to traditional mobile phones injected new momentum into mobile phone sensing research with a shift towards *Smartphone Sensing* [65].

People of various age groups have different lifestyles, behavioral patterns, thought processes, and biological characteristics [32, 96]. Young adults (16-35 years old) go through different circumstances in life compared to older generations, and this is reflected in the activities, social interactions, food preferences, and even the phyisical and mental health conditions they have to face [38, 61, 83, 87, 107, 108]. If we consider young adults, many of them are doing their undergraduate or post-graduate studies, in their early career, in the first few years of marriage, unemployed, or a combination of the aforementioned. Considering this stage in life, challenges like stress, anxiety, depression, obesity, alcohol/smoking/drug addictions, unhealthy food habits are common among youth [59, 60, 88, 107], and the reasons why they face these issues might be different to why someone from another age group would face the same type of issue. Further, young people use social media and smartphones far more than older generations [90], and prior research suggests differences on the way people use the phone depending on age [12, 118]. Hence, the criteria to quantify various health/well-being related conditions of people using smartphone sensing would in principle be different, and needs special consideration for a range of smartphone related issues, from sensor selection, app design, and deployment strategies to data analysis techniques, while keeping in mind the *diversity* regarding the target audiences of smartphone well-being applications. Hence, in this section, we focus on smartphone sensing research that has dealt with health and well-being, specifically of *youth*.

2.1 Technical Approach

2.1.1 Socio-Psychology for Human Behavioral Modelling. We identify two main elements for this review.

Pillars of Data for Smartphone Sensing. In 1938, Lewin [67] proposed a theory for behavioral modelling, today known as Lewin's Field Theory, which can be expressed by the formula: $B = f(P, E)$, where B,P, and E stand for behavior, person, and environment respectively. This theory has been used as the foundation for many socio-psychology studies later. In this theory, environment corresponds to conscious as well as unconscious entities in a human's environment, person corresponds to characteristics of the individual in terms of physicality, mentality, and sociability, and behavior corresponds to changes in the life space of an individual as a result from changes in either the environment, or the person. In summary, this theory suggests that behavior is a function of how a person interacts with the environment.

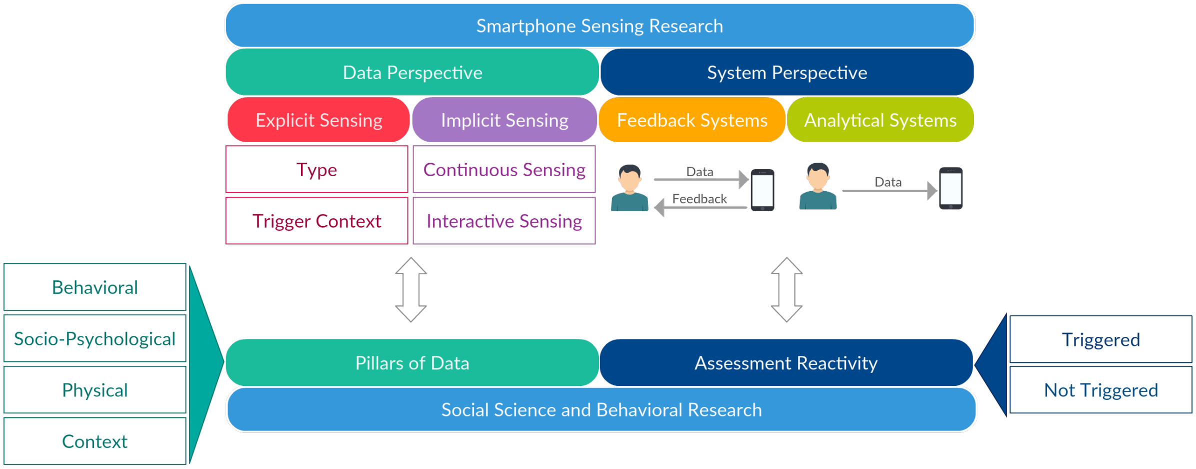


Fig. 1. Taxonomy of Smartphone Sensing Studies for the Well-being of Youth from a Human-Centred Perspective

Drawing motivation from Lewin's field theory, we consider 4 "pillars of data" to study smartphone sensing research for the well-being of young adults, namely (1) Behavior (B), (2) Physical (P), (3) Socio-Psychological (SP), and (4) Contextual (C). We mapped *person* from Lewin's theory into Physical and Socio-Psychological to represent the mental and physical aspects of people, and this mapping allows us to represent sensors and data sources in mobile sensing literature as proxies to different pillars of data. Hence, the equation can be re-written as $B = f(P, SP, C)$, and we use this adaptation of Lewin's Field Theory as a framework for the review. This allows to analyze studies in the domain using a well-grounded framework, where all the sensing datasets collected during the study fall into one or more of the pillars. As an example, data from the accelerometer sensor can be taken as a proxy to physical activity of an individual, hence falling under the physical pillar. But, if accelerometer data is processed to provide more details regarding the activity the user is doing (e.g. walking, running, in a vehicle, etc), this hence provides information regarding the context, and to an extent the behavior. If we consider location sensing as another example, pure location coordinates fall under context pillar. If they are further processed to obtain a semantic meaning, location can be used to determine travel patterns of the users, which can be categorized under the behavioral pillar. As explained by these two examples, various features generated from a single sensor might provide data regarding different pillars. Hence, we use these 4 pillars of data to study the current body of research from the data perspective as explained in section 2.1.2.

Howthorne Effect and Assessment Reactivity Howthorne Effect [77] is the change in a person's behavior due to the effect of being observed, compliance with the wishes of the researchers because of the special attention they have given, or positive response to any stimulus which is introduced. In modern literature, this has been called as Assessment Reactivity [91], or Reactivity [122] in some contexts. In the context of smartphone sensing, if researchers want to trigger assessment reactivity, then the smartphone applications can provide users with feedback regarding their health or behavior. Further, manual interventions can be done by researchers. Then, researchers can measure how well their feedback has influenced the participant app users. Drawing motivation from the concept of assessment reactivity, we study the current body of research from 2 system perspectives as discussed in section 2.1.3.

2.1.2 What is Data Perspective? The Data Perspective considers the data flow in smartphone sensing studies. Under this perspective, we propose a novel taxonomy that involves 2 main components: **Implicit sensing**

Co-funded by the Horizon 2020
Framework Programme of the European Union

involves sensors that acquire data about smartphone users passively (continuous sensing) and based on natural user interactions/usage with/of the phone (interactive sensing); and **Explicit sensing** involves users voluntarily providing data as self-reports through the smartphone application dedicated for sensing. The provided data could belong to either one or more of the pillars mentioned in section 2.1.1. Even though continuous and interactive sensing have been named passive sensing in prior computing literature, we make this distinction to highlight important aspects regarding how smartphone sensing studies on health and well-being can benefit young adults to greater extents.

It is important to understand how datasets are used in smartphone sensing studies. In all the studies, data from different sensors are used as proxies to different phenomena [79] related to context, behavior, physical, and mental aspects of people. For example, if the sensed phenomena are activity level, brightness of the environment, sociability, and indoor mobility, sensing modalities could be accelerometer, ambient light sensor, messaging app usage, and WiFi access point connectivity. Hence, as per this analogy, and pillars of data as given in section 2.1.1, the data perspective of a smartphone sensing study involves obtaining data from the smartphone regarding the 4 pillars of data in behavioral modelling. Implicit sensing directly obtains data regarding these pillars from users passively, while explicit sensing data act as proxies and ground truth for traits belonging to the 4 pillars, and are obtained explicitly from users. Hence, as a summary, smartphone sensing studies for the well-being of youth involves obtaining data from young adults regarding traits belonging to 4 pillars using implicit and explicit means in order to analyze an unknown trait in these pillars.

2.1.3 What is System Perspective? Smartphone Sensing studies involve smartphone app based systems that people use. Using implicit and explicit sensing techniques mentioned in section 2.1.2, smartphones acquire multi-dimensional user data. The majority of such systems stores these information for off-the-shelf analysis that is done at the end of the deployment phase. In contrast, some systems process these data on-device or in cloud, to provide feedback to users regarding their health and well-being state. If we combine features of smartphone sensing studies in computing such as (a) data acquisition by the smartphone; (b) value-added feedback given to users, together with socio-psychology theories regarding assessment reactivity (explained in section 2.1.1), it is possible to segregate literature as FSys, where assessment reactivity is triggered, and ASys, where assessment reactivity should not be triggered.

Feedback systems (FSys) are systems which do data analysis, training, or inference (either in-device or on servers; done manually or in an automated manner) in the deployment phase. These systems often use results of the data analysis or inference, primarily to give user in-app feedback, and also to understand compliance. Some systems provide users with feedback regarding the user behavior, health, or routines in order to motivate them to use the app further [68, 71], develop self insights [69, 100], provide in-app behavioral intervention strategies [93], and for clinical interventions [47]. In other terms, this type of systems affect normal user behavior, and can be used for interventions as well. However, it should also be noted that not all feedback systems are used for interventions, because some apps do not necessarily provide any active intervention even though they summarize sensed details in the app. An ideal Feedback system with triggered assessment reactivity should provide app users with feedback and observe how people benefit from the feedback. Moreover, from a socio-psychological perspective, systems of this sort cannot be used to evaluate a hypothesis regarding the presence or absence or a certain trait, because assessment reactivity is triggered in these studies (hence people know that they are monitored), and it affects the behavior of people.

Analytical systems (ASys) are systems that do data analysis and inference off-the-shelf after collecting the data from the study. These systems often run with no (or less) data analysis during the deployment (with the exception being activity inference from accelerometer data [47, 93, 123, 124]), and even if data analysis is done, results are not directly conveyed to app users, hence avoiding any behavioral change. An ideal ASys should not have any assessment reactivity, and hence smartphone app users should not feel that they are being

Pillars of Data	Continuous Sensing	Interactive Sensing
Behavioral	Accelerometer [69, 71, 98, 123], Proximity Sensor [13], Location [23, 54, 62, 98, 100, 123], Ambient Light [123], Audio [69, 70, 94, 123], Gyroscope [13, 69], Bluetooth [], WiFi [100]	Phone Calls [13, 21–23, 54, 68, 71, 98, 100], Messages [13, 21–23, 54, 68, 71, 98, 100], Email [68], App Usage [13, 68, 81, 99], Browsing History [68], Calendar [], Typing Events [13], Touch Events [], Lock/Unlock or Screen On/Off Events [11, 13, 98, 99, 124], Push Notifications [], Battery Events [69, 99], Other []
Physical	Accelerometer [13, 23, 47, 54, 69, 71, 93, 98–101, 123, 124], Proximity Sensor [], Location [98, 123], Ambient Light [], Audio [], Gyroscope [], Bluetooth [], WiFi [], Other [19, 70, 124]	Phone Calls [], Messages [], Email [], App Usage [], Browsing History [], Calendar [], Typing Events [], Touch Events [], Lock/Unlock or Screen On/Off Events [], Push Notifications [], Battery Events []
Socio-Psychological	Accelerometer [], Proximity Sensor [], Location [], Ambient Light [], Audio [], Gyroscope [], Bluetooth [22], WiFi []	Phone Calls [22, 23, 54], Messages [22, 23, 54], Email [], App Usage [68, 81], Browsing History [], Calendar [], Typing Events [], Touch Events [], Lock/Unlock or Screen On/Off Events [], Push Notifications [], Battery Events [], Other [70]
Contextual	Accelerometer [13, 47], Proximity Sensor [13], Location [13, 19, 23, 28, 47, 54, 62, 68, 71, 93, 98, 99, 101, 123, 124], Ambient Light [13, 124], Audio [69, 94, 100, 123, 124], Gyroscope [13], Bluetooth [13, 21, 22, 80, 99, 123], WiFi [13, 99]	Phone Calls [], Messages [71], Email [], App Usage [71], Browsing History [], Calendar [], Typing Events [13], Touch Events [], Lock/Unlock or Screen On/Off Events [], Push Notifications [], Battery Events [13], Other [70]

Table 1. Implicit sensing in smartphone studies for well-being of youth. Empty bracket [] correspond to cases where an example was not found in the literature.

monitored or controlled in any manner. Moreover, for an ideal ASys, it is better if app users do not necessarily know what specific scientific hypothesis is being tested by researchers, to avoid potential biases in the mind of study participants (given of course that the general purpose of the research is made explicit to obtain informed consent of participants). This would help to obtain better quality datasets that reflect the real behavior of study participants, hence leading to more reliable findings from studies. Hence, there should be minimum interventions throughout the study period.

Table 1 shows how implicit sensing modalities fall into different pillars of data, and Table 2 shows how explicit sensing techniques fall into different pillars of data.

2.2 Discussion

2.2.1 Interactive Sensing to Better Understand Young Adults. Current work in the domain reflects that priority has been given to continuous sensing techniques. It should be understood that these continuous sensing techniques act in a similar way regardless of the user, as in whether they are young adults or older people, and hence, the sensing data can be obtained with good quality regardless of the age group, or any other diversity attributes. But, when we consider interactive sensing techniques, it is mainly based on how people use the smartphone, and interact with it. Prior research [4, 6] suggests that there are significant differences in the smartphone and

Pillars of Data	Type	Trigger Context	Pre and Post Deployment
Behavioral	Structured QnA [11, 13, 69, 80, 93, 94, 98, 99, 123], Pictures from Camera [19, 93], Videos from Camera [], Audio [], Diary [81], Other []	In-situ [11, 19, 99], Retrospective [13, 19, 80, 81, 94, 99], Self-Initiated [80, 94], Reminders [13, 19, 81, 99]	Interviews/Focus Groups [13, 23, 68, 81], Filling Survey [98]
Physical	Structured QnA [11, 19, 93, 98, 123], Pictures from Camera [], Videos from Camera [], Audio [], Diary [69], Other []	In-situ [11, 69, 93], Retrospective [93, 123], Self-Initiated [69], Reminders [123]	Interviews/Focus Groups [], Filling Survey [100]
Socio-Psychological	Structured QnA [11, 21–23, 28, 47, 71, 80, 93, 94, 98, 101, 123, 124], Pictures from Camera [], Videos from Camera [], Audio [], Diary [], Other [11, 68, 81, 93, 100, 123]	In-situ [11, 23, 68, 71, 81, 100, 101, 123, 124], Retrospective [21–23, 28, 80, 94, 101, 123], Self-Initiated [21, 22, 71, 80, 81, 94, 100, 123], Reminders [23, 28, 47, 68, 101]	Interviews/Focus Groups [47, 68, 81], Filling Survey [21, 22, 54, 62, 98, 100, 123, 124]
Contextual	Structured QnA [13, 19, 94, 98, 99, 101, 123], Pictures from Camera [19, 93], Videos from Camera [101], Audio [101], Diary [69], Other []	In-situ [19, 69, 99, 101], Retrospective [13, 19, 94, 99, 101], Self-Initiated [19, 69, 94], Reminders [13, 99]	Interviews/Focus Groups [], Filling Survey [100]

Table 2. Explicit Sensing in Smartphone Sensing Studies for Well-being of Youth. Empty bracket [] corresponds to cases where an example was not found in the literature.

app usage behavior of young adults compared to older generations. We believe researchers should leverage this in order to understand and support young adults better by using interactive sensing. In the current day and age, smartphone applications are tied to young people, and these applications have a direct impact on physical, socio-psychological, behavioral patterns of young adults. Leveraging this prominent characteristic would allow a paradigm shift from continuous sensing to interactive sensing. While we understand the importance of continuous sensing techniques, interactive sensing offers much room for improvement, and a lot of open research questions in smartphone sensing, specially when targeting young adults and their well-being aspects.

2.2.2 Assessment Reactivity in Smartphone Sensing: Trigger or Not? As per prior research, it is unclear whether researchers in computing have adopted principles and techniques from human psychology and behavioral research into smartphone sensing research. Given that smartphone sensing systems are sensing the behavior of young adults, it is essential to understand behavioral dynamics and psychology of young adults. When looking at this problem from the system perspective: (1) If we want to understand behavior in everyday life, if we let app users know the test hypothesis explicitly (e.g. that we are testing for stress, alcohol usage etc.), it might impose a bias on people (e.g. I am reporting that I am stressed everyday, I should relax today, etc.). These biases may get reflected in the data we collect if people alter their behavior, resulting in researchers drawing possibly wrong conclusions from the studies; (2) If we want to understand how the behavior of users changes over time due to the use of smartphone sensing based feedback systems, it is necessary to create a bias in users by using intervention and feedback mechanisms. A good example for this is BeWell [69], which measures how well young people adapt their lifestyle based on feedback they get. Another example is Farhan et al. [47], where they had real-time clinical interventions for participants who reported higher stress levels.

