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

WENET
INTERNET OF US

D2.2 Advanced Individual Learning Methods

Revision: v.1.0

Work package	WP2
Task	Task 2.1, 2.2
Due date	30/06/2021
Submission date	30/06/2021
Deliverable lead	IDIAP
Version	1
Authors	Lakmal Meegahapola (Idiap Research Institute, Switzerland) William Droz (Idiap Research Institute, Switzerland) Daniel Gatica-Perez (Idiap Research Institute, Switzerland) Andrea Bontempelli (University of Trento, Italy) Fausto Giunchiglia (University of Trento, Italy) Salvador Ruiz-Correa (IPICYT, Mexico)
Reviewers	Carles Sierra (CSIC)

Abstract	This deliverable will outline the results from the mexico pre-pilot and an initial analysis of the ilog data from the diversity pilot.
Keywords	Mobile sensing, Machine Learning, Smartphone Sensing, Behaviour, Routines

Document Revision History

Version	Date	Description of change	List of contributor(s)
V0.1	30/06/2021	1st version with table of content and structure	Lakmal Meegahapola (IDIAP)
V0.2	30/06/2021	First draft of the deliverable	Lakmal Meegahapola (IDIAP) William Droz (IDIAP) Daniel Gatica-Perez (IDIAP) Andrea Bontempelli (UNITN) Fausto Giunchiglia (UNITN) Salvador Ruiz-Correa (IPICYT)

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

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, UNITN, and IPICYT.

As stated in the proposal, WP2 has three tasks:

T2.1. Diversity-aware routine learning *[Lead: IDIAP]*. 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].* 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].* 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 a set of individual learning methods for the project. In summary, the work described in this document spans six outcomes:

(1) Inferring Food Consumption Level Using Smartphone Sensing (related to Task T2.1, conducted by IDIAP and IPICYT).
(2) Understanding Eating Episodes with Mobile Sensing (related to Task T2.1, conducted by IDIAP and IPICYT).
(3) Privacy Protection of Mobile Food Diaries (related to Task T2.1, conducted by IDIAP and IPICYT).
(4) Handling Human Annotator Mistakes and Knowledge Drift (related to Task T2.2, conducted by UNITN).
(5) First Analysis of WeNet Pilots in the UK, Denmark, Mongolia, and Paraguay (related to Task T2.1, conducted by IDIAP).

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

Contents

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

List of Figures

List of Tables

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

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

2 INFERRING FOOD CONSUMPTION LEVEL USING SMARTPHONE SENSING

While the characterization of food consumption level has been extensively studied in nutrition and psychology research, advancements in passive smartphone sensing have not been fully utilized to complement mobile food diaries in characterizing food consumption levels. In this study, we examine the WeNet Mexico pre-pilot dataset, first introduced in Wenet's Deliverable D2.1 regarding the holistic food consumption behavior of 84 college students that was collected using a mobile application combining passive smartphone sensing and self-reports. We show that factors such as sociability and activity types and levels have an association to food consumption levels. Finally, we define and assess a novel ubicomp task, by using machine learning techniques to infer self-perceived food consumption level (eating as usual, overeating, undereating) with an accuracy of 87.81% in a 3-class classification task by using passive smartphone sensing and self-report data. Furthermore, we show that an accuracy of 83.49% can be achieved for the same classification task by using only smartphone sensing data and time of eating, which is an encouraging step towards building context-aware mobile food diaries and making food diary based apps less tedious for users.

Many young adults show a tendency to adopt unhealthy eating practices during college years, when they undergo significant lifestyle changes such as leaving home, meeting new friends, starting a career, and developing relationships [102, 112]. Even though young adults are relatively healthy compared to other older populations, unhealthy eating habits at this age could lead to adverse health outcomes such as cardiovascular diseases, overweight conditions and obesity in the long term [13, 52, 102]. Due to these reasons, researchers in nutrition, behavioral science, and psychology are extensively studying causes and contexts of food consumption, specially among college students [66, 112, 154, 161]. Moreover, prior research in these domains have linked factors such as social context [64], eating location [48], availability and types of food [134], and psychological aspects [62] to food consumption behavior. With increasing smartphone coverage among young adults and the availability of a plethora of mobile health (mHealth) applications [97], smartphones have become a ubiquitous tool that can help young adults adhere to healthier food consumption practices [90].

To our knowledge, while food intake recognition has been studied in ubicomp research [24], the specific overeating phenomenon has not been studied using passive smartphone sensing and self-report datasets. Using such a rich combination of data sources allows to analyze eating behavior of college students using knowledge from nutrition and mobile sensing research by associating food consumption levels to aspects such as mobile app usage, location, activity levels, sociability, and food types. This approach allows for comparisons with findings about self-perceived food consumption levels in prior nutrition and behavioral science research (which validates some of the observed trends), and also to provide novel insights regarding techniques to build mobile food journaling systems that leverage passive sensing to identify behaviors of college students associated to overeating, and to provide them with valuable insights and interventions regarding their food consumption.

Further details about this work can be found in this publication: Lakmal Meegahapola, Salvador Ruiz-Correa, Viridiana del Carmen Robledo-Valero, Emilio Ernesto Hernandez-Huerfano, Leonardo Alvarez-Rivera, Ronald Chenu-Abente, and Daniel Gatica-Perez. 2021. One More Bite? Inferring Food Consumption Level of College Students Using Smartphone Sensing and Self-Reports. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5, 1, Article 26 (March 2021), 28 pages. DOI: https://doi.org/10.1145/3448120