Co-funded by the Horizon 2020 Framework Programme of the European Union

These two types of systems are in-fact the two corners of the spectrum in terms of system perspective, and there can be studies that have features of both. This said, we believe that proper attention should be given to system design and experimentation in the domain of smartphone sensing for young adults by drawing principles from behavioral research, keeping in mind what each type of system offers in terms of testing a hypothesis. Researchers should be aware of how experiments get affected by human bias, behavior, and psychology. As a summary, ASys are suitable to test hypotheses regarding human behavior if done without any assessment reactivity, FSys are suitable to test how behavioral change occurs based on app based feedback or interventions (hence, when assessment reactivity is present).

It should also be highlighted that FSys are better used when the basic hypothesis regarding the human behavior has already been established using ASys or clinical methodologies. For example, if the experiments are done with a FSys (with assessment reactivity) to analyze the relationship between well-being of people and physical activity levels, more often than not it would lead to wrong conclusions because assessment reactivity alters the behavior of people as they know they are being tested for a specific hypothesis. This kind of relationships are better examined first with ASys to establish relationships clinical research, and then tested for occuring behavioral change using FSys.

2.2.3 Lack of Systematic Participant Recruitment Strategies Compared to Clinical Research, Leading to Issues Regarding Scientific Validity. It was particularly highlighted that computer science researchers have not paid enough systematic attention to participant recruitment strategies. Most studies used whoever researchers could find for the study, without a proper recruitment strategy. We believe one open area of research would be to compare and contrast different recruitment strategies that researchers could use by having different participant pools, and recruiting them based on different systematic strategies. Research along this line would not only help research targeting young adults, but also mobile sensing research as a whole. While smartphone sensing studies in our scope have focused mainly on *Cost* in terms of money and time when recruiting participants, socio-psychology research suggest that focus should be on not only cost, but also on aspects such as scientific validity and ties with clinical research [16, 86]. Moreover, in psychology and social science research, in addition to proper recruitment strategies, corrections are used to unbalanced variables such as age, gender, or ethnicity, with rigour in statistical analysis [44]. This makes sure that results provided are scientifically valid.

2.2.4 Diversity-Aware Research in Ubiquitous and Mobile Computing. Diversity in Machine Learning [55] is an important topic which has been growing with popularity during the last few years. While traditional machine learning focuses on data, model, and inference; diversity-aware machine learning has components such as data diversification, model diversification, and inference diversification. Specially in the computer vision domain, these issues have been discussed in depth, where some examples are biases in facial recognition datasets [120], [95], and self driving car accidents and findings that self-driving cars are more likely to hit people of dark skin color [24, 36]. The key drawbacks and findings from our analysis are;

Using diverse people in the sample population in small numbers [68, 69, 93] - Using diverse people in very small numbers would contribute to researchers obtaining different types of data, belonging to diverse people, in small amounts. This might lead to models that learn wrong correlations/features from the data, which might lead to wrong conclusions. For example, in [69], the study had just 27 people in the sample population while it had 9% from CS department, 34% doctors/medical researchers, and 57% graduate students, in an experiment to understand well being. If we consider these sample cohorts, the behavioral routines of undergraduates, doctors, medical researchers and graduate students could be highly different. The question would be whether the models they used considered these diversity aspects of people.

Biased training data might lead to possibly unreliable findings/conclusions [98, 123, 124] - Some studies contain extremely biased sample populations. As concluded by Santani et al. [99], diversity aspects such as gender of young adults could play a huge role in determining their drinking patterns. Yet, some research have used extremely biased gender ratios in their studies which might have lead to biased conclusions. For example, [98] had 18 young people out of which 15 were males. [123] used a dataset of 48 people out of which 38 were males. Hence, it is fair to say that having a more in-depth analysis regarding the results with diversity in mind would have been very informative.

No diverse sample sets, but general claims/conclusions [62] - Another issues is when the sample sets are not diverse enough. For example, stress levels of students could be different based on the classes they are in, and the college subjects which they take. In [62], the authors recruited all the students from a psychology class, while the conclusions they make are regarding all the university students. Whether this is a valid conclusion remains a question.

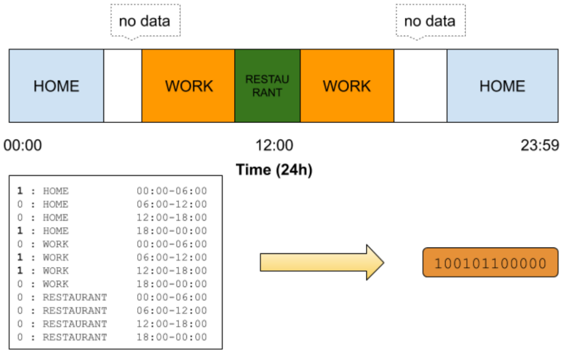
Geographical Diversity - Smartphone usage behavior of young adults in different countries could be different too [76] from the type of apps they use, to the time of phone consumption due to a plethora of factors such as cost of phones, unique lifestyles, culture of the society etc. For example, for youth in western countries, Friday night would be a relaxing day where they drink and party; the situation could be totally different in Asian countries such as India, where drinking is not yet socially accepted. This kind of geographical diversity has not been thoroughly studied in the literature so far, and it could be mainly because most studies were done in one or two countries with limited youth participation. As EmotionSense study [94], Sea Hero Quest [114] have demonstrated, with wide smartphone adoptability, availability of internet in many corners of the world, change in the attitudes of people regarding using smartphones and availability of app based ecosystems, it has now become possible to conduct large scale smartphone based sensing for wider audiences. This also makes it possible to conduct smartphone sensing studies for youth, specifically considering geographical diversity.

3 BASIC ROUTINE MODELS INTEGRATED IN WENET PLATFORM BASED ON SPATIOTEMPORAL BAG-OF-WORDS AND TOPIC MODELS

Routine modelling is an important aspect in behavioural modelling in mobile sensing. Location is one of the most important sensors used in routine models. Routine models help to understand where users spend their time the most, where they travel next and other important aspects regarding the behaviour of people. In the context of WeNet, this way of modelling is important because some WeNet tasks are location-dependent, and hence it is important for the platform to have an idea regarding the routine of users. We developed two kind of routine modelling approaches; (1) **Embedded routines:** consist of abstract representation of the routines, and (2) **Semantic routines:** consist of high-level, human-readable, representation of the routine. For both cases, we use location data as input (collected via the iLog mobile app, and shared after an anonymization procedure implemented by U. Trento, which decreases the precision of the location data). On the current WeNet scenarios, the routine modelling can be useful for (a) filtering volunteers based on routine similarity; (b) suggesting optimal time for certain individual or group activities (e.g. eating) based on group routines.

3.1 Technical Approach and Results

To provide **embedded routines** from the data, we follow 4 steps: (1) Re-sample the locations for each user, for each day. We are interested in where the user is a specific time, but we do not always have the data for this specific time slot. We solve this by interpolating the locations using the median value. At the end of this step, we have well-formed, discretized locations for all days for all users; (2) Create BOW (Bag-Of-Words) - A BOW helps to meaningfully vectorize the previous locations from the above step. To build this BOW, we combine all possible labels with all possible time slots; (3) Train models - As we have BOWs as features, we can use models that come from NLP (Natural Language Processing), such as LDA (Latent Dirichlet Allocation) or HDP (Hierarchical Dirichlet Process). This step can be run each week to update the routine model of the users. At the end of this step, we save the trained models; (4) Predict and save - The final step consist in predicting the routine for each user (based on the previous information) and saving them to a database. The routine is the output of a model learned for a given user.



```
1 : HOME          00:00-06:00
0 : HOME          06:00-12:00
0 : HOME          12:00-18:00
1 : HOME          18:00-00:00
0 : WORK          00:00-06:00
1 : WORK          06:00-12:00
1 : WORK          12:00-18:00
0 : WORK          18:00-00:00
0 : RESTAURANT    00:00-06:00
0 : RESTAURANT    06:00-12:00
0 : RESTAURANT    12:00-18:00
0 : RESTAURANT    18:00-00:00
```

Fig. 2. Example of a BOW representation.

To provide **semantic routine**, the pipeline is slightly different. Instead of having a flattened representation of the days, we group them by weekday. The motivation behind this is that the routines of people are more related to the day of the week. We also changed the output, as we are seeking to create high-level, human-friendly representation. The output of the semantic model is a distribution of probability of each label for a given period, for a given weekday, for a given user. e. g. for the user ID-001 on Friday at 5 pm.

Label	Probability
HOME	0.2
WORK	0
UNIVERSITY	0.6
TRAVELLING	0.2
RESTAURANT	0

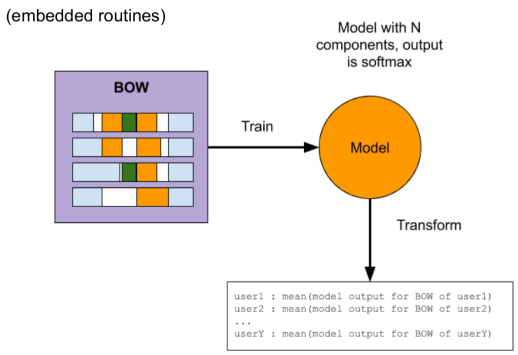
Table 3. Example for a given user, for a given day of the week at a given time.

3.1.1 Integration. We created several APIs to integrate our work with the different partners. We dockerized our component and use CI/CD to update the version and running tests. We may change our interaction with the partners in the future. For instance, instead of offering APIs, our component could send the routines to the partners using their APIs.

3.2 Discussion and Conclusion

Based on the data from the first pre-pilot, the semantic routine modelling is possible for the models we described in this section. As for the embedded routines, we have put them aside for now as partners do not intend to use them at the moment. With data produced by consecutive pre-pilots, we will start more in-depth testing of these algorithms. In the future, we will explore context-aware semantic routine prediction, with additional contextual cues such as activity, real-time locations, and other factors.

Co-funded by the Horizon 2020
Framework Programme of the European Union



```
user1 : mean(model output for BOW of user1)
user2 : mean(model output for BOW of user2)
...
userY : mean(model output for BOW of userY)
```

Fig. 3. Schema of the flow of data to predictions.

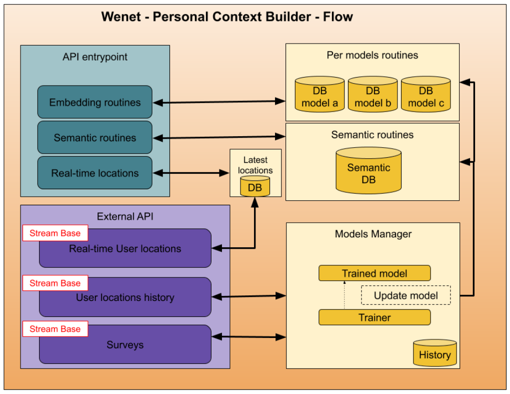


Fig. 4. Flow architecture of the Personal Context Builder component

Method	Path	Description
GET	/routines	Get the list all embedded routines for all users with all models
GET	/routines/{user_id}	Get specific user embedded routine
GET	/models	Get the list of all available models for the embedded routine

Table 4. APIs for Embedded routines.

Method	Path	Description
GET	/semantic_routines/{user_id}/{weekday}/{time}	Get the semantic routine for a given user, weekday and time
GET	/semantic_routines_transition/leaving/{user_id}/{weekday}/{label}	Get the information about what when the user_id is leaving the label on the given weekday
GET	/semantic_routines_transition/entering/{user_id}/{weekday}/{label}	Get the information about what when the user_id is entering the label on the given weekday

Table 5. APIs for Semantic routines.

4 WORK ON ANALYSIS OF MOBILE FOOD JOURNALS TO PERFORM APPLICATION-RELATED INFERENCES WHILE PRESERVING DIVERSITY ATTRIBUTES

There is an increasing interest in smartphone applications that use passive sensing to support human health and well-being. Such applications primarily rely on generating low-dimensional representations from raw data, making inferences regarding user behaviour, and using those inferences to benefit application users, while sometimes datasets are shared with third parties as well. The goal of these applications is to increase the utility of application-related inferences, while modern ubiquitous systems must also ensure that sensitive attributes regarding users are preserved. In this section, we analyse the eating behaviour of 122 university students, and show a potential privacy risk, namely the possibility of inferring the gender of students with an accuracy of 77% using low-dimensional and sparse data obtained by processing high-resolution mobile sensing data and self-reports. Then, we demonstrate how deep learning techniques can be used for feature transformation to preserve the sensitive information of users (decreasing the accuracy for gender inference to a random guess (48%)) while achieving high accuracies for application inferences. We believe that researchers should be aware of these implications and take necessary precautions to preserve sensitive information of mobile food diary users from unexpected results when creating machine learning pipelines, storing data, and sharing even anonymized data with third parties.

There is a booming industry around mobile sensing and behavioural analysis applications for human health and well-being. There are studies that attempt to infer health and well-being related attributes such as stress [52, 70, 98], emotions and mood [94, 106], well-being [66, 69, 123], alcohol consumption [13, 99], and other behaviour [93] using smartphone sensor data and self-reports. In the specific case of food and nutrition, even though many studies demonstrate the potential to identify eating occasions using wearable cameras and sensors [15, 109, 110], relatively less attention has been given to explore food consumption behaviour of people primarily using smartphone sensing systems, even though there is evidence in nutrition research that establishes that eating practices are embedded into everyday routines of people [20, 35, 37, 78].

Most commercial mobile food diary based health and well-being applications such as Samsung Health [10], Google Fit [7], and Apple Health [8] passively sense activity information regarding users by transforming high-dimensional sensor data from accelerometer, location, gyroscope, and other sensors into low-dimensional features such as step count, semantic location, and activity type. Moreover, they collect data regarding food intake as food diaries. Such applications usually provide an option for the users to provide sensitive information such as gender, body mass index (BMI), and age claiming that if they provide such data, personalised services could be provided with better quality of service [3, 5, 9]. In practice, there are some users who provide such data, while other users may prefer to use the application without providing sensitive information, and it should be understood that this represents a trade-off among personalization, privacy, and utility when using applications and services [29, 111, 121]. How this conundrum plays a role in ubiquitous computing is described in a recent paper [14], which emphasizes the need for privacy preserving pervasive systems. Moreover, according to the terms of use of several mobile health applications [3, 5, 9], this is exactly why they use personalised algorithms for users who provide such sensitive information, and have non-personalised algorithms for users who refuse to provide such data, but still opt to use the mobile health app.

We use the term "sensitive inference" for inferences of information regarding mobile health application users. These inferences might reveal private or health related information regarding users (e.g. gender, BMI, weight, height, etc.), and hence are sensitive in nature. Hence, we first examine whether mobile food diaries might reveal such sensitive information, and if yes, how to mitigate such risks. According to many personal data related guidelines, personal data or sensitive information could be data that can be connected to a person, even if the possibility of identifying a person based on such data is limited [1, 2, 115]. Following this notion, there is an increased possibility of identifying an individual due to inferring more sensitive attributes regarding users, in addition to all the non-sensitive data that are already available.

We use the term "application inference" for inferences that are done in mobile food diaries and mobile health applications to benefit users. A basic example is inferring the step count using raw accelerometer traces, where high-dimensional data are used in the inference process. In the context of this research, we use three useful inferences done using low-dimensional data such as; (1) meal vs. snack, which focuses on inferring if an app user takes a meal or a snack at a particular moment [19]; (2) sweet vs. non-sweet food inference, which focuses on inferring if an user would take sugary food or not [25, 45, 64]; and (3) dairy vs. non-dairy inference, which focuses on inferring whether users take dairy food with their meals or snacks [17, 75]. Hence, if most of these food intake related variables can be inferred using contextual and activity related data, it would be valuable for future food diary based mobile health applications [19, 101, 110].

Co-funded by the Horizon 2020
Framework Programme of the European Union

4.1 Technical Approach and Results

In this work, we used a smartphone sensing dataset called *Bites'n'Bits* collected in a previous project [19, 50]. It contains smartphone sensor data, self-reported data, and activity data from fitbit wearables from 122 students of a Swiss university. The smartphone application allowed users to report their food intake by uploading a photo. Moreover, users reported food type, the context of eating by describing factors such as with whom they were eating (e.g. date, family, friends, alone, etc.), what they were doing while eating (e.g. commuting, reading, watching TV etc.), and also semantic location (e.g. university, restaurant, home, etc.). Further, their activity levels were captured using a fitbit wearable. After the end of the data collection campaign, there were 1208 food user days from 122 users, 3414 meal reports, 1034 snack reports, 5097 photo reports, and 998 fitbit user days. In addition, basic demographic data and additional information were provided by each participant, including age (in years) and sex (men/woman). In this study, we use the term "gender" to designate the variable *sex*, aware of the fact that the two terms are not identical [112, 129]. Furthermore, participants self-reported their BMI, an indicator commonly used in nutrition and health which is also a personal health related attribute. All the students who took part in the study were between 18-26 in age, with a mean age of 20.5 years, and there were 65% men and 35% women.

Initially, we trained and tested the MT-NN using binary cross entropy loss function for both sensitive inference and application inference. Then we created an AE with an equal number of dense neurons in the input/output layers (also equal to the number of features in the dataset); with 12,10,8,10,12 dense-neurons in each intermediate layer, elu activations for intermediate layers, and sigmoid activations for the output layer. The AE and MT-NN based architecture is depicted graphically in Figure 5. We locked the weights of the MT-NN so that weights do not get tuned during the training process, and then trained the AE using the training dataset.

If we define our dataset as X_n, the two functions for sensitive and application inferences can be defined as $F_{sen}(.)$ and $F_{app}(.)$. The objective is to find a feature transformation function for AE as $F_{ae}(.)$ where the resultant dataset from the autoencoder is $X_n^* = F_{ae}(X_n)$ such that $F_{sen}(X_n^*)$ accuracy is not high (closer to 50%), hence preserving sensitive attributes about users, and $F_{app}(X_n^*)$ is high (closer to 100%), providing high inference accuracies for application inferences from the same dataset. In the training phase of the AE, for a given data point B_i, the output of the MT-NN for the sensitive inference would be $F_{sen}(B_i)$, and the application inference output would be $F_{app}(B_i)$ whereas the two losses are indicated by Equations 1 and Equation 2 respectively. The objective of the autoencoder is represented by Equation 3 which combines the losses from the two inferences in the MT-NN, and aims at minimising the loss for the training dataset.

$$L_{sen} = |0.5 - F_{sen}(B_i)| \tag{1}$$

$$L_{app} = -(F_{app}(B_i) \times log(p) + (1 - F_{app}(B_i)) \times log(1-p)) \tag{2}$$

$$F_{obj} = \arg\min_{B_i}(L_{sen} - L_{app}) \tag{3}$$

To make sure that the AE learns its weights and biases to create a dataset that provides higher accuracies for application inference and lower accuracies for sensitive inferences, we used a modified loss function as in Equation 1 for gender (we use the value 0.5 because it is desired accuracy for the binary classification task to make sure that it has a lower accuracy), and traditional binary cross entropy (given in Equation 2) for application inference. Hence, the loss for the AE was derived from the two output losses of the MT-NN as given in Equation 3, whereas no matter how high the loss for gender classification is, it is not conveyed as it is to the AE due to the modified objective. This makes sure that the AE tunes its weights such that resultant dataset after the feature transformation does not care about the accuracy of sensitive inferences, and the features are transformed to ensure reasonable accuracies for application inferences. After the training process, we obtain the testing dataset with transformed features using the AE, and this dataset is ideal to potentially store in cloud services or to share with third parties (of course when the user agrees with such use). For the experiment, we used gender as an example of sensitive inference, and meal_snack, sweet, and dairy as application inferences for the three MT-NNs.

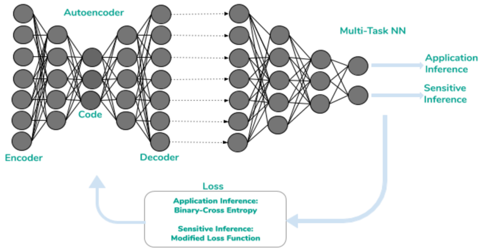

Autoencoder
Multi-Task NN
Application Inference
Sensitive Inference
Code
Encoder
Decoder
Loss
Application Inference: Binary-Cross Entropy
Sensitive Inference: Modified Loss Function

Fig. 5. AE and MT-NN based architecture for privacy preserving feature transformation. The output of the AE is directly mapped to the input of MT-NN. The AE's loss function is based on the losses of sensitive inference and application inference.

Feature Group	Classification	MT-NN Before AE	MT-NN After AE	RF Before AE	RF After AE
C+A	Meal vs. Snack	86%	81%	86%	85%
	Male vs. Female	67%	51%	77%	48%
C+A+D	Meal vs. Snack	86%	82%	88%	84%
	Male vs. Female	75%	47%	91%	54%
C+A	Sweet vs. Non-Sweet	83%	79%	82%	81%
	Male vs. Female	69%	53%	77%	48%
C+A+D	Sweet vs. Non-Sweet	86%	80%	84%	81%
	Male vs. Female	76%	51%	91%	56%
C+A	Dairy vs. Non-Dairy	78%	78%	73%	71%
	Male vs. Female	76%	51%	77%	57%
C+A+D	Dairy vs. Non-Dairy	82%	80%	87%	82%
	Male vs. Female	81%	47%	91%	53%

Table 6. Accuracy for Application Inferences and Gender Inference using MT-NN and RF, before and after feature transformation using the AE.

After training the AE to transform dataset features so that sensitive inferences are made difficult following the procedure given, we measure both the application inference and sensitive inference accuracies for the transformed dataset using newly trained RF and MT-NN. Table 6 shows the results using a comparison between accuracy results before and after the use of autoencoder for MT-NN and RF, for all three inference tasks. Application inference accuracies have been kept reasonably high for all three inference tasks (above 81% for MT-NN and 85% for RF in meal vs. snack, and similar results hold for other two application inferences as well). At the same time, we were able to reduce the gender inference accuracy from 67% to 51%, 77% to 48%, 75% to 47%, and 91% to 54% for MT-NN for C+A, RF for C+A, MT-NN for C+A+D, and RF for C+A+D respectively, when using gender inference together with meal vs. snack inference. A similar pattern in results can be seen for other two applications inferences, showing the generalisation of our approach to different application inferences. Importantly, the output from this procedure is still low-dimensional (similar to the original dataset), but also privacy-preserving because gender attribute can not be directly inferred from the resulting features even if a model is newly trained.

4.2 Discussion

Using Feature Transformation Techniques on High Dimensional vs. Low Dimensional Data. We emphasise the importance of understanding the difference between high-dimensional/high-resolution data and low-dimensional/low-resolution data. High-dimensional data either spans many features by design, or are raw traces that can be processed in numerous ways to generate novel features. For example, raw data traces of accelerometer are typically high-resolution because these values preserve detailed information regarding the context at which data were collected. As revealed in prior research [26, 63, 84, 119, 119, 130], these high-resolution traces can be used to discriminate attributes like gender using neural networks. Moreover, such data allows to engineer novel features (e.g. statistics of accelerometer trace along x,y and z axes, activity types in certain time windows, activity levels in time windows, step counts, etc). Hence, even though privacy-preserving techniques such as [72–74] can be used for these data traces, they still could allow the creation of novel features of low-dimensionality and low-resolution, which would in turn still contain discriminative features to identify potentially sensitive information of mobile app users. Further, if we just consider the dimensionality of raw data traces, the higher the number and diversity of features in the data, the higher the potential amount of information available in the dataset, thus increasing the ability of discriminating sensitive attributes.

On the other hand, low-dimensional or low-resolution datasets are already processed in some way, reducing the information embedded in them, and thus typically have smaller number of features. For example, the step count of a person is derived by processing high-resolution data traces from the accelerometer and gyroscope sensors where many features (x,y,z axis of accelerometer and gyroscope, time) are combined to derive one single value i.e. the step count in a particular time window. Because step counts are low-resolution, it is comparatively difficult to engineer more features by processing them with different techniques. Therefore, from our study findings, we advocate the idea that preserving sensitive attributes from high-dimensional or high-resolution datasets might have some limitation if novel discriminative features can still be generated. Moreover, preserving sensitive attributes from low-dimensional or low-resolution data might mitigate privacy risks to a larger extent. Researchers and developers who use mobile sensing datasets should be aware of these findings, specially when they store data, or share data with other parties or to the public.

Datasets Before and After Feature Transformation. The feature transformation process proposed in this work makes significant changes to dataset features after transformation. One such change is the conversion of categorical variables to numerical variables. For example, during an experiment, the dataset had two values each for the categorical variables "withwhom-mod_family", "withwhom-mod_friends" and "withwhom-mod_date" before the transformation, and after the transformation resulted in 2492, 4751, and 3481 unique values respectively. This is because the feature transformation happens to each data row separately, and not to each column separately unaware of the categorical nature of the dataset. Hence, it creates a situation where the dataset after feature transformation would be uninterpretable unless users of the dataset have prior knowledge regarding the feature transformation process. However, this naturally protects the dataset from privacy risks from third parties who may access the data. For example, if a transformed dataset is shared with a third party by the data owner together with instructions for creating useful application inferences, it is difficult for the third party to interpret data for purposes other than what they were informed of from the data owner. As another example, if the data is stored in the processed form by the data owner, even if the dataset falls into the hand of a third party (e.g. due to a data breach, etc), since the data is only interpretable to the data owners, the dataset would become useless for the third party. Hence, the techniques we propose would create uninterpretable datasets for sharing and storage, and it would ensure that the data are used only for the required purposes.

Dataset Diversity. A limitation of our study is the relative homogeneity in the study cohort in terms of participants. The dataset used is from one signle university, while the eating habits of students might be different across universities in other world regions. While the results show initial evidence of gender-specific eating behaviour, and that a feature transformation technique can preserve privacy, we believe that conducting a larger scale experiment across countries with different behavioural habits would shed more light into the preliminary evidence we present here. This multi-site effort, within and outside Europe, is indeed one of the objectives of WeNet. We hypothesise that even though using more diverse user populations might demonstrate varieties of eating behaviours, the privacy preservation technique we have proposed might still be useful as it is agnostic to the dataset used. This said, we also speculate that additional work will be needed to understand whether country-specific models are more appropriate than multi-country models.

Personalisation, Privacy, and Utility. Researchers in computing usually strive to enhance utility of applications and algorithms, and often use personalisation as a tool to increase utility. While this is important for the advancement of technology in many disciplines, recent trends emphasize the priority of privacy [14, 18, 43, 74]. Personalisation and privacy preservation could be seen at the two ends of the spectrum because personalisation requires more personal data to provide high utility, while privacy preservation aims at having a reasonable utility from the application, while preserving privacy from known and unknown risks. The trade-offs between these variables also influence people who value different aspects while using mobile health applications, and online applications in general. Hence, it should be understood that while some users might prefer to distribute their personal information and health related information for personalised services, there are other users who have concerns regarding application developers, and also regarding how their personal data might be used. Hence, in this work we discussed the trade-off between utility and privacy. As seen from the results above, application inference utility slightly drops when privacy is preserved (after feature transformation), and application inference (and sensitive inference) accuracies increase when adding more sensitive attributes to the dataset (C+A+D feature group instead of C+A feature group). While personalisation of algorithms and services have a place in the ecosystem of mobile computing, we endorse the idea that app users, app developers, and data owners should be aware of the risks they might face when sharing and storing personal information under foreseen and unforeseen circumstances. We believe that having a clear idea about personalisation, privacy, and utility is important for the advancement of the field in an ethical manner. Recent literature in pervasive computing further explains this conundrum, and discusses why privacy preservation techniques are important for modern ubiquitous system by pointing out to the fact that simple anonymization techniques are no longer enough to preserve user privacy [14].

5 WORK ON SKEPTICAL LEARNING TO COPE WITH UNTRUTHFULNESS OF HUMAN ANNOTATION DURING REAL-TIME REPORTING OF ACTIVITIES USING MOBILE APPLICATIONS

A key aspect in learning accurate user profiles is the reliability of user feedback. In WeNet, one of the main ways to collect user feedback is by administering periodic questionnaires to users about aspects like current location, activity, social interactions and so on. However, the reliabililty of this type of self-reported information will typically be rather low. Research in the Social Sciences provides evidence of the unreliability of people when required to compile self-reports such as time diaries [105] describing their behavior. A preliminary study on the habits of a group of students at the University of Trento confirmed these findings, showing a substantial proportion of mislabellings in the annotations, with largely different behaviours among students. While modern machine learning approaches are conceived to be robust to label noise, this robustness critically depends on the amount of noise in the training data [40, 82] and can be severely affected by the amount of noise that is found in these self-reported data.

In order to overcome these problems, we developed a novel learning framework named *skeptical learning* [127, 128]. In skeptical learning the machine interactively learns a user profile from sensor data and user feedback. As learning proceeds, the machine increases the confidence in its own predictions, and starts acting skeptically and questioning user feeback when it sees it as inconsistent with its own judgment. A conflict resolution strategy allows the machine and the user to find a consensus between their respective beliefs, and learning proceeds.

In the following we describe the skeptical learning framework, its implementation on top of the i-log activity logging application and report experimental results showing substantial advantages over non-skeptical alternatives and identifying a number of prototypical users.

5.1 Technical approach

Figure 6 shows the skeptical learning architecture (SSML stands for Skeptical Supervised Machine Learning). Sensor data continuously collected by the logging application are processed by a set of predictors (machine learning algorithms) predicting properties of interest for characterizing the user, e.g. the current location or activity. Both sensor data and predicted labels are stored in memory (stream data storage) where they enrich

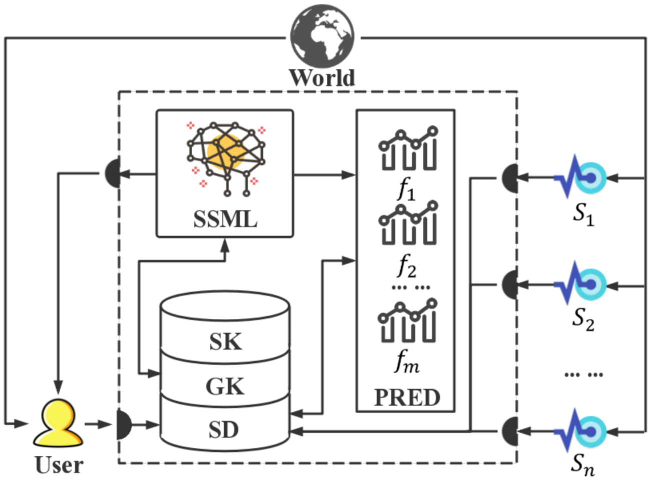


Fig. 6. The SSML Architecture. The disk-like component marked with SK, GK, SD contains the prior knowledge (SK, GK) and the streaming data coming from the sensors and the user labeling activity (SD); PRED is the main machine learning algorithm which in turns is an ensemble of simpler machine learning algorithms; SSML is the main Skeptical Supervised Machine Learning algorithm.

the knowledge of the system, integrating the schematic and ground knowledge components. The SSML module is the core reasoning component of the architecture. It gathers predictions from sensors, decides whether to query the user for feedback on the properties being predicted, and whether to be skeptical on the user feedback. The pseudocode of SSML is shown in Algorithm 1, where a single property with a set of $\mathcal{Y}$ possible values is assumed for simplicity (i.e. a multiclass classification task). The algorithm can be in one of three modalities. The *train* mode is the initial one, where the algorithm behaves as a standard (online) supervised learning algorithm, collecting user feedback for each sensor reading. In the *refine* mode, the algorithm checks the quality of the user feedback and depending on its own confidence can question the feeback. In the *regime* mode the algorithm starts being autonomous and only queries the user for particularly ambiguous instances.

The algorithm takes as input a confidence threshold θ, and keeps a confidence vector for the user ($|\mathbf{c}^{\mathrm{u}}|$, initialized to one for all classes) and the predictor ($|\mathbf{c}^{\mathrm{p}}|$, initialized to zero for all classes). Then the training phase begins. The algorithm collects sensor readings ($\mathbf{x}_t$), passes them to the predictor that returns the highest scoring class ($\hat{y}_t$) and asks the user for her own label (y_t). The user feedback is used to update the predictor and and its confidence $|\mathbf{c}^{\mathrm{p}}|$. When the predictor is confident enough to start challenging the user on the correctness of a certain labeling, the training stage is stopped and the system enters the refinement mode. In this mode, the system keeps asking the user for labels, but it starts to compare them with its own predictions.