2.1 Technical Approach and Results

2.1.1 Defining Food Consumption Level. In prior research, food consumption level has had both objective interpretations (nutrition science-based) [26, 133, 156] and subjective ones (nutrition and psychology-based) [45, 121, 144, 145, 151, 156], and there is no unique way to define it [60, 81]. The *objective* food consumption level attempts to capture the exact calorie consumption during eating episodes from a purely nutritional standpoint. In lab studies, calorie intake is pre-calculated before offering food to participants [26]. Under this objective interpretation, a person should eat only as much as is necessary to offset her/his caloric demands, and overeating occurs if the food intake exceeds this amount [60]. Many currently available mobile food diaries such as MyFitnessPal [5], Samsung Health [10], and other research studies [41] attempt to capture this attribute using self-reports by requesting the users to enter each food type and the amount they eat. Even though the target here is to capture the objective calorie intake, there is by design a subjective element because users self-report it, and it is known that people often fail to report volume/weight of a dish accurately [156]. However, even if caloric intake is correctly reported and calculated, defining food consumption level as overeating and undereating according to this approach is complicated according to Herman et al. [60, 145], because it depends on a plethora of factors: (a) individual factors such as metabolic rates, activity levels, age, gender, height, weight; and (b) measurement factors, i.e., the unit of calculation for overeating is usually caloric deficit per day (nutritionists often do it at the day, week, or meal/snack episode level). This implies that for the same person, eating the exact same amount of food on a more active day could be overeating on a slow day. Hence, the process gets more complex as factors add on, and it could get particularly difficult and inaccurate if overeating and undereating episodes are determined based on self-reports that reflect food types and volumes. In addition, a recent study by Jung et al. [69] emphasized how currently available mobile food logging systems can be troublesome to users because of the tedious manual data entry process, hence leading to low adoption rates.

Contrary to the objective view of food consumption level, nutrition researchers have also widely used *subjective* measures to capture food consumption levels of people by considering the psychology of food consumption [71, 81, 119, 121, 134, 143–146], also known as self-perceived food consumption level. This view is primarily based on the idea that, when you ask people whether they overate or not, more often than not, the answer would be based on an eating episode level, and the self-perceived amount of food they have eaten [60]. This measure has often been used as a proxy to the actual amount of food people have eaten. Field et al. [45] showed that self-perceived food consumption level can be similar to real food consumption levels, and these self-reports are valid to determine bulimic episodes in adolescents. Moreover, Williamson et al. [156] examined the relation between self-reported caloric intake (similar to self-reports regarding caloric intake in MyFitnessPal and Samsung Health) and self-perceived overeating, concluding that there is a positive relationship between the two variables for all four groups of people they considered: (1) suffering from bulimia nervosa, (2) compulsive binge eaters, (3) obese, and (4) not having any of the three previous conditions. Due to the above mentioned factors, many prior studies have used self-perceived food consumption level as a proxy to the objective food consumption level, although we are not aware of any study that establishes detailed guidelines of when self-reported subjective overeating and objective overeating coincide or not. Furthermore, prior work in nutrition research suggests that adverse behavioral and emotional effects of overeating arise not only after people eat an objective large amount of food, but even when people simply think

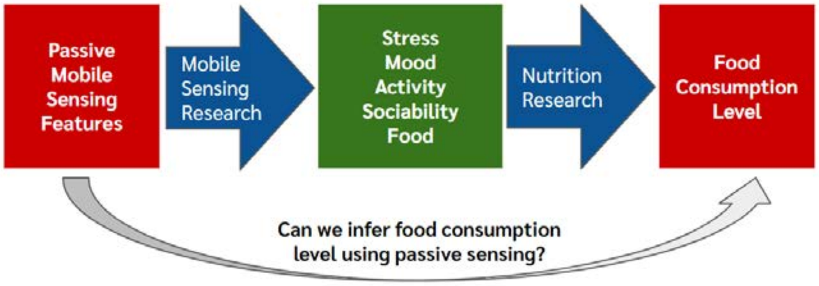

Fig. 1. Objective of the Study

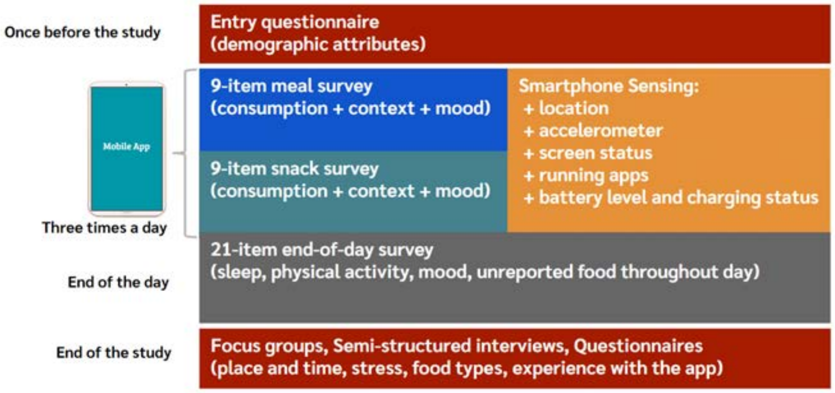
Once before the study
Entry questionnaire (demographic attributes)
Mobile App
Three times a day
9-item meal survey (consumption + context + mood)
9-item snack survey (consumption + context + mood)
Smartphone Sensing:
+ location
+ accelerometer
+ screen status
+ running apps
+ battery level and charging status
End of the day
21-item end-of-day survey (sleep, physical activity, mood, unreported food throughout day)
End of the study
Focus groups, Semi-structured interviews, Questionnaires (place and time, stress, food types, experience with the app)

Fig. 2. Block Diagram of Data Collection

that they have overeaten (self-perceived overeating) compared to their prior beliefs or current social context [81, 107, 108]. This is why many studies regarding psychology and eating behavior consider self-perceived food consumption level to be an important attribute, specially when considering eating as a holistic process to understand eating behavior [25, 68, 139].

Williamson et al. [156] captured overeating episodes by asking participants to report their perception on whether they overate or not (a binary choice). In a study by Ruddock and Hardman [121], self-perceived food consumption level was examined using a three-level coding system (eating more-than/less-than/as usual). Moreover, Vartanian et al. [145] used a five-point likert scale (1-5) in their study regarding food consumption levels where 1, 3, and 5 corresponded to "ate much less than I normally eat", "ate similar to the amount I normally eat", and "ate much more than I normally eat" respectively. By *normal* or *as usual*, what these studies meant is in comparison to their past behavior, and how they perceive societal norms regarding normal food intake. Following this literature, in this paper we define self-perceived food consumption level as *"eating more than (overeating condition), less than (undereating condition), or roughly the same (as usual condition) amount of food during an eating episode, in relation to the person's own estimated consumption, beliefs, and norms"*. Hence, from here onwards in this paper, we use the terms "food consumption level", "overeating", "eating as usual" or "undereating" to denote the self-perceived and self-reported attributes.