Algorithm 2 shows the SOLVECONFLICT procedure which deals with this comparison. The procedure compares the predictor and the user label according to the ISCOMPATIBLE procedure. In the simplest case, this outputs true if the two labels are identical, and false otherwise. In more complex scenarios, this procedure can use existing knowledge, as stored in the SK or in the GK, to decide whether two distinct labels are compatible, e.g., if the concept denoted by one is a generalization of the concept denoted by the other. In case the labels are compatible, a consensus label is taken as the ground truth, and the predictor and user confidences are updated accordingly. A natural choice for the consensus (and the one we use in our experiments) is being conservative and choosing the least general generalization of the two concepts. A labeling conflict arises in the case of the two

Algorithm 1. Skeptical Supervised Learning (SSML)

```
procedure SSML($\theta$)
    init $\mathbf{c}^{\mathrm{u}} = 1$, $\mathbf{c}^{\mathrm{p}} = 0$
    while TRAINMODE($\mathbf{c}^{\mathrm{p}}, \mathbf{c}^{\mathrm{u}}, \theta$) do
        $\mathbf{x}_t$ = SENSORREADING()
        $y_t$ = ASKUSER()
        $\hat{y}_t$ = PRED($\mathbf{x}_t$)
        TRAIN($\mathbf{x}_t, y_t$)
        UPDATE($\mathbf{c}^{\mathrm{p}}, \hat{y}_t, y_t$)
    while REFINEMODE($\mathbf{c}^{\mathrm{p}}, \mathbf{c}^{\mathrm{u}}, \theta$) do
        $\mathbf{x}_t$ = SENSORREADING()
        $y_t$ = ASKUSER()
        $\hat{y}_t$ = PRED($\mathbf{x}_t$)
        SOLVECONFLICT($\mathbf{c}^{\mathrm{p}}, \mathbf{c}^{\mathrm{u}}, \mathbf{x}_t, \hat{y}_t, y_t$)
    while True do
        $\mathbf{x}_t$ = SENSORREADING()
        $\hat{y}_t$ = PRED($\mathbf{x}_t$)
        if CONF($\mathbf{x}_t, \hat{y}_t, c^{\mathrm{p}}_{\hat{y}_t}$) $\leq \theta$ then
            $y_t$ = ASKUSER($\hat{y}_t$)
            SOLVECONFLICT($\mathbf{c}^{\mathrm{p}}, \mathbf{c}^{\mathrm{u}}, \mathbf{x}_t, \hat{y}_t, y_t$)
```

Algorithm 2. Procedure for solving labeling conflicts.

```
procedure SOLVECONFLICT($\mathbf{c}^{\mathrm{p}}, \mathbf{c}^{\mathrm{u}}, \mathbf{x}, \hat{y}, y$)
    if ISCOMPATIBLE($\hat{y}, y$) then
        $y^*$ = CONSENSUS($\hat{y}, y$)
        UPDATE($\mathbf{c}^{\mathrm{p}}, \hat{y}, y^*$)
        UPDATE($\mathbf{c}^{\mathrm{u}}, y, y^*$)
    else if CONF($\mathbf{x}, \hat{y}, c^{\mathrm{p}}_{\hat{y}}$) $\leq c^{\mathrm{u}}_{y} \cdot \theta$ then
        TRAIN($f, \mathbf{x}, y$)
        UPDATE($\mathbf{c}^{\mathrm{p}}, \hat{y}, y$)
    else
        $y^*$ = ASKUSER($\hat{y}, y$)
        if $not$ ISCOMPATIBLE($\hat{y}, y^*$) then
            TRAIN($\mathbf{x}_t, y^*$)
        UPDATE($\mathbf{c}^{\mathrm{p}}, \hat{y}, y^*$)
        UPDATE($\mathbf{c}^{\mathrm{u}}, y, y^*$)
```

labels are not compatible. In case the confidence of the prediction is not large enough to contradict the user, the user label is taken as ground truth, the predictor is retrained with this additional feedback, and its confidence is updated accordingly. Otherwise, the system queries the user providing the two conflicting labels as input, asking her to solve the conflict. The user is free to stick to her own label, change her mind and opt for the label suggested by the predictor, or provide a third label as a compromise. As we are assuming a non-adversarial the

Table 7. Table showing the location labels that the users in the experiment could select and the mapping with the three superclasses we defined.

Bar	Gym	Shop	Outdoors	Workplace	Other Home	Home	Class	Canteen	Study hall	Library
Others						Home	University			

setting, the system eventually trusts the newly provided label (even if unchanged) which becomes the ground truth. At this point, a compatibility check is made in order to verify whether a retrain step is needed, and the predictor and user confidences are updated.

The refinement stage is stopped when the predictor is confident enough to stop asking for feedback to the user on every input, but selectively query the user on "difficult" cases. When leaving the refine mode, the system enters the regime one, where it remains indefinitely. In this mode, the system stops asking feedback for all inputs, and an (online) active learning strategy begins, combined with the SOLVECONFLICT procedure to deal with the comparison between the predicted and the user labels.

5.1.1 Experimental protocol. The evaluation of the SSML framework is based on a data set collected in an experiment with main objective was to understand the empirical gap concerning students' time allocation and academic performance. The data was collected using the i-Log mobile application [125] that can simultaneously acquire data from up to thirty sensors on the smartphone, both hardware (e.g., GPS) and software (e.g., running applications). The i-Log also allows to administer time diaries to the participants asking about their activities, location and social relations at fixed time intervals. The collected answers are then used as the user's labels in the machine learning algorithms of the SSML. We decided to focus on location labels for this experiment, as this is the property for which we could produce a reliable oracle emulating the ground truth labels (as opposed to the noisy labels provided by the user annotations). The first row or Table 7 presents the pre-defined labels that the user could choose one of them as their current location annotation.

The data processing stage generated a set of 122 feature vectors for each user, using all available sensors inputs. The features were calculated using a window size of 30 minutes, which is the time between two consecutive annotations. Our analysis focuses on locations because it is easier to verify the correctness of such data with respect to activities or social relations data. To this aim, we created an additional element, *the oracle*, which provides ground truth labels independently of both the predictor and the user annotations. The oracle relies on information regarding the location of the University buildings, and identifies the home of a user by clustering the locations she labels as home via DBSCAN [46] and choosing the cluster where she spends most of the time during the night. Note that SSML has no access to this information. The oracle is used for the evaluation of the performance of the system in predicting actual labels, and to simulate a non-adversarial, collaborative user as detailed in the next section.

We implemented the predictor as a random forest classifier [27] (with batch training), which is robust to labeling noise [51], in order to evaluate the ability of SSML to improve over an already noise-robust baseline. For simplicity, we used an infinite window ($d = \infty$) for the confidence update, also given the relatively short duration of the experiment. The confidence parameter θ was set to 0.2 in order to achieve a reasonable trade-off between accurate training and reasonable cognitive effort for the user, considering the complexity of the learning task.

5.2 Results and discussion

We firstly evaluated the robustness of SSML and compare the its performance with a non-skeptical algorithm. In this experiment, user's labels are replaced with three general labels (the second row of Tabel 7), namely *Home*, *University* and *Others*, basing on the semantic of user's original labels. The non-skeptical alternative is a solution that never contradicts the user. It is obtained by replacing the SOLVECONFLICT procedure with only train and update steps. This non-skeptical solution is refered as SML (Supervised Machine Learning). Figure 7 reports the comparing of performance between SSML and SML. The time axis represents the number of iterations that the algorithm is going through. Figure 7(a) reports the f_1 score of the SSML and SML predictors with increasing time. We use the latest 15% of the all data available for each user as test data set, and the f_1 score at each time was computed on this test set. This provides an estimate of the performance of the algorithms when doing predictions on future data. Figure 7(b) reports the number of queries that SSML and SML respectively send to the user (red solid and red dashed). According to these two figures, the results shows that the system can achieve a 34.2% relative improvement in performance (from $f_1 = 0.38$ to $f_1 = 0.51$) with a slight increase of queries to the user. It also reports, for the SSML case, that most of the times that the user is contradicted by SSML (green), she ends up agreeing with the predictor (brown). In other words, it turns out that the machine's prediction is aligned with the oracle label in most of the cases in which the machine challenges the user.

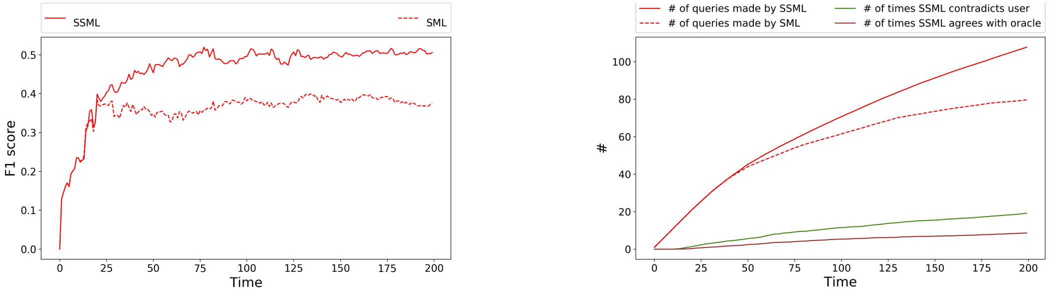


Fig. 7. The comparing between SSML and SML: (a) f_1 scores for an increasing number of iterations, for SSML (solid red) and SML (dashed red) respectively. (b) number of queries made by SSML (solid red) and SML (dashed red), number of times the user is contradicted by SSML (green) and number of times SSML ends up being right (brown). The Time axis represents the number of iterations the algorithm is going through.

By looking at the performance of each individual user, we notice that they have very different behaviours and we can clearly identify a number of prototypical users. Therefore, we secondly investigate the performance of SSML and SML on two type of labels, namely objective labels and subjective labels. The *objective labels* are the ones provided by oracle, while the *subjective labels* are provided by the user. Analysing the performance of different users by identified four patterns as highly common. Each row in Figure 8 refers to a specific user. The figures on the left side show the f_1 score in different settings (i.e., objective vs subjective labels, SSML vs SML). The figures on the right side report information on the number of queries and agreement with the user and with the oracle, same as in Figure 7(b).

Inattentive User The first row of Figure 8 shows the performance of an *inattentive user*. The highest score is achieved by the SSML algorithm evaluated on objective labels, which means the user often provides subjective labels that are very different from the objective labels. (difference between red and blue curves in the left graph). The user's inconsistency is also reflected in the right graph. The SSML contradicts the user quit often (the green curve showing the high number of queries), and almost half times the SSML agrees with the oracle. Therefore, this type of user is a "detectable" inconsistent one and she benefits the most from SSML.

Predictable User In the second case, the highest score is achieved by SSML on subjective labels in the initial phase, and at a certain point SSML learns to predict objective labels with a higher accuracy with respect to subjective ones (crossing between blue solid and red solid). This happens because the user is consistent in providing feedback, but her subjective labels are largely different from the objective ones. We refer to this user as "predictable". When the system receives the appropriate feedback, objective labels can be

Co-funded by the Horizon 2020
Framework Programme of the European Union

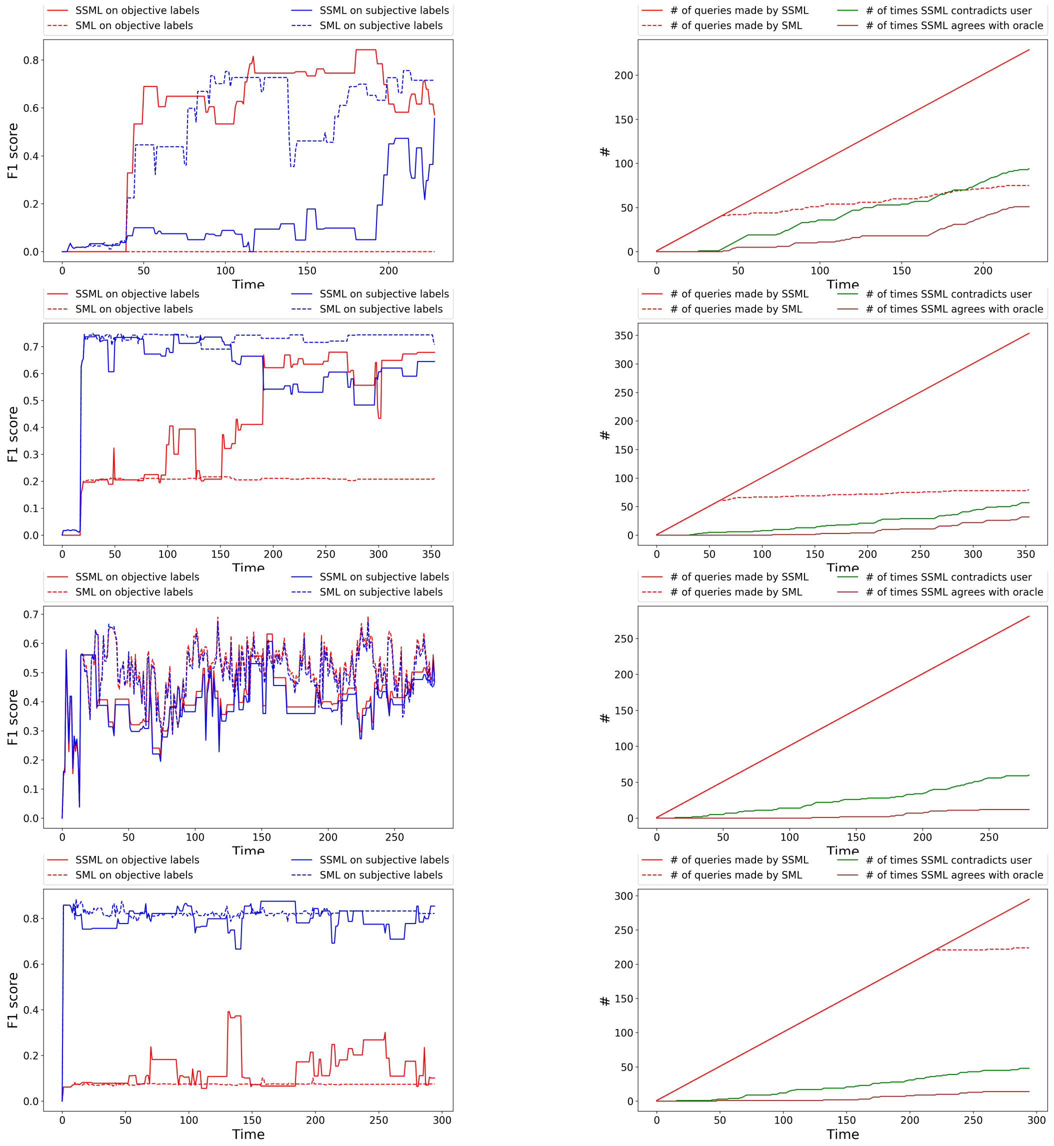


Fig. 8. Results for four different prototypical users: from the first row to the last row they are *inattentive user*, *predictable user*, *reliable user* and *tricksy user*. The images on the left report the f_1 scores in different settings while the ones on the right report information about the number of queries and agreement with the user and the oracle. The Time axis represents the number of iterations the algorithm is going through.

predicted with high accuracy. A predictable user is thus another case in which the benefits of SSML are substantial, even if it takes some time for the system to figure out the discrepancy between subjective and objective labels.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Reliable User The third row of Figure 8 shows that, for this user, the performance of the SSML on objective and subjective labels are roughly the same and have similar trend. This is because the user is already reliable in providing initial feedback, as can be seen by the substantial overlap between the red and blue curves. Indeed, the user is contradicted only occasionally (green curve in the right figure), and even rarer are the cases in which the oracle agrees with the predictor against the subjective label of the user (brown curve). This is a user for whom SSML is not helpful, but also not harmful.

Tricksy User The last one is a case in which the SSML algorithm completely fails to predict user actual behavior. The big gap between between blue and red curves shows the difference performance between subjective and objective labels. The algorithm keeps learning from subjective labels, even when it goes into the next mode and is given the chance to question user labeling. The right figure shows that this chance is rarely taken by the algorithm, and almost never leads to discovering the cases in which subjective and objective labels disagree. The user here succeeds in fooling the system by convincing it of the correctness of her own feedback.

This section described the skeptical learning framework, a framework that we developed in order to account for the unreliability of user feedback when learning user profiles from sensor data. We reported preliminary experiments showing how skeptical learning is capable of identifying inaccurate feedback from the user with good accuracy and how the utility of the framework strongly depends on the type of user being interacted with. This research led to two publications where further details can be found:

- Zeni, M., Zhang, W., Bignotti, E., Passerini, A. and Giunchiglia, F., *Fixing Mislabeling by Human Annotators Leveraging Conflict Resolution and Prior Knowledge.* Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 3, 1, Article 32 (March 2019).
- Zhang, W., Passerini, A. and Giunchiglia, F. *Dealing with Mislabeling via Interactive Machine Learning.* Künstl Intell (2020).

As next steps we plan to further improve the reliability of the system by a more principled modeling of the uncertainty of the predictor, and to integrate this module into the WeNet platform.

6 WORK ON USER CONTEXT RECOGNITION USING AN ONTOLOGY-BASED MODEL OF DAILY LIFE

A core goal of WP2 is to develop diversity-aware approaches for learning the patterns and routines of users from data. These are instrumental for describing and analyzing diversity of individuals and groups of individuals. The notion of personal context is key in this endeavor. Loosely speaking, "context" refers to any kind of information necessary for characterizing an individual [39]. Here we are concerned with formal, ontology-based models of context. A well-defined context model is necessary to study how people perceive their own state and each other's states, and hence how social ties are built. The context model also supplies a necessary bridge between users and machines and it forms the basis for human-machine communication, in particular for transferring the user's requirements to the machine.

The **first goal** of this section is to outline the ongoing work on user *context modeling* based on ontology-based model of daily life. Traditionally, such information are collected by self-report, e.g., time diaries. The **second goal** of this section is to overview progress on *context annotation and analysis*, i.e., our effort of collecting contextual annotations from individuals and analyzing them. This is a prerequisite for studying the proposed context model.