2.1.2 Study Objective and Hypothesis. The primary objectives of this study are to investigate links between food consumption level and features derived using passive sensing, and to leverage such links to automatically infer food consumption level, as summarized in Figure 1. Prior literature has shown that passive sensing features can be used to infer psychological and contextual aspects such as stress [82, 127], mood [79], activity types [7, 10, 24], sociability [14, 21, 57, 58], and food types [24, 96, 128]. In addition, a plethora of prior nutrition and behavioral science studies have linked the above aspects to food consumption levels [60, 61, 94, 146]. Knowing that smartphone sensing features have shown correlations to certain attributes, which have also been connected to food consumption levels as shown in Figure 1, our objective is to leverage these relationships studied in prior literature to use passive sensing for inference of food consumption levels. In other words, as passive sensing features have been linked to aspects such as stress, mood, activity, sociability, and food; and as these aspects have been linked in nutrition literature to food consumption levels like overeating; our hypothesis is that mobile sensing features could be used to infer self-percieved food consumption levels.

2.1.3 Mobile Application. We used a native android mobile application called i-Log to collect data from volunteers [162]. The app was developed at the University of Trento with Java, and data were initially stored in a SQLite database in the smartphone. Moreover, the system uses Google Firebase as a notification broker to send push notifications. When the phone is connected to a WiFi network and the phone has sufficient battery capacity, anonymized data were uploaded to Cassandra DB database in secure servers, hence freeing up the internal storage. The app has three main components: (a) push notification system to prompt users to complete questionnaires; (b) mobile surveys to record self-reports; and (c) passive smartphone sensing component to log sensor data.

2.1.4 Pre-Processing the Dataset for Analysis. The goal of our analysis was to investigate eating episode level data. Hence, we chose each food intake self-report as a data point in our dataset. To integrate sensor and survey data, we followed an approach suggested in prior mobile sensing literature [24, 124, 127], where for each event of focus, in this case for each eating episode, passive sensing data would be aggregated using a defined *time window*. We selected a time window of one hour which would mean that for each food intake event, we aggregate passive sensing data half an hour before and after the event starting time. We chose this time window considering prior research regarding characterizing eating events [24] and from a preliminary analysis regarding food consumption level. We started the procedure by finding the adjusted eating time because self-reports were done retrospectively. As mentioned in the previous section, we asked users "how long before they had the last meal". Using the answer for this question, we adjusted the timestamp of each food intake report to estimate the actual time of the eating episode. As an example, if the time of the self-report is 2pm, the answer for the question is 30-60 minutes ago (on average 30+60/2 = 45 minutes ago), the adjusted time of eating is estimated as 1.15pm (2pm - 45 minutes). Hence, using the one-hour time window, each eating event would be considered as a one hour eating episode. If the adjusted eating time is denoted by T, the time window would be the one hour from $T - 30$ minutes to $T + 30$ minutes. Next, we describe how each data modality was processed to associate it with eating episodes.

Accelerometer: following an approach similar to [24], for each 10-minute slot of the day, we generated features (aggregated sum of all values and sum of absolute values) using accelerometer value for axes x, y, and z. Then, depending on the adjusted time of an eating event (T), we considered three 10-minute bins before that eating episode (T-30 to T-20, T-20 to T-10, and T-10 to T), and three

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

Table 1. Pearson and Point-Biserial correlation analysis for some self-report features and food consumption level.

Features	Value	Features	Value
food consumption level	1 (+)	food meat sausages	.31827 (+)
mood	.35927 (+)	food fats oils	.32637 (+)
stress	.29719 (+)	food starches lugumes	.30412 (+)
food type	.22355 (+)	food softdrinks sugery juice	.25710 (+)
social context	.29165 (+)	food prepared dishes	.24144 (+)

10-minute bins after the start of the eating episode (T to T+10, T+10 to T+20, and T+20 to T+30). This way of pre-processing led to creating 18 features using accelerometer values. We use abbreviations to name the features generated using this methodology: (a) abs - calculated using absolute values of the accelerometer data; (b) bef - feature is calculated considering data before T, from T-30 to T; and (c) aft - feature is calculated considering data after T, from T to T+30.

Apps: we selected the ten most frequently used apps in the dataset. Then, during the hour associated with the eating episode, we determined whether each of those apps were used or not, hence resulting binary values for features in feature group *App*.

Location: using location traces, we calculated radius of gyration (a commonly used metric in mobile sensing [16, 160]) within the hour of consideration associated to the eating episode. Moreover, for each user, we generated stay regions throughout the whole day. Hence, using self-report labels (home, university, etc.), we generated labels for passively sensed stay regions of users, and we call that feature as *location* in our analysis. Moreover, for the location feature, we only used location degraded in precision for location privacy reasons (keeping only 4 decimal points).

Screen: using screen-on/off events in the dataset, we calculated the number of times the screen was turned on during the time slot, similar to prior literature [12, 15].

Battery: Similar to [15], we calculated the average battery level and also whether any charging events were detected during the time of eating episode. Battery and Screen events are used as proxies to smartphone usage behavior [12, 15].

2.1.5 Three-class Food Consumption Level Inference. The three-class inference task uses different subsets of features in the training set, and calculates classification accuracy, precision, and recall. The target classes were *overeating*, *undereating*, and *as usual*. We used python with scikitlearn and keras in this phase, and we conducted experiments using several model types (in the decreasing order of accuracies for inference task G5): random forest, naive bayes, gradient boosting, neural networks, XGboost, AdaBoost, and support vector classifiers. However, considering space limitations and aspects such as interpretability and model personalization, we present inference results for two models as follows:
(a) Random forest classifier (RF) with ntree values between 50 - 500: we got the highest accuracy values for inference tasks using RFs. More importantly, RFs models output the feature importance values used in inference, hence enabling us to understand and interpret the results.
(b) Neural network (NN) with 3-4 layers with relu activations, dropout for regularization, and binary cross entropy loss function (number of layers, number of nodes in intermediate layer/s changed

depending on the feature group and the input dimensions; hyper-parameter tuning was done for each feature group with the goal of obtaining the best model for the task).

We followed leave k-participants out strategy (k = 15) for all the experiments when preparing the dataset, where training and testing splits did not have data from the same user, hence avoiding this possible source of bias in the evaluation procedure. Moreover, when preparing the dataset, we made sure that the classes are balanced by up-sampling the minority classes and down-sampling majority class to get a balanced dataset of 2400 records. The baseline for experiments is 33.3% since the classes were balanced in all inference tasks. We conducted experiments for individual feature groups and meaningful feature group combinations:

Self Reports without FOOD (G1): This corresponds to self-reports that would not be available in a traditional mobile food diary, such as reports regarding eating context (sociability, concurrent activities), psychological state (mood, stress) together with the time of eating. The objective is to show that even without capturing the types and amounts of food, it is still possible to infer food consumption level. An envisaged application scenario of this inference is where the mobile health app simply captures these few self-reports instead of all the food consumption details, hence making the user experience better in terms of lower burden of manual data input.

Self Reports (G2): This corresponds to all the self-reported features including food types and categories. In addition to features in G1, this also captures the types of food consumed by participants. This inference would reaffirm the relationship shown in Figure 1 with regard to the association between food consumption level and aspects such as mood, stress, sociability, activities, and food.

Passive Smartphone Sensing with TIME (G3): This feature group combination contained a single self-reported variable (time of eating), and a set of passively sensed features without any user input, such as accelerometer, app usage, location, screen usage, and battery events. Importantly, this group reflects an envisaged mobile health application usage scenario where participants only report that they ate, hence capturing the time of eating, without typing all the details about the food types and amounts, sociability, concurrent activities, hence making it less tedious. In addition, prior work has examined the use of passive mobile sensing features to determine the time of eating [20, 115, 135], which is a separate open research question. Hence, this feature group combination denotes a envisaged use case that depends on *near-passive* sensing.

All Feature Groups without FOOD (G4): This contained all feature groups except for FOOD. Hence, this set of feature would require the same set of user involvement/effort as in G1. As we are following a holistic approach regarding food consumption, the goal here is to evaluate whether only knowing about the contextual factors and sensing data without knowing the food types and amounts could characterize the food consumption levels.

All Features (G5): This used all the available features to demonstrate the potential of a future mobile food diary that is driven by passive smartphone sensing in addition to traditional self-reports. This feature group captures food related details, contextual and socio-psychological attributes, and passive sensing data.

All Features w/ PCA (G6): Out of the 3 commonly used multi-modal fusion techniques: (a) data, (b) feature, and (c) decision) [39, 129], all inference tasks in this study except for G6 used feature-level fusion, that feeds a processed feature map into a classifier. In G6, we examined feature extraction

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

Table 2. Three-class food consumption inference (overeating, undereating, as usual) accuracy, precision, and recall obtained with a random forest classifier (RF) and a neural network (NN) using different feature group combinations.

Feature Group Name (# of Features)	**RF**			**NN**		
	Accuracy	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
Baseline	33.33%	-	-	33.33%	-	-
SCR (1)	40.26%	42.65%	40.97%	31.35%	33.46%	15.91%
APP (10)	47.52%	51.29%	46.74%	45.21%	48.68%	44.05%
PSY (2)	50.82%	52.36%	50.59%	44.88%	51.81%	44.54%
LOC (2)	56.43%	56.31%	56.68%	32.34%	29.63%	33.94%
CON (3)	62.37%	62.27%	62.37%	45.87%	55.08%	45.70%
BAT (2)	63.03%	61.58%	62.69%	50.03%	40.36%	51.64%
FOOD (15)	65.34%	66.11%	65.38%	60.72%	60.83%	60.78%
TIME (2)	67.65%	67.14%	67.01%	56.30%	56.04%	56.14%
ACC (18)	76.89%	83.89%	69.76%	47.52%	47.01%	57.89%
G1: CON + PSY + TIME (7)	81.19%	81.45%	80.91%	62.19%	62.39%	62.11%
G2: CON + PSY + TIME + FOOD (22)	82.50%	82.61%	83.56%	66.68%	67.29%	67.25%
G3: ACC + APP + LOC + SCR + BAT + TIME (35)	83.49%	83.19%	82.84%	73.26%	73.33%	72.20%
G4: ACC + APP + LOC + SCR + BAT + CON + PSY + TIME (40)	83.61%	83.99%	83.57%	79.20%	79.26%	79.17%
G5: All Features (55)	87.81%	87.97%	88.37%	82.17%	82.19%	82.95%
G6: All Features w/ PCA (principle components=4) (4)	83.53%	83.46%	83.54%	76.67%	76.94%	76.53%

and dimensionality reduction with principal component analysis (PCA), that fuses the features before feeding them into the classifier.

Results of the experiments (Table 6) show that RFs perform better than NNs across all feature groups and evaluation measures. Hence, in this section, only the results from RFs are discussed. Table 6 shows that individual feature groups such as ACC (accelerometer data), TIME (time of the day), FOOD (types of food consumed – similar to a traditional food diary), BAT (battery events), and CON (context when consuming food) have accuracies above 60%, while the highest accuracy of 76.89% corresponded to ACC feature group. This suggests that activity levels derived from the smartphone can be used to distinguish food consumption levels, to some degree. This is justifiable because prior work in nutrition and behavioral sciences have discussed the relation between food consumption levels and activity levels [60, 153]. G2 provides an idea regarding accuracies that can be obtained with currently available mobile food diaries that fully rely on participant self-reports. Accuracy, precision, and recall had values in the range 81% to 83% suggesting that self-reports are able to classify the three food consumption levels.

G3 shows that it has an even higher accuracy when compared to G1 or G2. Given that most prior research in nutrition has relied on self-reports regarding food categories and volumes to analyze food consumption behavior, this result shows that it is worth looking into passive smartphone sensing for cues regarding food consumption levels, given that there are many aspects such as app use, screen time, sociability that relate to the way people consume food in modern societies. Moreover, this result aligns with the hypothesis we presented in Section 2.1.2 and Figure 1 with regard to the possibility of inferring food consumption levels primarily using mobile sensing features.

G4, with an accuracy of 83.61% shows the benefit of considering eating as a holistic event as compared to just food categories and volumes. This accuracy, which is above the accuracy obtained with G2, again shows the importance of the holistic view of eating. Finally, by combining all the feature groups in G5, the model achieved an accuracy of 87.81%, precision of 87.97%, with a recall of 88.37%, all of which are encouraging. In addition, for G6, we got the best results with 4 principle components, an accuracy

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

of 83.53%, which is higher than G1 and G2, although it is still lower than the corresponding feature level fusion (G5) that used the same set of features. These result suggests that passive smartphone sensing can be of great value when incorporated to mobile food diaries that are currently based only on self-reports. Further, these results also highlight the potential that passive smartphone sensing has as part of mobile applications for food monitoring with less intrusive usage scenarios.