Self-reporting, however, requires substantial effort from the user and fails to scale to continuous monitoring settings. The solution adopted in WeNet is to leverage ubiquitous computing, mobile sensing and machine learning technologies, to automatically recognize the user's personal context from real-time recordings collected

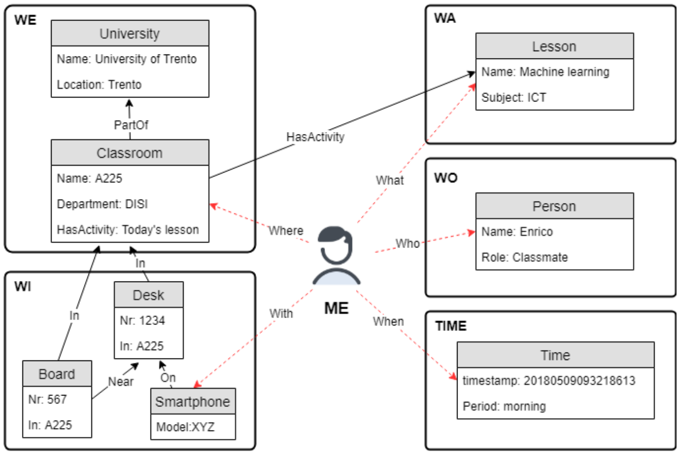


Fig. 9. Example personal context as a knowledge graph.

from smart devices (e.g., smartphones and wearable sensors). The **third goal** of this section is to provide an update on the our current effort on automated *context recognition*.

6.1 Technical approach

6.1.1 Context Model. Basic questions like: “What time is it?”, “Where are you?”, “What are you doing?”, and “Who are you with?” form the backbone of any narrative. Being able to formulate and pose these questions is fundamental for describing and collecting the daily habit and behaviors in the life of individuals. Motivated by this observation, we designed an ontology-based context model organized according to the aforementioned dimensions of the world: time, location, activity, and social relations.

We formalize the context model as a tuple:

$$Context = \langle TIME, WE, WA, WO, WI \rangle \tag{4}$$

Where:

- TIME indicates the *temporal context*, or the answer to "What **TIME** is it?". It represents the exact time in which that context was observed, e.g., "morning";
- WE is the *endurant context*, or the answer to "**W**h**E**re are you". It indicates the relevant location that a person is at, e.g., "classroom";
- WA is the *perdurant context*, or the answer to "**W**h**A**t are you doing?". It consists in the main activity taking place, e.g., the lesson;
- WO is the *social context*, or the answer to **W**h**O** are you with?". It covers all the relevant people for a person in the current context, e.g., the "teacher" and "classmates";
- WI is the *object context*, or the answer to "**W**hat are you w**I**th?", it indicates the relevant objects for a person in the current context, e.g., his or her own smartphone.

Figure 9 illustrates an example scenario. The knowledge graph represents the personal context of a Ph.D. student. In this example, the attributes of WO are “Class”, “Name”, and “Role”, and their corresponding values are “Person”, “Shen”, and “Ph.D. student”, respectively. Edges represent relations between entities, e.g., “Shen” is in relation “Attend” with “Lesson”.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Types	TIME	WE	WA
Objective Context	2020-02-17, 11am	Via Sommarive, 9, 38123 Povo TN	Shen
Machine Context	1581938718026	46°04’01.9"N 11°09’02.4"E	“Shen” is in Contact list
Subjective Context	Morning	Classroom	Friend

Table 8. Examples of three-partition context.

In this example, the context is given from an objective perspective, that is, in terms of facts that are independent from personal conscious experiences. However, individuals interpret the world based on their own personal knowledge, mental characteristics, states, *etc*. For instance, while in our example “Shen” has the objective role of a Ph.D. student, from others’ subjective perspective he has very different roles, such as ”friend” or “classmate”. In addition, the individual’s view of her own context is radically different from that of, e.g., her handheld personal assistant or any other machine. Machines observe the world via sensors, while humans interpret the world with their perceptions and also with their knowledge. For instance, location can be represented as coordinates for a machine, but humans interpret locations via functions such as home or office.

To model context precisely and completely, therefore, besides explicitly including the four dimensions of context, we also consider the perspective of the agent from which the annotations are supplied. We model context from three types of perspectives, which are objective context, subjective context and machine context. Table 8 shows the same scenario as Figure 9, but viewed from these three types of context. Consider the temporal and spatial contexts. What can be subjectively described as “morning” in the objective context, can be described as a timestamp string “1581938718026” in the machine context. Analogously, a location “Via Sommarive, 9, Trento, Italy” can be described as “classroom” subjectively by a student and sensed as coordinates “46°04’N,11°09’E” by her smartphone. As for social context, “Shen” can be described as “friend” subjectively by the user and as a mere contact list by somebody else’s email client.

6.1.2 Context Recognition. Context recognition is the task of inferring a person’s situation such as environment, physical state, social state and activity performed automatically [116]. Standard approaches to context recognition rely on supervised machine learning: a statistical model is supplied with a large set of training examples in the form of context annotations and fitted to this data. A number of machine learning approaches have been employed, both shallow models like logistic regression [116] and deep neural networks like multi-layer perceptrons [117], LSTMs [58], and CNNs [97].

A great number of studies have been done since [31] introduced baseline and recognized a person’s situation from only a wearable camera and microphone. [57] used sensors from smartphones and smartwatches for predicting basic movement more composite activities, such as drinking coffee. Work in [116] focuses on context recognition in-the-wild, that is, in real-world conditions featuring substantial variability; multi-label classification is used for recognizing combinations of context labels such as {“Running”, “Outside”, “Exercise”, “Talking”, “With friends”}. Existing work on context recognition, however, is severely limited:

(1) The issue of diversity, both objective and subjective, has never been considered. For instance, “University of Trento” is a “study place” for students but “workplace” for professors subjectively. To improve the performance of context recognition, annotation must be aligned to fit the machine learning model.
(2) The context model used is incomplete, as it targets a single modality out of four (TIME, WA, WE, WO). This is a major issue, as it completely ignores correlations between different modalities. In other words, models trained on a single modality (e.g. location) are completely blind to the others. For instance, the predictor should be aware that "Classroom" and "Housework" cannot occur at the same time.

Co-funded by the Horizon 2020
Framework Programme of the European Union

In stark contrast, our context model is designed specifically for diversity and includes four aspects (TIME, WE, WA, and WO) and has native support for subjective annotations, as discussed above.

6.2 Results and Discussion

6.2.1 Context annotation. In order to validate the proposed ontology-based context model, we mapped it to the state of the art in sociological approaches and used the mapping to collect context annotations from individuals. Time-use surveys are particularly relevant approaches, since they are widely used to investigate a specific aspect of people's time management, e.g., working, academic performance, and so on [30]. For this reason, we based our modelling of activities on several time-use surveys, especially the American Time Use Survey (ATUS) [102].

The mapping between our ontology and the sociological methodology encompasses three different lists of annotations:

Locations (WE): Figure 10 (Top) shows the mapping from the locations of the perdurant (WE) context to questions about locations. Here the mapping is almost one-to-one with the lowest tier, except for "Other University" and "Other Home", since they group more specific types of buildings. Notice that, even though "En route" is an activity, here it is treated as a location. If a student chooses "En route", instead of the options in Figure 10, a list of means of transportation is provided and the question is "How are you travelling?". The possible means of transportation are listed exactly as suggested by the sociology experts, i.e., "By Foot", "By Bus", "By Train", "By Car", "By Motorbike", and "By Bike".

Activities (WA): Figure 10 (Middle) shows the mapping of activities. Here the annotations are adapted by the first tier of activities, especially for "Relax", which maps to 4 annotations, i.e., "Hobbies", "Cultural Activity", "Other Free Time", and "Social Life". This coarseness in the mapping is due to the fact that, in order to capture high level patterns, activities are required to be very general. Furthermore, more detailed activities, as underlined by the sociology experts, would cause more cognitive load in terms of memory for students and force them to answer more questions to reach an unnecessary fine grained level of detail.

Social relations (WO): In the case of social relations, unlike locations and activities, the mapping is one to one, since they are a simple list in our current version of the WO context, as shown in Figure 10 (Bottom).

To test and apply our methodology, we interacted with sociology experts in the SmartUnitn project [53] to design a data collection and learn students' behavioral pattern and predictability. Participants were asked to install the i-Log app [126] on their smartphone to keep the app running and carry the device with them for the duration of the experiment. The i-Log app records sensor measurements from both hardware (e.g., GPS, accelerometer, gyroscope, among others) and software (e.g., running applications, nearby devices). The questionnaires were supplied every 30 minutes during the first part of the experiment and every 2 hours during the second part, and comprise three questions, one for each element of the context (WE, WA, WO): "Where are you?", "What are you doing?", and "Who are you with?".

The possible answers used the mapping above. More specifically, i) answers to "Where are you?" from the bottom tier of Figure 10 (Top), i) answers to "What are you doing?" are taken from the bottom tier in Figure 10 (Middle), and iii) answers to "Who is with you?" from the from the bottom tier of Figure 10 (Bottom). The link between the fourth question "How are you travelling?" and the "En route" activity is shown via an asterisk at the end of the latter.

6.2.2 Analysis. Next, we carried out an extensive quantitative analysis on the collected annotations. The goal was to determine whether annotations taken using our context model capture salient patterns of the individual's life, whether patterns of personal context capture the essential elements of diversity and thus act as fingerprints of individuals [113], and whether our context model is capable of supporting advanced context recognition.

Figure 11 illustrates the daily contexts of the two participants with high (top) and low (bottom) diversity, measured using information theory tools (namely, entropy). The data clearly shows that the behavioral patterns

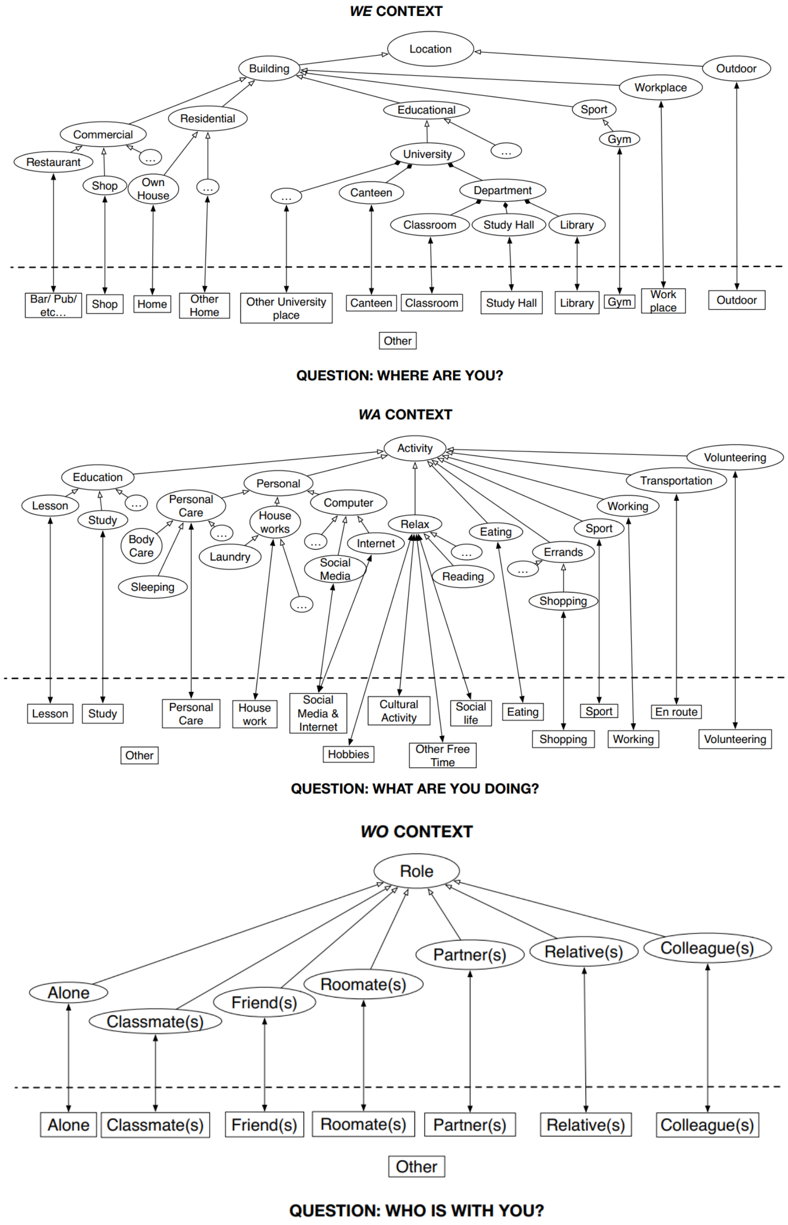


Fig. 10. Mapping between context ontology and social concepts. The concepts used as possible answers to the thee questions in the SmartUnitn questionnaires. From top to bottom: location (WE), activity (WA), and social context (WO).

differ substantially between the two users (cf. [41]). This indicates that even in a rather homogeneous user group (university students from the same city and age group) and at a rather coarse level of detail, there us a substantial amount of diversity across individuals.

Co-funded by the Horizon 2020
Framework Programme of the European Union

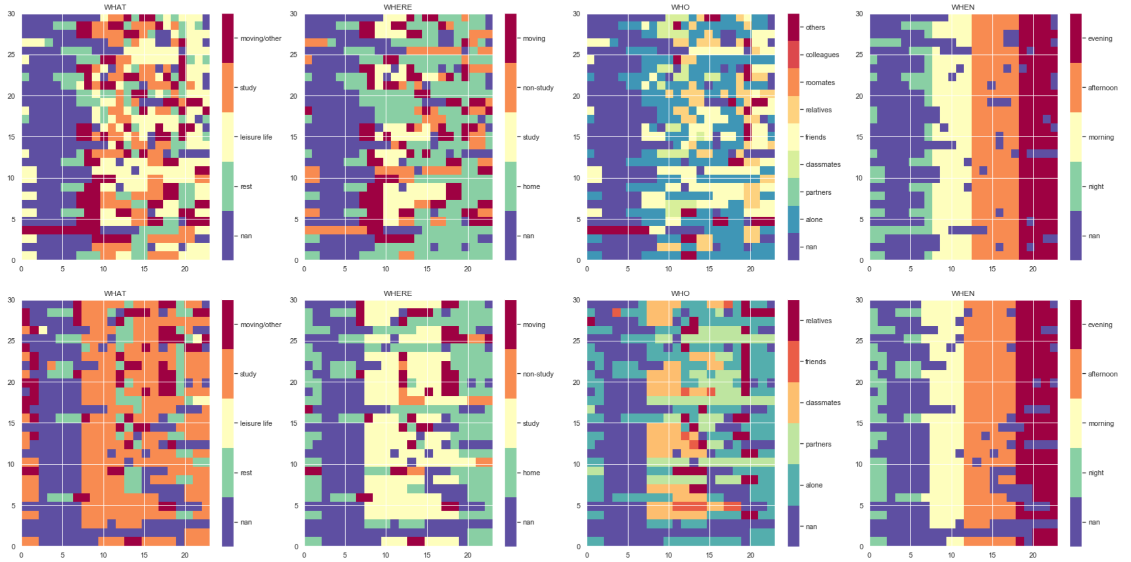


Fig. 11. Annotations of the least (top) and most (bottom) predictable users. Each row is a day, columns are hours. Colors indicate (simplified) annotations for, from left to right, WA, WE, WO, and TIME, respectively. Missing values are in blue.

Following previous work on human mobility and behavior [56, 92, 103, 104], our analysis makes use of techniques from information theory [34]. This enables us to examine the potential predictability of personal context patterns in a model-agnostic fashion. We measured the randomness of the value distribution of each aspect of the context model (TIME, WA, WE, WO) in terms of *entropy*. The higher the entropy, the less predictable the behavior. The entropy distribution averaged over users is reported in Figure 12. The three histograms in each plot refer to three progressively more informed forms of entropy, introduced in [104]:

- **Orange**: This form of entropy assumes that all values (e.g. locations) annotated by an individual appear equally likely. In other words, the red curve ignores the distribution of observed values and is useful as a baseline. Past annotations (e.g., taken in the previous questionnaire) are not taken into consideration.
- **Blue**: This other form of entropy takes into consideration the actual frequency of observed values. No time information is used.
- **Green**: This more refined form of entropy takes into consideration both the actual frequency of observed values and previous annotations.

The plots highlight the following phenomena:

- a) All modalities are to some extent predictable: For instance, observing the intra-modal state of a subject reveals b bits about what the next intra-modal state will be, where on average $b > 1$ in every modality [104].
- b) Some modalities are *intrinsically* more uncertain than others. This is largely governed by the number of unique values, i.e., more alternatives imply lower predictability, as expected.
- c) Finer-grained entropy values reveal a higher degree of predictability: uniform entropy (blue) is often higher than non-sequential entropy (orange), which is itself much higher than the sequential entropy (green). This means that a lot structure is provided by the sequence of events. The spread depends on the aspect.

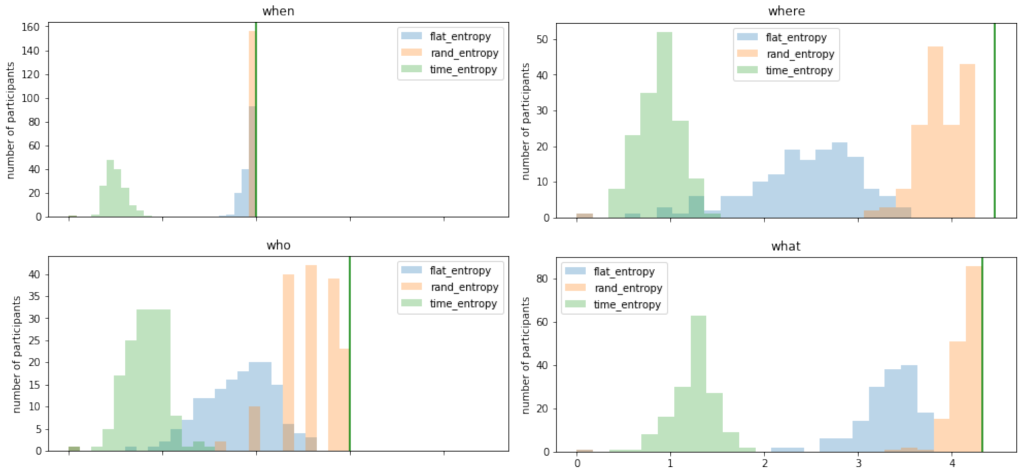


Fig. 12. Histogram of (three types of) entropy. Counts are averaged over users.

These results extend the observations of [104] from mobility alone to all aspects of the personal context. Next, we studied whether correlations existing across different aspects of our context model are beneficial for predictability. In order to do so, for each modality (TIME, WA, WE, WO) we computed the difference in entropy before and after disclosing information about one or more of the remaining modalities. This quantity, called information gain, quantifies how much information a modality reveals about another modality, e.g., location about activity. Figure 13 shows the average (across users) information gain of each modality conditioned on all subsets of other modalities. The results show immediately that *in all cases, knowledge about one modality provides information about the others.* This is a strong message, and provides ample support for using multi-modal context models like ours, rather than existing uni-modal models. Given that entropy and information gain represent upper bounds on predictive performance (the link is due to Fano inequality, cf. [104]), we expect these results to carry over to machine-learning based context recognition.

Our work so far has focused on designing a formal ontology-driven context model appropriate for understanding human behavior and diversity, on mapping the context model to the state of the art sociology, on validating the context model in a data collection experiments, and on analyzing the behavior of participants. The results highlighted the ability of our context model to capture relevant information, including correlations between time, location, activity, and social context, which are fundamental for both analysis and prediction. Indeed, taking multiple aspects into consideration dramatically improves the predictability of personal contexts. Work on combining our context model with machine learning for carrying out context recognition is underway.

7 DESIGN, IMPLEMENTATION, AND INITIAL ANALYSIS OF THE MEXICO PRE-PILOT

We tested the full cycle of the WeNet data collection protocol (experimental design; GDPR compliance and institutional ethical approval; i-Log mobile sensing application adaptation; student recruitment; field data collection and feedback; and data analysis) through two phases of pre-pilot data collection in two universities in Mexico. This was a joint effort involving several WeNet partners (IPICYT, IDIAP, UNITN, and UHOPPER).

There were three main considerations regarding this pre-pilot; (1) Finding a topic that significantly matters to people in Mexico, so that they naturally tend to use the mobile application. (2) Engaging a youth population

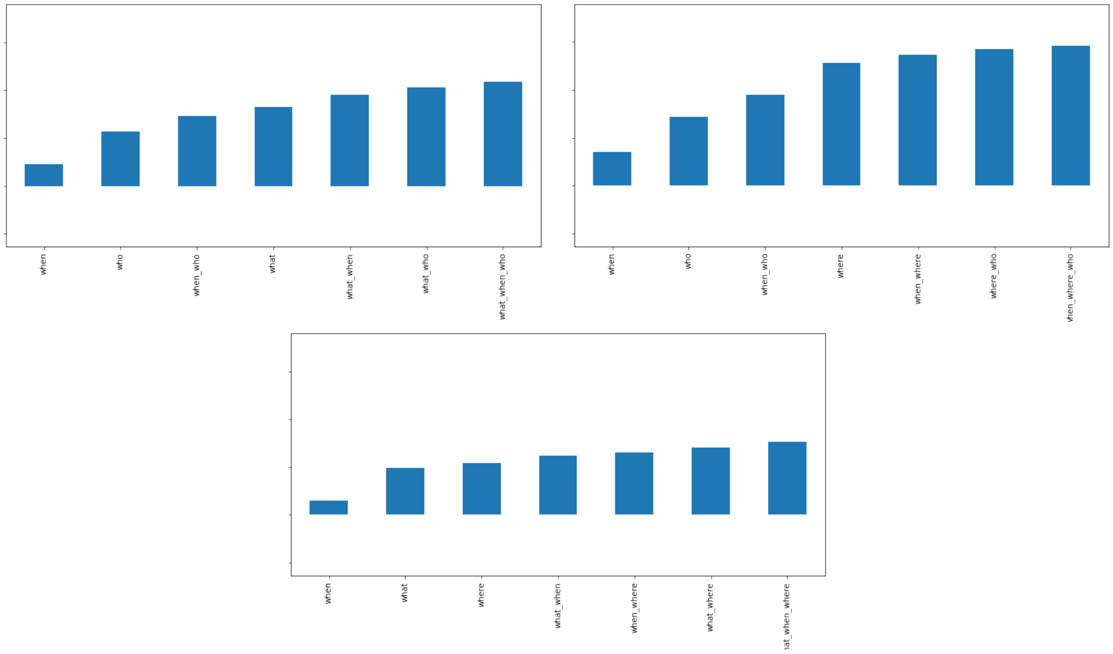
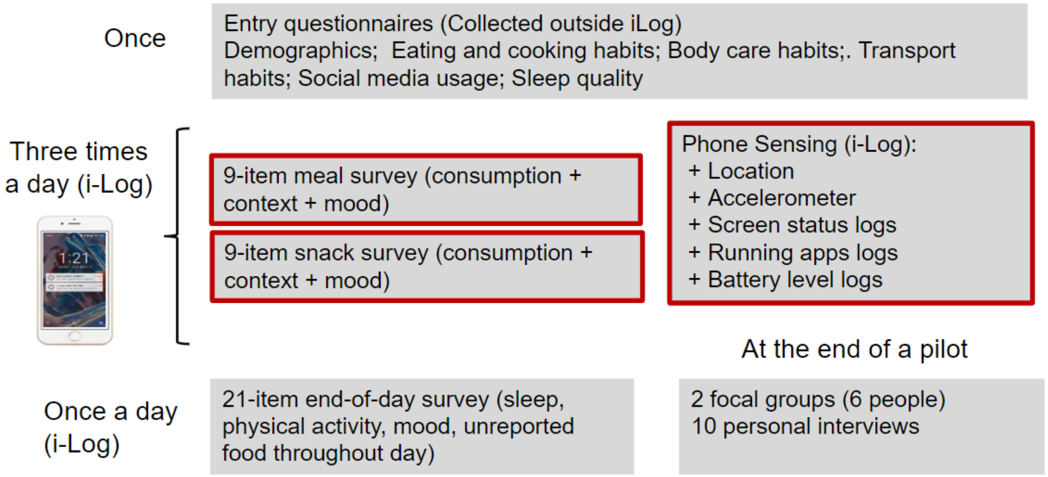


Fig. 13. Information gain: decrease in entropy in a modality due to conditioning on a subset of the other modalities. Top to bottom, left to right: information gain for WE, WA, and WO, respectively.


Once

Entry questionnaires (Collected outside iLog)
Demographics; Eating and cooking habits; Body care habits;. Transport habits; Social media usage; Sleep quality

Three times a day (i-Log)

9-item meal survey (consumption + context + mood)

9-item snack survey (consumption + context + mood)

Phone Sensing (i-Log):
+ Location
+ Accelerometer
+ Screen status logs
+ Running apps logs
+ Battery level logs

At the end of a pilot

Once a day (i-Log)

21-item end-of-day survey (sleep, physical activity, mood, unreported food throughout day)

2 focal groups (6 people)
10 personal interviews

Fig. 14. Block Diagram of Data Collection in Mexico Pre-Pilot

who are willing to participate in this campaign, and (3) Engaging people into a local initiative that can gradually grow at a regional or even national level. Considering these circumstances, overweight and obesity was chosen as a topic of interest for the mobile sensing application.

Regarding the topic of interest, overweight and obesity, both related to food consumption, are pressing challenges specially for young adults. In Mexico, obesity is considered a huge societal issue. According to statistics, seven out of ten adults, and one out of three children are overweight or obese. Obesity can lead to conditions such as hypertension, low self-esteem, depression, diabetes, and heart problems, and 70 million people have a sickness associated with obesity. For the pilot, the area of San Luis Potosi, which is home for 1.2 million people, was chosen. Figure 15 shows some pictures captured during awareness sessions and focus groups. In this context, we chose food consumption and obesity as the primary topic of interest.

Regarding the population of participants, we chose university students for several reasons. People of this age are more curious and tech savvy. Hence, deploying a mobile sensing application among students of this age group is a natural choice. In addition, it is easier to motive young adults and they are open to change their behaviour compared to adults. Hence, considering all these factors, the overall goal of WeNet in Mexico was to use new mobile technologies to study this pervasive problem affecting young people, specifically in relation to their eating and physical activity habits.

Considering all the above factors, a WeNet Mexico campaign was launched on 28^{th} June 2019 by IPICYT's general director in Mexico. A launching event presented WeNet to several local and national media outlets. Moreover, in order to engage people with the pilot, we created awareness campaigns, social media marketing, and media briefings that were reported through national media outlets of mexico [1]. One such workshop was held on August 24, 2019 with four main activities: (1) WeNet.MX was introduced to participants including the goals of the study (i.e., to sense eating behaviour of university students using a mobile platform); (2) participants were informed about how the pilot study would be carried out, and how their data will be used for research; (3) participants voluntarily filled out a consent form; and (4) participants who filled the consent form and entered the pilot, filled an entry questionnaire (see Figure 14).

7.1 Technical Approach and Results

We collected data from users with two main techniques: passive sensing and self-reporting. Passive sensing involved collecting data including location, accelerometer, screen status log, running apps log, and battery level logs. Self-reporting involved users answering questionnaires regarding food consumption (meals or snacks, their mood), end of the day survey (sleep, physical activity, mood, unreported food throughout) on a daily basis. Moreover, they provided basic demographic information at the beginning of the pilot, and took part in interviews and focus groups at the end of the pilot. The types of data that were collected are summarized in Figure 14. The pilot was carried out in two phases as described in Table 10, and the number of surveys sent to users and answered are mentioned in Table 11. Table 9 shows a set of features that can be derived from the collected data.

Figure 16 shows locations at which people have responded to questionnaires. In the morning and afternoon, the majority of responses have come from home while in the evening (5pm to 7pm) most responses responses are scattered compared to the other two time periods. Based on the reported locations, stay regions were calculated for all users. As seen in Figure 17, the majority of users have only 0-5 unique stay points while higher number of unique stay points have been reported only by fewer number of students.

Moreover, we calculated linear acceleration of all users during the time period of the experiment, for each 1 hour window. This value acts as a proxy regarding the activity level of users. As it is seen from Figure 18, during the time period between 1 am to 6 am acceleration values are low, probably because students are sleeping.

A full analysis of the data set will be reported in the next WP2 deliverable.

[1] A video from an awareness campaign: https://www.facebook.com/watch/?v=513199302861930

Co-funded by the Horizon 2020 Framework Programme of the European Union

Feature	Description	Type	Feature Group
gender	Whether the user is male or female	categorical(2)	D
age	Age of the user	numerical	D
bmi	Body Mass Index of the user	numerical	D
time_since_last_meal	Time in minutes, since the last meal	numerical	C
time_in_min	Time of the day at which the eating event took place	numerical	C
meal_snack	Whether it is a meal or a snack	categorical(2)	C
where	Semantic meaning of the location at which meal/snack event took place	categorical(10)	C
withwhom	Social context of a meal/snack (alone, with a group, etc)	categorical(4)	C
whatelse	What were the user doing when having the meal/snack	categorical(17)	C
app_usage	Type of apps used by people during different times of the day	categorical(5)	C
battery_event	Battery related charging events in the smartphone	categorical(2)	C
screen_event	Screen related on/off events in the smartphone	categorical(2)	C
battery_level	Battery level in the smartphone	numerical	A
activity_levels	Activity levels derived from accelerometer traces	numerical	A

Table 9. Summary of the features used for the analysis. Three types of feature groups are Demographic Information (D), Contextual Information (C), and Activity Data (A). Type describes whether the feature is categorical or numerical, and if it is categorical, how many categories are represented by the feature.

Phase	# of People in Recruiting Workshop	# of Recruited Volunteers	Population
I	32	29 (90.6%)	College Students UASLP
II	90	55 (61.1%)	College Students UTAN (Nutrition)

Table 10. Details of Two Phases

Pilot	Participants	Sent Surveys	Answered Surveys
I	29	6466	3436
II	55	2177	1718

Table 11. Summary of Collected Data

7.2 Discussion

There were several key lessons from the pilot. One key aspect regards the importance of considering diversity when creating machine learning models for diverse populations. As understood during the pilot, the Mexican student population is highly diverse in terms of their behaviour and societal background, hence leading to different ways of living. This requires diversity-aware mobile sensing techniques. This kind of technology can also support the development of citizen-based solutions to challenging health and social challenges faced by people in México in their daily lives. Second, it was also emphasized that the i-Log mobile sensing application might need changes to suit local contexts. This lesson would be helpful in deploying the application during the future pre-pilots scheduled in Denmark, UK, and Italy.

During the interviews, the pilot received many positive responses. Below, we illustrate some of the suggestions and feedback. Many students appreciated the feedback they could potentially got from a mobile application to improve their behaviour leading to healthier lifestyles.

Fig. 15. Participants in Mexico Pre-pilot

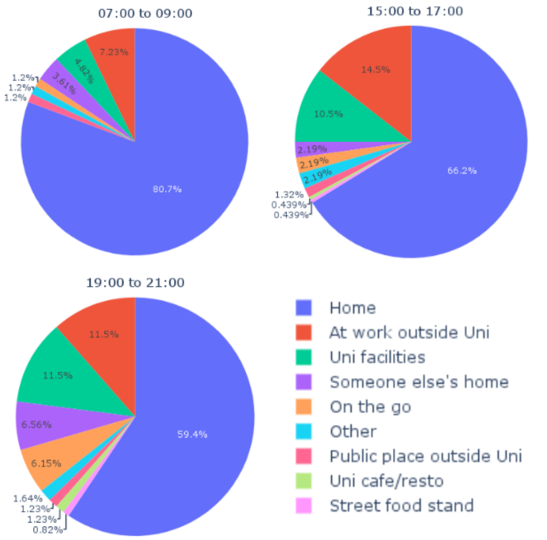


Fig. 16. Time of the day and self-report Locations

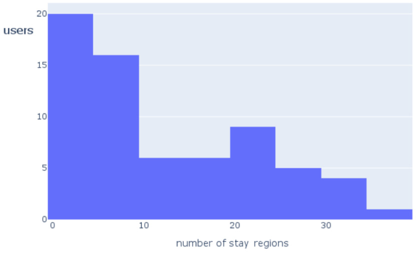


Fig. 17. Distribution of number of stay regions per user over the campaign

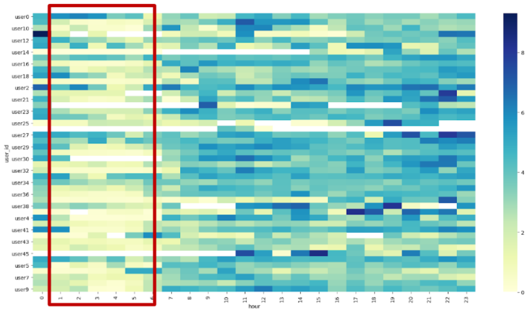


Fig. 18. 2D acceleration mean for different users for one-hour windows

> “The study encouraged me to think about the way I eat. I think I do not exercise enough, and spent too much time sitting and during my free time I watch TV.”
> – A 22 years old female student

> “The pilot made me reflect on the number of meals I have during a day. Sometimes because I’m so busy with school, I eat twice a day and I should at least have three meals a day. I felt bad when I had to answer the morning questionnaire, yet I did not have breakfast!.”
> – A 24 years old female student

> “They could provide feedback every week on how it went. For example, the app could report information such as : "Well, this week in general these were the results: You consumed so much carbohydrates, lots of protein” and so on. Or ”You didn’t exercise much this week”... It could also give feedback on know how you compare yourself with the rest. That is, or seeing what pattern you are following. ”
> – A 23 years old female student

> “I would like that the app to provide feedback on how many calories you are consuming, maybe not exactly but an approximation, so we have an idea on what we are eating by the end of the week, you make a count at the end of the week."
> – A 22 years old male student

Moreover, the Mexico pre-pilot is the first of several pre-pilots that WeNet plans to deploy in Y2. We believe that doing pilots in different countries with diverse user groups would allow us to identify fine-grained behavioral patterns of people using diversity aware machine learning techniques.