2.2 Discussion and Conclusion

2.2.1 Passive Smartphone Sensing for Characterizing Food Consumption Levels. Results presented in prior sections confirm that our hypothesis regarding food consumption levels and passive sensing features (Section 2.1.2) is valid. Most smartphones are capable of both continuous sensing (feature groups APP and SCR) and interaction sensing (feature groups such as ACC, BAT, TIME, LOC) for behavioral modeling. Obviously, these sensing modalities do not directly capture the food type or internal aspects that nutrition and behavioral science researchers have linked to food consumption levels in the past. However, in Section 2.1.5, we showed the potential of using passive smartphone sensing to infer food consumption level. What these modalities sense is not the food type or psychological aspects, but the physical activity levels and smartphone usage behavior. Given that physical activity levels have been linked to stress and mood in prior mobile sensing literature [114, 127], we believe these passive sensing modalities contain contextual information that could directly relate to food consumption behavior, and that is the reason why inferring food consumption levels with an accuracy of 83.49% using passive smartphone sensing and time related features was feasible. This is one of the first studies in this direction, and there are plenty of opportunities to explore *eating as a holistic event* as we did here. Given that computer vision researchers are focused on identifying food intake types and levels using images of the food portion, we could expect food consumption self-reports to get automated in future mobile food diaries. However, considering that eating is a holistic event driven by many factors, further research to determine food consumption behavior could enable advanced mobile food diaries that do no solely depend on user input to generate recommendations and interventions.

2.2.2 Further Informative Features Regarding Food Consumption Levels. We acknowledge that the features we generated from passive modalities are simple and easier to interpret when associated with eating episodes. However, there is ample opportunity to build upon these findings, and develop novel features that could discriminate food consumption levels with higher accuracies. For example, in this study, all the accelerometer features are single dimensional and we did not use linear acceleration or 2D resultant acceleration features due to some limitations in the data collection process (not having data from gyroscope to match accelerometer traces so that gravity biases could be removed [19, 59]). Moreover, when considering app usage behavior, the features we used only determined whether a particular app was used or not during the time window of the eating episode. However, advanced research could be done to determine usage times of each app during eating episodes, hence obtaining a comprehensive understanding regarding app usage behavior related to eating. Moreover, using a low-power API such as Google Activity Recognition API to detect activity types could generate new features that might be beneficial in characterizing eating events.

2.2.3 Accounting for Diversity. The eating behavior of people in different countries vary depending on a plethora of factors such the culture, type of food they consume, concurrent activities while eating, and how they perceive events such as eating [40, 76, 147]. Hence, it is important to clarify that the results from the deployment of our application in Mexico are exploratory and not representative of the food consumption behavior of people from other regions. Moreover, if other aspects apart from food are considered, there are already known differences with regard to factors such as sociability [105, 122], activity levels [18, 55], and phone usage [80, 106] in different countries, and these aspects could get reflected in smartphone sensing datasets. For example, a study regarding sociability of university students in Mexico and USA showed that Mexican students perceived themselves to be less sociable compared to how Americans perceived themselves, although in reality Mexicans were more sociable [116]. Moreover, results show that Americans socialize more in private environments or by interacting through social media. On the other hand, Mexican students preferred to be more social in-person with people who are around them [116]. Hence, we could expect differences in passive sensing data obtained from students in these two countries. It is fundamental to consider human diversity in smartphone sensing studies, and we believe more studies should integrate these diversity aspects in the future. Hence, future research could look into deploying mobile food diaries with sensing capabilities in diverse user groups based on ethnicity, culture, and geographic regions. In our opinion, the goal of such studies is to build models for mobile food diaries that generalize well enough to cater and adapt to diverse user populations. Even though our study is focused on college students of a Latin American country, we believe that this is a first step in this direction.

3 UNDERSTANDING THE SOCIAL CONTEXT OF EATING EPISODES WITH MOBILE SENSING

Understanding food consumption patterns and contexts using mobile sensing is fundamental to build mobile health applications that require minimal user interaction to generate mobile food diaries. Many available mobile food diaries, both commercial and in research, heavily rely on self-reports, and this dependency limits the long term adoption of these apps by people. The social context of eating (alone, with friends, with family, with a partner, etc.) is an important self-reported feature that influences aspects such as food type, psychological state while eating, and the amount of food, according to prior research in nutrition and behavioral sciences. In this work, we use two datasets regarding the everyday eating behavior of college students in two countries, namely Switzerland (N_{ch}=122) and Mexico (N_{mx}=84), as an additional analysis on the WeNet Mexico pre-pilot dataset, to examine the relation between the social context of eating and passive sensing data from wearables and smartphones. Moreover, we design a classification task, namely inferring *eating-alone vs. eating-with-others* episodes using passive sensing data and time of eating, obtaining accuracies between 77% and 81%. We believe that this is a first step towards understanding more complex social contexts related to food consumption using mobile sensing. In addition, this is a first exploration of diversity as reflected in data from two rather different countries.

Food and Nutrition has risen as the second most common category of apps used by mHealth app users according to recent reports [97]. Most of these apps have already incorporated mobile food diaries/journals to provide basic temporal insights to users. Even though many food diary-based studies have been carried out in the past with encouraging results for food consumption related interventions [95, 165], passive smartphone sensing has just begun to be widely used, comparatively speaking, in conjunction with food diaries, to understand contextual aspects that affect food consumption [24, 128].

Mobile food diaries use two main techniques to capture data [24, 124, 128]. They are: (a) Passive Sensing - using embedded sensors in smartphones and wearables (accelerometer, gyroscope, location, etc.) and events generated from the phone (app usage, screen-on time, and battery charging events), models can unobtrusively generate behavioral and contextual insights; and (b) Self-Reports - capture details regarding daily behavior and context related to eating. According to prior literature in mobile sensing [24, 128] where eating is considered as a holistic event with interconnected dimensions [25, 139], the social context of eating is an important variable that is self-reported, as it is a factor that relates to many aspects regarding food consumption episodes such as location, time, psychological state while eating, physical conditions, and food amount.

Studies have found that eating in highly social contexts (partying, celebrations, gatherings, etc.) can influence the amount of food consumed, which might lead to overeating in the short term [26, 31, 103] and to eating disorders in the long term [35, 131]. Further, concepts such as *social facilitation* and *impression management* emphasize how the presence of one or more people when eating can lead to overeating and undereating, respectively [60, 61, 63]. Moreover, studies have examined the effects of *eating-alone* and *eating-with-others* as fundamental aspects regarding eating behavior [63, 159]. Hence, understanding the social context of eating has been outlined as an important component of food consumption research [32, 38, 42, 50, 56, 139]. Furthermore, automatically inferring attributes related to social context would enable mobile food diaries to send context-aware notifications [70] and to support interventions [104], and also to help users adhere to healthy eating practices [50, 64]. In this study, similar to prior mHealth sensing studies with food diaries [24, 92, 128], we consider eating to be a holistic event, and use a binary categorization for the social context of eating – *eating-alone* vs. *eating-with-others* as a construct to understand food consumption behavior of college students in two countries.

Further details can be found in this publication: Lakmal Meegahapola, Salvador Ruiz-Correa, and Daniel Gatica-Perez. 2020. Alone or With Others? Understanding Eating Episodes of College Students with Mobile Sensing. In 19th International Conference on Mobile and Ubiquitous Multimedia (MUM 2020). Association for Computing Machinery, New York, NY, USA, 162–166. DOI: https://doi.org/10.1145/3428361.3428463

3.1 Datasets and Pre-Processing

The feature groups we used are temporal (**T**), contextual (**C**), and activity (**A**). Further, data sources are denoted by the by sub indices denoting self-reports ($_{sr}$), fitbit ($_{fb}$), smartphone ($_{sp}$), and other passive sensing modalities ($_{ps}$).

Switzerland Dataset (CH-Dataset): We used a mobile sensing dataset called *Bites'n'Bits* collected in our group's previous work [24, 49]. It contains smartphone sensor data, self-reported data, and activity data of fitbit wearables from 122 students of a Swiss university. The smartphone application allowed users to self-report details regarding eating events in-situ (**T**: time of eating; $\mathbf{C}_{sr}$: social context of eating, food type, concurrent activities, etc.). Further, their activity levels were captured using a fitbit wearable ($\mathbf{A}_{fb}$: step count, physical activity level), and activity level features were derived using the minutes spent on each of the four levels: sedentary, lightly active, fairly active, and very active. Moreover, passive sensing data regarding the context such as location were captured ($\mathbf{C}_{ps}$). In the final dataset, there are 4448 reports (3414 meals, 1034 snacks). All the participants 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.

Mexico Dataset (MX-Dataset): We used the MX dataset from 84 university student in San Luis Potosi, Mexico as mentioned in [93] and the previous section. The dataset had self-reported features similar to CH-Dataset ($\mathbf{C}_{sr}$ and **T** feature groups), and instead of the FitBit wearable (which we left out for cost reasons), activity levels of students were captured using the phone accelerometer ($\mathbf{A}_{sp}$: x, y, and z axis values of the accelerometer). Further, this dataset contained additional features about the participant context ($\mathbf{C}_{ps}$**:** app usage, radius of gyration, screen/battery charging events). The dataset contained 3278 reports (1911 meals, 1367 snacks). The average age of study participants was 23.4 years, and the cohort had 44% men and 56% women. A more detailed feature summary with naming conventions is available in [92].

Data Pre-Processing: During the feature extraction phase, standard datasets were created with one entry per eating event using a similar procedure to that given in [24]. For both datasets, if the eating event occured at time T, we aggregated sensing data from T-α to T+α (α_{CH} = 2 hours as chosen in [24], α_{MX} = 30 minutes, we present results for this value after examining performance of the model for different α).

Activity (CH and MX Datasets): For both datasets, (CH-Dataset: step counts and activity levels from fitbit; MX-Dataset: phone accelerometer), initially, the physical activity related features were calculated for 10-minute slots throughout the day. For the CH-Dataset: the total, median, mean, and standard deviation (*sd*) values of these features were calculated for α_{CH} before (*bef*) and after (*aft*) each eating

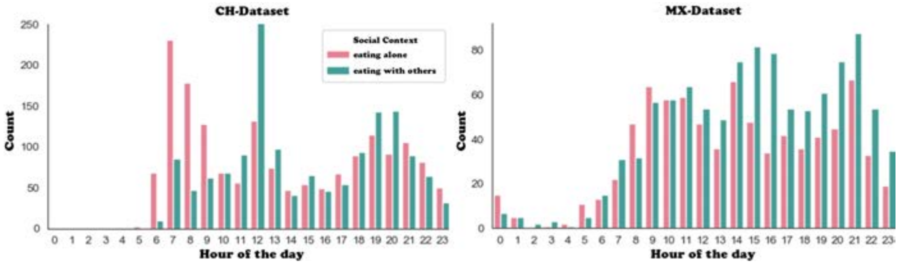

Fig. 3. Temporal variation for social context of eating

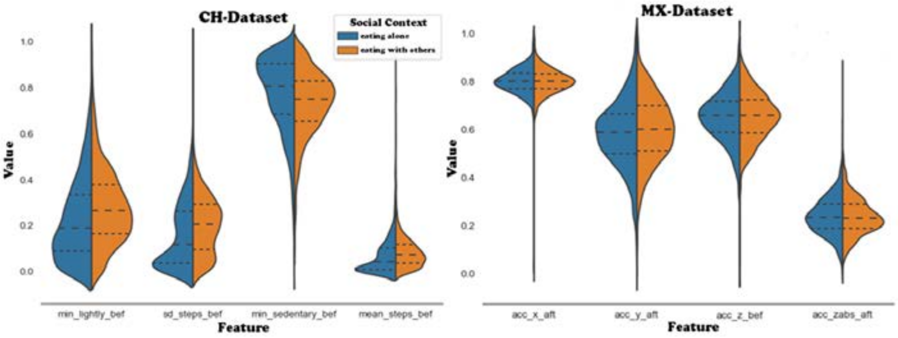

Fig. 4. Violin plots for selected activity features

event using 10-minute based values. For MX-Dataset: features were derived from three axes of the accelerometer sensor using absolute (*abs*) values and real values for α_{MX} before (*bef*) and after (*aft*) each eating event. Then, for the time window corresponding to the eating event, mean of feature values was calculated using 10-minute based values.
Apps (MX-Dataset): We selected the ten most frequently used apps in the dataset. Then, during the eating time window, we determined whether each of the apps have been used or not, hence resulting in binary values for all app related features.
Location (CH and MX Datasets): using location traces, we calculated radius of gyration [16, 160] within the time period associated to the eating episode.
Screen (MX-Dataset): using screen-on/off time slots, we calculated the screen-on time during the hour of consideration, and also the number of times the screen was turned on, similar to prior literature [12].
Battery (MX-Dataset): we calculated the average battery level during the hour of consideration and also whether any charging events were detected during the time of eating episode.