8 CONCLUSION

This deliverable describes the initial individual learning methods developed in WeNet. The work included a systematic review of the domain; work on algorithms for routine learning; work on algorithms for learning while preserving sensitive attributes; work on skeptical learning to deal with real-time mobile reports; work on user context recognition using ontology models; and design, implementation, data analysis of the Mexico pre-pilot, which represents the first WeNet dataset collected in the project.

The work in WP2 has progressed at a good pace thanks to the collaboration of the partners. One key objective for the next period will be the application of these methodologies to the datasets to be collected in Year 2 of the project by the next pre-pilots, both in Europe and outside Europe. The second objective is the development of additional methods to improve the algorithmic capabilities of the WP2 technologies, and to adapt and integrate the models to use them in the specific scenario for the first full pilot.

REFERENCES

[1] 1999. *Proposition. No. 92 (1998-99) About the Act on the Processing of Personal Data (the Personal Data Act)*. Retrieved Nov 14, 2020 from https://www.regjeringen.no/no/dokumenter/otprp-nr-92-1998-99-/id160088/

[2] 2016. *APPFAIL: Threats to Consumers in Mobile Apps*. Retrieved Nov 14, 2020 from https://fil.forbrukerradet.no/wp-content/uploads/2016/03/Appfail-Report-2016.pdf

[3] 2019. *FitBit Privacy Policy*. Retrieved Nov 11, 2019 from https://www.fitbit.com/eu/legal/privacy-policy

[4] 2019. *Mobile Fact Sheet*. Retrieved July 24, 2019 from https://www.pewinternet.org/fact-sheet/mobile/

[5] 2019. *MyFitnessPal*. Retrieved Nov 06, 2019 from https://www.myfitnesspal.com/

[6] 2019. *Smartphone Addiction Facts Phone Usage Statistics*. Retrieved July 25, 2019 from https://www.bankmycell.com/blog/smartphone-addiction/

[7] 2020. *Google Fit - Coaching you to a healthier and more active life*. Retrieved February 12, 2020 from https://www.google.com/fit/

[8] 2020. *A more personal Health app. For a more informed you*. Retrieved February 12, 2020 from https://www.apple.com/ios/health/

[9] 2020. *S Health Terms of Use*. Retrieved Feb 13, 2020 from https://account.samsung.com/membership/etc/specialTC.do?fileName=shealth.html

[10] 2020. *Samsung Health App*. Retrieved February 12, 2020 from https://www.samsung.com/us/support/owners/app/samsung-health

[11] Saeed Abdullah, Elizabeth L. Murnane, Mark Matthews, Matthew Kay, Julie A. Kientz, Geri Gay, and Tanzeem Choudhury. 2016. Cognitive Rhythms: Unobtrusive and Continuous Sensing of Alertness Using a Mobile Phone. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. ACM, New York, NY, USA, 178–189. https://doi.org/10.1145/2971648.2971712

[12] Ionut Andone, Konrad Błaszkiewicz, Mark Eibes, Boris Trendafilov, Christian Montag, and Alexander Markowetz. 2016. How Age and Gender Affect Smartphone Usage. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 9–12. https://doi.org/10.1145/2968219.2971451

[13] Sangwon Bae, Denzil Ferreira, Brian Suffoletto, Juan C. Puyana, Ryan Kurtz, Tammy Chung, and Anind K. Dey. 2017. Detecting Drinking Episodes in Young Adults Using Smartphone-based Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2, Article 5 (June 2017), 36 pages. https://doi.org/10.1145/3090051

[14] B. Baron and M. Musolesi. 2020. Interpretable Machine Learning for Privacy-Preserving Pervasive Systems. *IEEE Pervasive Computing* (2020), 1–10. https://doi.org/10.1109/MPRV.2019.2918540

[15] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Yan Beh, Mayank Goel, Thad Starner, and Gregory Abowd. 2017. EarBit: Using Wearable Sensors to Detect Eating Episodes in Unconstrained Environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 37 (Sept. 2017), 20 pages. https://doi.org/10.1145/3130902

[16] Lisa K. Berger, Audrey L. Begun, and Laura L. Otto-Salaj. 2009. Participant recruitment in intervention research: scientific integrity and cost-effective strategies. *International Journal of Social Research Methodology* 12, 1 (2009), 79–92. https://doi.org/10.1080/13645570701606077 arXiv:https://doi.org/10.1080/13645570701606077

[17] Jennifer Berry. 2019. *Is dairy good or bad for your health?* Retrieved Jan 22, 2020 from https://www.medicalnewstoday.com/articles/326269.php

[18] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. *CoRR* abs/1707.00075 (2017). arXiv:1707.00075 http://arxiv.org/abs/1707.00075

[19] Joan-Isaac Biel, Nathalie Martin, David Labbe, and Daniel Gatica-Perez. 2018. Bites&Lsquo;N&Rsquo;Bits: Inferring Eating Behavior from Contextual Mobile Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 125 (Jan. 2018), 33 pages. https://doi.org/10.1145/3161161

[20] Carole A. Bisogni, Laura Winter Falk, Elizabeth Madore, Christine E. Blake, Margaret Jastran, Jeffery Sobal, and Carol M. Devine. 2007. Dimensions of everyday eating and drinking episodes. *Appetite* 48, 2 (2007), 218 – 231. https://doi.org/10.1016/j.appet.2006.09.004

[21] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex (Sandy) Pentland. 2014. Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits. In *Proceedings of the 22Nd ACM International Conference on Multimedia* (Orlando, Florida, USA) *(MM '14)*. ACM, New York, NY, USA, 477–486. https://doi.org/10.1145/2647868.2654933

[22] A. Bogomolov, B. Lepri, and F. Pianesi. 2013. Happiness Recognition from Mobile Phone Data. In *2013 International Conference on Social Computing*. 790–795. https://doi.org/10.1109/SocialCom.2013.118

[23] Mehdi Boukhechba, Alexander R. Daros, Karl Fua, Philip I. Chow, Bethany A. Teachman, and Laura E. Barnes. 2018. DemonicSalmon: Monitoring mental health and social interactions of college students using smartphones. *Smart Health* 9-10 (2018), 192 – 203. https://doi.org/10.1016/j.smhl.2018.07.005 CHASE 2018 Special Issue.

[24] Russel Brandom. 2018. Self-driving cars are headed toward an AI roadblock. https://www.theverge.com/2018/7/3/17530232/self-driving-ai-winter-full-autonomy-waymo-tesla-uber

[25] George A. Bray and Barry M. Popkin. 2014. Dietary Sugar and Body Weight: Have We Reached a Crisis in the Epidemic of Obesity and Diabetes? *Diabetes Care* 37, 4 (2014), 950–956. https://doi.org/10.2337/dc13-2085 arXiv:https://care.diabetesjournals.org/content/37/4/950.full.pdf

[26] Sanja Brdar, Dubravko Ćulibrk, Vladimir Crnojević, and Trg Dositeja Obradovića. [n.d.]. Demographic Attributes Prediction on the Real-World Mobile Data.

[27] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (01 Oct 2001), 5–32. https://doi.org/10.1023/A:1010933404324

[28] Luca Canzian and Mirco Musolesi. 2015. Trajectories of Depression: Unobtrusive Monitoring of Depressive States by Means of Smartphone Mobility Traces Analysis. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. ACM, New York, NY, USA, 1293–1304. https://doi.org/10.1145/2750858.2805845

[29] Ramnath K. Chellappa and Raymond G. Sin. 2002. Personalization versus Privacy: An Empirical Examination of the Online Consumer's Dilemma. *Information Technology and Management* 6 (2002), 181–202.

[30] Brigitte JC Claessens, Wendelien Van Eerde, Christel G Rutte, and Robert A Roe. 2007. A review of the time management literature. *Personnel review* (2007).

[31] Brian Clarkson, Kenji Mase, and Alex Pentland. 2000. Recognizing user context via wearable sensors. In *Digest of Papers. Fourth International Symposium on Wearable Computers*. IEEE, 69–75.

[32] Lisa Cohen, Gary C. Curhan, and John P. Forman. 2012. Influence of Age on the Association between Lifestyle Factors and Risk of Hypertension. *J Am Soc Hypertens* (2012). https://doi.org/10.1016/j.jash.2012.06.002

[33] Sunny Consolvo, David W. McDonald, Tammy Toscos, Mike Y. Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, Ian Smith, and James A. Landay. 2008. Activity Sensing in the Wild: A Field Trial of Ubifit Garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. ACM, New York, NY, USA, 1797–1806. https://doi.org/10.1145/1357054.1357335

[34] Thomas M Cover and Joy A Thomas. 2012. *Elements of information theory*. John Wiley & Sons.

[35] Tegan Cruwys, Kirsten E. Bevelander, and Roel C.J. Hermans. 2015. Social modeling of eating: A review of when and why social influence affects food intake and choice. *Appetite* 86 (2015), 3 – 18. https://doi.org/10.1016/j.appet.2014.08.035 Social Influences on Eating.

[36] Anthony Cuthbertson. 2019. Self-driving cars more likely to drive into black people, study claims. https://www.independent.co.uk/life-style/gadgets-and-tech/news/self-driving-car-crash-racial-bias-black-people-study-a8810031.html

[37] John M. de Castro, George A. King, Maria Duarte-Gardea, Salvador Gonzalez-Ayala, and Charles H. Kooshian. 2012. Overweight and obese humans overeat away from home. *Appetite* 59, 2 (2012), 204 – 211. https://doi.org/10.1016/j.appet.2012.04.020 The 36th annual meeting of the British Feeding and Drinking Group, March 29th and 30th 2012, Brighton, UK.

[38] Marco De Nadai, Angelo Cardoso, Antonio Lima, Bruno Lepri, and Nuria Oliver. 2019. Strategies and limitations in app usage and human mobility. *Scientific Reports* 9 (07 2019), 10935. https://doi.org/10.1038/s41598-019-47493-x

[39] Anind K Dey. 2001. Understanding and using context. *Personal and ubiquitous computing* 5, 1 (2001), 4–7.

[40] Thomas G. Dietterich. 2000. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning* 40, 2 (01 Aug 2000), 139–157. https://doi.org/10.1023/A:1007607513941

[41] Nathan Eagle and Alex Sandy Pentland. 2009. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology* 63, 7 (2009), 1057–1066.

[42] Nathan Eagle and Alex (Sandy) Pentland. 2006. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing* 10, 4 (01 May 2006), 255–268. https://doi.org/10.1007/s00779-005-0046-3

[43] Yanai Elazar and Yoav Goldberg. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 11–21. https://doi.org/10.18653/v1/D18-1002

[44] Jon D. Elhai, Jason C. Levine, Robert D. Dvorak, and Brian J. Hall. 2017. Non-social features of smartphone use are most related to depression, anxiety and problematic smartphone use. *Computers in Human Behavior* 69 (2017), 75 – 82. https://doi.org/10.1016/j.chb.2016.12.023

[45] Alex Elliott-Green, Lirije Hyseni, Ffion Lloyd-Williams, Helen Bromley, and Simon Capewell. 2016. Sugar-sweetened beverages coverage in the British media: an analysis of public health advocacy versus pro-industry messaging. *BMJ Open* 6, 7 (2016). https://doi.org/10.1136/bmjopen-2016-011295 arXiv:https://bmjopen.bmj.com/content/6/7/e011295.full.pdf

[46] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Kdd*, Vol. 96. 226–231.

[47] Farhan, Yue, Morillo, Ware, Lu, Bi, Kamath, Russell, Bamis, and Wang. 2016. Behavior vs. Introspection: refining prediction of clinical depression via smartphone sensing data. In *2016 IEEE Wireless Health (WH)*. 1–8. https://doi.org/10.1109/WH.2016.7764553

[48] J. Farringdon, A. J. Moore, N. Tilbury, J. Church, and P. D. Biemond. 1999. Wearable sensor badge and sensor jacket for context awareness. In *Digest of Papers. Third International Symposium on Wearable Computers*. 107–113. https://doi.org/10.1109/ISWC.1999.806681

[49] Jon Froehlich, Mike Y. Chen, Sunny Consolvo, Beverly Harrison, and James A. Landay. 2007. MyExperience: A System for in Situ Tracing and Capturing of User Feedback on Mobile Phones. In *Proceedings of the 5th International Conference on Mobile Systems, Applications and Services* (San Juan, Puerto Rico) *(MobiSys '07)*. ACM, New York, NY, USA, 57–70. https://doi.org/10.1145/1247660.1247670

[50] Daniel Gatica-Perez, Joan-Isaac Biel, David Labbe, and Nathalie Martin. 2019. Discovering eating routines in context with a smartphone app. In *UbiComp/ISWC Adjunct*.

[51] Aritra Ghosh, Naresh Manwani, and P. S. Sastry. 2017. On the Robustness of Decision Tree Learning Under Label Noise. In *Advances in Knowledge Discovery and Data Mining*, Jinho Kim, Kyuseok Shim, Longbing Cao, Jae-Gil Lee, Xuemin Lin, and Yang-Sae Moon (Eds.). Springer International Publishing, Cham, 685–697.

[52] Henner Gimpel, Christian Regal, and Marco Schmidt. 2015. myStress: Unobtrusive Smartphone-Based Stress Detection. In *ECIS*.

[53] Fausto Giunchiglia, Mattia Zeni, Elisa Gobbi, Enrico Bignotti, and Ivano Bison. 2018. Mobile social media usage and academic performance. *Computers in Human Behavior* 82 (2018), 177–185.

[54] Jiaqi Gong, Yu Huang, Philip I. Chow, Karl Fua, Matthew S. Gerber, Bethany A. Teachman, and Laura E. Barnes. 2019. Understanding behavioral dynamics of social anxiety among college students through smartphone sensors. *Information Fusion* 49 (2019), 57 – 68. https://doi.org/10.1016/j.inffus.2018.09.002

[55] Zhiqiang Gong, Ping Zhong, and Weidong Hu. 2019. Diversity in Machine Learning. *IEEE Access* 7 (2019), 64323–64350. https://doi.org/10.1109/access.2019.2917620

[56] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *nature* 453, 7196 (2008), 779.

[57] John J Guiry, Pepijn Van de Ven, and John Nelson. 2014. Multi-sensor fusion for enhanced contextual awareness of everyday activities with ubiquitous devices. *Sensors* 14, 3 (2014), 5687–5701.

[58] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880* (2016).