3.2 Descriptive Analysis

Temporal Variations. Figure 3 shows the temporal variation of the reported eating episodes for different social contexts of eating in both datasets. For both figures, three peaks can be seen that correspond to breakfast, lunch, and dinner, even though times at which these peaks occur are different in the two datasets. This could be due to cultural differences between the students in the two countries. Note that students in MX often live with their family while attending college, while this is less common with CH students. In the CH-Dataset, the time period from 6.00AM to 9.00AM (breakfast) results in

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

significantly high number of eating-alone episodes compared to eating-with-others episodes. However, in the MX-Dataset, breakfast peak occurs later closer to 9.00AM to 11.00AM and the differences in terms of social context of eating are minimal, even though it still favors eating-alone. In CH-Dataset, a significantly high number of eating-with-others episodes are present during the lunch peak, and a similar pattern can be seen in the MX-Dataset as well. This could be because students were eating in the university, with their friends. In the CH-Dataset, dinner episodes are more or less even in terms of social context. However, the MX-Dataset show that even the dinner episodes are reported in highly social contexts, again partly explained for the living- with-parents situation. Hence, as a summary, Swiss students have reported high number of eating-alone episodes for breakfast, high number of eating-with-others episodes for lunch, and evenly distributed dinner episodes. For the MX students, except for the slight lean towards eating-alone in the morning, most other eating episodes have been reported to be with others. This could also be justified by prior research in psychology that has shown that Mexicans are highly social [116].

Passive Activity Variations. Figure 4 shows differences in distributions of sensed activity levels for the two social contexts of eating, for few selected features (due to space limitations) using Violin plots [11]. In the CH-Dataset, activity level features captured using the wearable show significant mean differences for all four features shown here. For example, the distributions of feature *mean steps bef* (see Section 3.1 and [92] for feature descriptions) are different in terms of the shape, where eating-alone distribution is highly skewed towards lower values, meaning that activity levels are low for eating-alone episodes. A similar pattern can be seen for features such as *min lightly bef* and *sd steps bef*. The feature *min sedentary bef*, that corresponds to time spent in sedentary state, show higher values for eating-alone episodes. As a summary, lower physical activity levels around eating episodes correspond to eating-alone in the CH-Dataset. This is consistent with prior work that has shown that low physical activity levels correspond to less social contexts [130, 152]. However, in the MX-Dataset, the activity levels were sensed via the smartphone, and the mean differences are smaller for both social contexts of eating. Hence, for the MX case, it is less clear whether there are significant activity level differences depending on eating social context. However, it should be noted that the features generated regarding activity levels (based on availability) from the two datasets are different (fitbit for CH, phone accelerometer for MX), and this could have influenced the results.

3.3 Statistical Analysis

Table 3 shows statistics such as t-statistic [72], p-value [53], and cohen's-d (effect size) [77] for all the features in both datasets for two groups: eating-alone and eating-with-others. The objective is to identify features that discriminate between the two social contexts. In the table, the features are ordered by the descending order of cohen's-d value. We calculated cohen's-d [117] to help understand the statistical significance of the features because p-values are not sufficiently informative [78, 158]. To interpret cohen's-d, we used a commonly used rule-of-thumb: small effect size = 0.2, medium effect size = 0.5 and large effect size = 0.8. Moreover, we calculated 95% confidence interval for cohen's-d. A confidence interval that includes zero suggests statistical non-significance [78].

In the CH-Dataset, the feature with highest cohen's-d (medium effect size) is *time since last meal*, which is derived from temporal aspects regarding food consumption. Moreover, physical activity features such as *min lightly bef*, *sd steps bef*, *min sedentary bef*, and *mean steps bef* that were derived

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

Table 3. Comparative statistics of top 10 features across classes *eating-alone* and *eating-with-others*: t-statistic, p-value, and cohen's-d with 95% confidence intervals. Features are sorted based on the decreasing order of cohen's-d; 95% confidence interval of cohen's-d CI includes 0 = ^{x}; when considering p-values, p<0.0001=*, p<0.001=**, p<0.01=***.

Feature	**Feature Group**	**CH-Dataset**		**Feature**	**Feature Group**	**MX-Dataset**	
		cohen's-d	**t-statistic**			**cohen's-d**	**t-statistic**
time since last meal	T	0.57276	16.77194*	location	C_{ps}	0.37116	5.86325*
min lightly bef	A_{fb}	0.32612	9.54787*	time	T	0.25114	4.02973*
sd steps bef	A_{fb}	0.31644	9.27218*	charging event	C_{ps}	0.16151	2.70769**
location	C_{ps}	0.26917	7.87712*	acc y aft	A_{sp}	0.15354	2.45258
min sed bef	A_{fb}	0.26841	7.86013*	acc y abs aft	A_{sp}	0.12842	2.05463
time	T	0.21020	6.15727*	app google search	C_{ps}	0.11903^x	1.87712
concurrent activity	C_{sr}	0.20756	6.09132*	acc z bef	A_{sp}	0.11629^x	1.86609
mean steps bef	A_{fb}	0.20101	5.88661*	acc z abs aft	A_{sp}	0.10989^x	1.76660
tot steps bef	A_{fb}	0.20079	5.88032*	screen on	C_{ps}	0.09536^x	1.51781
min lightly aft	A_{fb}	0.12267	3.59117***	app spotify	C_{ps}	0.09453^x	1.48093

from activity level by users before the eating event show cohen's-d values larger than 0.2. Moreover, all the features had confidence intervals that did not include zero. Other two features in the top ten are location and time, that also have cohen's-d values larger than 0.2. As a summary, the CH-Dataset contained several features derived from fitbit that show discriminating signs between eating-alone and eating-with-others episodes.