[59] HealthLinkBC. 2019. Alcohol and Drug Use in Young People. https://www.healthlinkbc.ca/health-topics/tp17749

[60] HealthLinkBC. 2019. Statistics of Alcohol and Drug Use in Young People. https://www.healthlinkbc.ca/health-topics/tp17749#uq2408

[61] Kelly Houston, Keith Hawton, and Rosie Shepperd. 2001. Suicide in young people aged 15–24: a psychological autopsy study. *Journal of Affective Disorders* 63, 1 (2001), 159 – 170. https://doi.org/10.1016/S0165-0327(00)00175-0

[62] Yu Huang, Haoyi Xiong, Kevin Leach, Yuyan Zhang, Philip Chow, Karl Fua, Bethany A. Teachman, and Laura E. Barnes. 2016. Assessing Social Anxiety Using Gps Trajectories and Point-of-interest Data. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. ACM, New York, NY, USA, 898–903. https://doi.org/10.1145/2971648.2971761

[63] A. Jain and V. Kanhangad. 2016. Investigating gender recognition in smartphones using accelerometer and gyroscope sensor readings. In *2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT)*. 597–602. https://doi.org/10.1109/ICCTICT.2016.7514649

[64] Fanyu Kong and Jindong Tan. 2012. DietCam: Automatic dietary assessment with mobile camera phones. *Pervasive and Mobile Computing* 8, 1 (2012), 147 – 163. https://doi.org/10.1016/j.pmcj.2011.07.003

[65] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. 2010. A survey of mobile phone sensing. *IEEE Communications Magazine* 48, 9 (Sep. 2010), 140–150. https://doi.org/10.1109/MCOM.2010.5560598

[66] N. D. Lane, M. Mohammod, M. Lin, X. Yang, H. Lu, S. Ali, A. Doryab, E. Berke, T. Choudhury, and A. Campbell. 2011. Bewell: A smartphone application to monitor, model and promote wellbeing.. In *Proceedings of the Conference PervasiveHealth (PervasiveHealth '11)*. http://pac.cs.cornell.edu/pubs/PervasiveHealth_BeWell.pdf

[67] Kurt Lewin. 1936. In *Principles of topological psychology.* https://doi.org/10.1037/10019-000

[68] Robert LiKamWa, Yunxin Liu, Nicholas D. Lane, and Lin Zhong. 2013. MoodScope: Building a Mood Sensor from Smartphone Usage Patterns. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services* (Taipei, Taiwan) *(MobiSys '13)*. ACM, New York, NY, USA, 389–402. https://doi.org/10.1145/2462456.2464449

[69] Mu Lin, Nicholas D. Lane, Mashfiqui Mohammod, Xiaochao Yang, Hong Lu, Giuseppe Cardone, Shahid Ali, Afsaneh Doryab, Ethan Berke, Andrew T. Campbell, and Tanzeem Choudhury. 2012. BeWell+: Multi-dimensional Wellbeing Monitoring with Community-guided User Feedback and Energy Optimization. In *Proceedings of the Conference on Wireless Health* (San Diego, California) *(WH '12)*. ACM, New York, NY, USA, Article 10, 8 pages. https://doi.org/10.1145/2448096.2448106

[70] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T. Chittaranjan, Andrew T. Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. StressSense: Detecting Stress in Unconstrained Acoustic Environments Using Smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (Pittsburgh, Pennsylvania) *(UbiComp '12)*. ACM, New York, NY, USA, 351–360. https://doi.org/10.1145/2370216.2370270

[71] Y. Ma, B. Xu, Y. Bai, G. Sun, and R. Zhu. 2012. Daily Mood Assessment Based on Mobile Phone Sensing. In *2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks*. 142–147. https://doi.org/10.1109/BSN.2012.3

[72] Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. 2018. Protecting Sensory Data Against Sensitive Inferences. In *Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems* (Porto, Portugal) *(W-P2DS'18)*. ACM, New York, NY, USA, Article 2, 6 pages. https://doi.org/10.1145/3195258.3195260

[73] Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. 2019. Mobile Sensor Data Anonymization. In *Proceedings of the International Conference on Internet of Things Design and Implementation* (Montreal, Quebec, Canada) *(IoTDI '19)*. ACM, New York, NY, USA, 49–58. https://doi.org/10.1145/3302505.3310068

[74] Mohammad Malekzadeh, Richard G. Clegg, and Hamed Haddadi. 2017. Replacement AutoEncoder: A Privacy-Preserving Algorithm for Sensory Data Analysis. *CoRR* abs/1710.06564 (2017). arXiv:1710.06564 http://arxiv.org/abs/1710.06564

[75] Roberta Masella and Walter Malorni. 2017. Gender-related differences in dietary habits. *Clinical Management Issues* 11, 2 (2017). https://doi.org/10.7175/cmi.v11i2.1313

[76] Akhil Mathur, Lakshmi Manasa Kalanadhabhatta, Rahul Majethia, and Fahim Kawsar. 2017. Moving Beyond Market Research: Demystifying Smartphone User Behavior in India. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 82 (Sept. 2017), 27 pages. https://doi.org/10.1145/3130947

[77] Jim McCambridge, John Witton, and Diana R. Elbourne. 2014. Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology* 67, 3 (2014), 267 – 277. https://doi.org/10.1016/j.jclinepi.2013.08.015

[78] Bent Egberg Mikkelsen. 2011. Guest Editorial. *Perspectives in Public Health* 131, 5 (2011), 206–206. https://doi.org/10.1177/1757913911419151 arXiv:https://doi.org/10.1177/1757913911419151 PMID: 21999023.

[79] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K. D'Mello, Munmun De Choudhury, Gregory D. Abowd, and Thomas Plötz. 2019. Prediction of Mood Instability with Passive Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 75 (Sept. 2019), 21 pages. https://doi.org/10.1145/3351233

[80] Sai T Moturu, Inas Khayal, Nadav Aharony, Wei Pan, and Alex (Sandy) Pentland. 2011. Using Social Sensing to Understand the Links Between Sleep, Mood, and Sociability. In *Proceedings of IEEE International Conference on Social Computing.*

[81] Elizabeth L. Murnane, Saeed Abdullah, Mark Matthews, Matthew Kay, Julie A. Kientz, Tanzeem Choudhury, Geri Gay, and Dan Cosley. 2016. Mobile Manifestations of Alertness: Connecting Biological Rhythms with Patterns of Smartphone App Use. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Florence, Italy) *(MobileHCI '16)*. ACM, New York, NY, USA, 465–477. https://doi.org/10.1145/2935334.2935383

[82] David F. Nettleton, Albert Orriols-Puig, and Albert Fornells. 2010. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review* 33, 4 (01 Apr 2010), 275–306. https://doi.org/10.1007/s10462-010-9156-z

[83] Michael D Newcomb and Peter M. Bentler. 1988. Impact of adolescent drug use and social support on problems of young adults: A longitudinal study. *Journal of Abnormal Psychology* (1988). https://doi.org/10.1037/0021-843X.97.1.64

[84] Toan Nguyen, Aditi Roy, and Nasir D. Memon. 2018. Kid on The Phone! Toward Automatic Detection of Children on Mobile Devices. *CoRR* abs/1808.01680 (2018). arXiv:1808.01680 http://arxiv.org/abs/1808.01680

[85] S Park, C Gopalsamy, R Rajamanickam, and S Jayaraman. 1999. The Wearable Motherboard: a flexible information infrastructure or sensate liner for medical applications. *Studies in health technology and informatics* 62 (1999), 252—258. http://europepmc.org/abstract/MED/10538367

[86] Maxine X. Patel, Victor Doku, and Lakshika Tennakoon. 2003. Challenges in recruitment of research participants. *Advances in Psychiatric Treatment* 9, 3 (2003), 229–238. https://doi.org/10.1192/apt.9.3.229

[87] Vikram Patel, Alan J Flisher, Sarah Hetrick, and Patrick McGorry. 2007. Mental health of young people: a global public-health challenge. *The Lancet* 369, 9569 (2007), 1302 – 1313. https://doi.org/10.1016/S0140-6736(07)60368-7

[88] Jennifer E. Pelletier, Dan J. Graham, and Melissa N. Laska. 2014. Social Norms and Dietary Behaviors among Young Adults. *American Journal of Health Behavior* 38, 1 (2014), 144–152. https://doi.org/doi:10.5993/AJHB.38.1.15

[89] A. Pentland. 2000. Looking at people: sensing for ubiquitous and wearable computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 1 (Jan 2000), 107–119. https://doi.org/10.1109/34.824823

[90] pewinternet. 2019. Social Media Fact Sheet. https://www.pewinternet.org/fact-sheet/social-media/

[91] Gregory W. Schrimsher PhD and Katie Filtz BS. 2011. Assessment Reactivity: Can Assessment of Alcohol Use During Research be an Active Treatment? *Alcoholism Treatment Quarterly* 29, 2 (2011), 108–115. https://doi.org/10.1080/07347324.2011.557983 arXiv:https://doi.org/10.1080/07347324.2011.557983

[92] Shao-Meng Qin, Hannu Verkasalo, Mikael Mohtaschemi, Tuomo Hartonen, and Mikko Alava. 2012. Patterns, entropy, and predictability of human mobility and life. *PloS one* 7, 12 (2012).

[93] Mashfiqui Rabbi, Min Hane Aung, Mi Zhang, and Tanzeem Choudhury. 2015. MyBehavior: Automatic Personalized Health Feedback from User Behaviors and Preferences Using Smartphones. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. ACM, New York, NY, USA, 707–718. https://doi.org/10.1145/2750858.2805840

[94] Kiran K. Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J. Rentfrow, Chris Longworth, and Andrius Aucinas. 2010. EmotionSense: A Mobile Phones Based Adaptive Platform for Experimental Social Psychology Research. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing* (Copenhagen, Denmark) *(UbiComp '10)*. ACM, New York, NY, USA, 281–290. https://doi.org/10.1145/1864349.1864393

[95] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. ACM, New York, NY, USA, 429–435. https://doi.org/10.1145/3306618.3314244

[96] Debora Rizzuto, Nicola Orsini, Chengxuan Qiu, Hui-Xin Wang, and Laura Fratiglioni. 2012. Lifestyle, social factors, and survival after age 75: population based study. *BMJ* 345 (2012). https://doi.org/10.1136/bmj.e5568 arXiv:https://www.bmj.com/content/345/bmj.e5568.full.pdf

[97] Aaqib Saeed, Tanir Ozcelebi, Stojan Trajanovski, and Johan Lukkien. 2018. Learning behavioral context recognition with multi-stream temporal convolutional networks. *arXiv preprint arXiv:1808.08766* (2018).

[98] A. Sano and R. W. Picard. 2013. Stress Recognition Using Wearable Sensors and Mobile Phones. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 671–676. https://doi.org/10.1109/ACII.2013.117

[99] D. Santani, T. Do, F. Labhart, S. Landolt, E. Kuntsche, and D. Gatica-Perez. 2018. DrinkSense: Characterizing Youth Drinking Behavior Using Smartphones. *IEEE Transactions on Mobile Computing* 17, 10 (Oct 2018), 2279–2292. https://doi.org/10.1109/TMC.2018.2797901

[100] Sandra Servia-Rodríguez, Kiran K. Rachuri, Cecilia Mascolo, Peter J. Rentfrow, Neal Lathia, and Gillian M. Sandstrom. 2017. Mobile Sensing at the Service of Mental Well-being: A Large-scale Longitudinal Study. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 103–112. https://doi.org/10.1145/3038912.3052618

[101] Edmund Seto, Jenna Hua, Lemuel Wu, Victor Shia, Sue Eom, May Wang, and Yan Li. 2016. Models of Individual Dietary Behavior Based on Smartphone Data: The Influence of Routine, Physical Activity, Emotion, and Food Environment. *PLOS ONE* 11, 4 (04 2016), 1–16. https://doi.org/10.1371/journal.pone.0153085

[102] Kristina J Shelley. 2005. Developing the American time use survey activity classification system. *Monthly Lab. Rev.* 128 (2005), 3.

[103] Roberta Sinatra and Michael Szell. 2014. Entropy and the predictability of online life. *Entropy* 16, 1 (2014), 543–556.

[104] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.

[105] Pitirim Aleksandrovich Sorokin and Clarence Quinn Berger. 1939. *Time-budgets of human behavior*. Vol. 2. Harvard University.

[106] Dimitris Spathis, Sandra Servia-Rodriguez, Katayoun Farrahi, Cecilia Mascolo, and Jason Rentfrow. 2019. Passive Mobile Sensing and Psychological Traits for Large Scale Mood Prediction. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare* (Trento, Italy) *(PervasiveHealth'19)*. ACM, New York, NY, USA, 272–281. https://doi.org/10.1145/3329189.3329213

[107] Hyuna Sung, Rebecca L Siegel, Philip S Rosenberg, and Ahmedin Jemal. 2019. Emerging cancer trends among young adults in the USA: analysis of a population-based cancer registry. https://doi.org/10.1016/S2468-2667(18)30267-6

[108] Megumi Tabuchi, Takeshi Nakagawa, Asako Miura, and Yasuyuki Gondo. 2013. Generativity and Interaction Between the Old and Young: The Role of Perceived Respect and Perceived Rejection. *The Gerontologist* 55, 4 (11 2013), 537–547. https://doi.org/10.1093/geront/gnt135 arXiv:http://oup.prod.sis.lan/gerontologist/article-pdf/55/4/537/19444595/gnt135.pdf

[109] Edison Thomaz, Irfan A. Essa, and Gregory D. Abowd. 2015. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. *Proceedings of the ... ACM International Conference on Ubiquitous Computing . UbiComp* 2015 (2015), 1029–1040.

[110] Edison Thomaz, Aman Parnami, Irfan Essa, and Gregory D. Abowd. 2013. Feasibility of Identifying Eating Moments from First-person Images Leveraging Human Computation. In *Proceedings of the 4th International SenseCam &#38; Pervasive Imaging Conference* (San Diego, California, USA) *(SenseCam '13)*. ACM, New York, NY, USA, 26–33. https://doi.org/10.1145/2526667.2526672

[111] Eran Toch, Yuhuai Wang, and Lorrie Faith Cranor. 2011. Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction* 22 (2011), 203–220.

[112] Britta N. Torgrimson and Christopher T. Minson. 2005. Sex and gender: what is the difference? *Journal of Applied Physiology* 99, 3 (2005), 785–787. https://doi.org/10.1152/japplphysiol.00376.2005 arXiv:https://doi.org/10.1152/japplphysiol.00376.2005 PMID: 16103514.

[113] Zhen Tu, Runtong Li, Yong Li, Gang Wang, Di Wu, Pan Hui, Li Su, and Depeng Jin. 2018. Your apps give you away: distinguishing mobile users by their app usage fingerprints. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–23.

[114] Alzheimers Research UK. 2019. Play Sea Hero Quest and gameforgood. https://www.alzheimersresearchuk.org/our-research/what-we-do/sea-hero-quest/

[115] European Union. 2019. *Reporton third gender markeror no gender marker options.* Retrieved Nov 13, 2019 from https://eugdpr.org/

Co-funded by the Horizon 2020
Framework Programme of the European Union

[116] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. 2017. Recognizing detailed human context in the wild from smartphones and smartwatches. *IEEE Pervasive Computing* 16, 4 (2017), 62–74.

[117] Yonatan Vaizman, Nadir Weibel, and Gert Lanckriet. 2018. Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–22.

[118] Alexander J.A.M. van Deursen, Colin L. Bolle, Sabrina M. Hegner, and Piet A.M. Kommers. 2015. Modeling habitual and addictive smartphone behavior: The role of smartphone usage types, emotional intelligence, social stress, self-regulation, age, and gender. *Computers in Human Behavior* 45 (2015), 411 – 420. https://doi.org/10.1016/j.chb.2014.12.039

[119] Tim Van hamme, Giuseppe Garofalo, Enrique Argones Rúa, Davy Preuveneers, and Wouter Joosen. 2019. A Systematic Comparison of Age and Gender Prediction on IMU Sensor-Based Gait Traces. *Sensors* 19, 13 (2019). https://doi.org/10.3390/s19132945

[120] James Vincent. 2018. IBM hopes to fight bias in facial recognition with new diverse dataset. https://www.theverge.com/2018/6/27/17509400/facial-recognition-bias-ibm-data-training

[121] Eugene Volokh. 2000. Personalization and Privacy. *Commun. ACM* 43, 8 (Aug. 2000), 84–88. https://doi.org/10.1145/345124.345155

[122] Scott T. Walters, Amanda M. Vader, T. Robert Harris, and Ernest N. Jouriles. 2009. Reactivity to alcohol assessment measures: an experimental test. *Addiction* 104, 8 (2009), 1305–1310. https://doi.org/10.1111/j.1360-0443.2009.02632.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1360-0443.2009.02632.x

[123] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) *(UbiComp '14)*. ACM, New York, NY, USA, 3–14. https://doi.org/10.1145/2632048.2632054

[124] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. 2018. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 43 (March 2018), 26 pages. https://doi.org/10.1145/3191775

[125] Mattia Zeni, Ilya Zaihrayeu, and Fausto Giunchiglia. 2014. Multi-device activity logging. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. 299–302.

[126] Mattia Zeni, Ilya Zaihrayeu, and Fausto Giunchiglia. 2014. Multi-Device Activity Logging. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication* (Seattle, Washington) *(UbiComp '14 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 299–302. https://doi.org/10.1145/2638728.2638756

[127] Mattia Zeni, Wanyi Zhang, Enrico Bignotti, Andrea Passerini, and Fausto Giunchiglia. 2019. Fixing Mislabeling by Human Annotators Leveraging Conflict Resolution and Prior Knowledge. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 32 (March 2019), 23 pages. https://doi.org/10.1145/3314419

[128] Wanyi Zhang, Andrea Passerini, and Fausto Giunchiglia. 2020. Dealing with Mislabeling via Interactive Machine Learning. *KI - Künstliche Intelligenz* (01 2020). https://doi.org/10.1007/s13218-020-00630-5

[129] Sean Zheng. 2015. *WHAT ARE GENDER STATISTICS.* Retrieved Nov 14, 2019 from https://unstats.un.org/unsd/genderstatmanual/What-are-gender-stats.ashx

[130] Erheng Zhong, Ben Tan, Kaixiang Mo, and Qiang Yang. 2013. User demographics prediction based on mobile data. *Pervasive and Mobile Computing* 9, 6 (2013), 823 – 837. https://doi.org/10.1016/j.pmcj.2013.07.009 Mobile Data Challenge.

Co-funded by the Horizon 2020
Framework Programme of the European Union