When considering the MX-Dataset, the feature with highest cohen's-d (0.37) was *location*. However, the only other feature from this dataset that had a cohen's-d higher than small effect size was *time* (cohen's-d = 0.25). Even though several passive sensing features related to activity levels were in the top ten, only two features (*acc y aft* and *acc z abs aft*) had confidence intervals that did not include zero. Another passive interaction sensing modality that had a closer to small effect size was charging events with a cohen's-d of 0.16. Moreover, two app usage related passive sensing features (*app google search* and *app spotify*) were in the top ten. However, both these had cohen's-d confidence intervals including zero. As a summary, these results show that passive sensing features that quantify the activity levels, time of eating, and location have shown signs of discriminating capability between eating-alone and eating-with-others episodes in both datasets.

3.4 Inferring Eating Episodes: Alone or With Others

The goal of the 2-class inference task was to use different subsets of features in the training set, and calculate the accuracy, precision, and recall. The target binary variable was eating-alone vs. eating-with-others, which indicates this fundamental aspect of eating. Moreover, we used random forest classifiers (RF) with ntree values between 100 - 150. We followed leave k-participants out strategy for all the experiments when preparing the dataset, where training and testing splits did not have data from the same user. Moreover, when preparing the dataset, we made sure that the classes are balanced by upsampling the minority classes to get balanced datasets (similar to [24]). Further, we conducted the experiment for different feature groups, and feature group combinations such as (a) A: these are features generated using the activity data from fitbit wearables and smartphones. Features in this group are passively sensed, hence not needing any user interaction; (b) A+T: when temporal features such as time of the day and time since last food intake are combined with activity data, it provides a temporal variation of the activity levels; and (c) A+T+C_{ps}: this feature group contains only passive

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

Table 4. **Eating-alone vs. eating-with-others inference task**

	Feature Group	Accuracy	Precision	Recall
	Baseline	50.00%	-	-
CH-Dataset	A_{fb}	75.54%	75.52%	75.53%
	A_{fb}+T	80.31%	80.35%	80.29%
	A_{fb}+T+C_{ps}	80.89%	80.96%	80.90%
	A_{fb}+T+C_{sr}	89.97%	90.31%	89.56%
	A_{fb}+T+C_{ps}+C_{sr}	90.88%	91.11%	90.69%
MX-Dataset	A_{sp}	70.57%	71.11%	71.10%
	A_{sp}+T	72.30%	73.03%	72.92%
	A_{sp}+T+C_{ps}	77.73%	77.74%	77.77%
	A_{sp}+T+C_{sr}	78.29%	78.59%	78.31%
	A_{sp}+T+C_{ps}+C_{sr}	84.08%	84.24%	84.03%

sensing features that are activity data and contextual data, and time of eating; (d) A+T+C_{sr}: this feature group contains only passive activity sensing features and contextual data that were self reported; (e) A+T+C_{ps}+C_{sr}: this combines all the available passive sensing and self-report features from wearables and smartphones to conduct the inference task. The baseline for experiments is 50% since the classes were balanced in the inference task.

Results are summarized in Table 6. As shown there, by only using activity data captured via wearables and smartphones, the models reached accuracies of 75.54% and 70.57% for the CH-Dataset and MX-Dataset, respectively. These accuracies are considerably increased when using temporal features. The A+T+C_{ps} feature group shows how models that primarily use passive sensing data (even though T is self-reported, prior work has shown that the time of eating can be inferred to some extent with mobile sensing [20, 136]. Hence, this feature group combination could be considered as near-passive.) without high user interaction to input details regarding details regarding food type or calorie levels, can infer eating social context with accuracies of 80.89% and 77.73% for CH-Dataset and MX-Dataset, respectively. For this feature group, features such as time, time since last meal, location, and other activity related features were among the top five for both datasets, when considering feature importance values derived from the RFs. Moreover, when all the features used in mobile food diaries with sensing capabilities are considered, the inference accuracies reached higher values of 90.88% and 84.08% for the CH and MX datasets, respectively. These results show that precise wearable sensing and even smartphone sensing (that can be obtained regardless of the smartphone type or brand) are both useful to infer eating-alone vs. eating-with-others episodes. This could be seen as a first step towards enabling holistic mobile food diaries with reduced user burden, by inferring attributes that are typically captured with self-reports.

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

4 PRIVACY PROTECTION OF MOBILE FOOD DIARIES

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 these data streams, making inferences regarding user behavior, and using those inferences to benefit application users, while sometimes these data are shared with third parties as well. Human-centered ubiquitous systems need to ensure that sensitive attributes of users are preserved when applications provide utility to people based on such behavioral inferences. In this paper, we demonstrate that inferences of sensitive attributes of users (gender and body mass index) is possible using low-dimensional and sparse data coming from mobile food diaries (a combination of sensor data and self-reports). After exposing this potential risk, we demonstrate how modified deep learning architecture based on autoencoders and multi-task neural networks can be used for feature transformation to preserve sensitive user information while achieving high accuracies for application-related inferences (e.g. inferring the type of consumed food). Our work is based on a dataset of daily eating behavior of 122 college students in Switzerland and the Mexico pre-pilot dataset collected from 84 students (same CH-Dataset and MX-Dataset described in the previous section). This deepens the initial analysis about this topic presented in WeNet's Deliverable D2.1 in terms of number of studied inference tasks and of comparison of performance across two countries. We believe that researchers in both nutrition and ubiquitous computing need to be aware of such implications, and should take necessary precautions to preserve sensitive information of mobile food diaries when creating machine learning pipelines, storing data, and sharing it (even when anonymized) with third parties.

Further details can be found in this publication: Lakmal Meegahapola, Salvador Ruiz-Correa, and Daniel Gatica-Perez. 2020. Protecting Mobile Food Diaries from Getting too Personal. In 19th International Conference on Mobile and Ubiquitous Multimedia (MUM 2020). Association for Computing Machinery, New York, NY, USA, 212–222. DOI: https://doi.org/10.1145/3428361.3428468

4.1 Introduction

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 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, personalized services could be provided with better quality of service [4, 5, 9]. While some users might be willing to provide such data, other users would prefer to use the application without providing sensitive information, thus setting a trade-off/conundrum between personalization, privacy, and utility, when using applications and services [34, 137, 150]. How this conundrum plays a role in ubiquitous computing is described in a recent study [17] which emphasizes the need for privacy-preserving systems. Moreover, according to the terms of use of several mobile health apps [4, 5, 9], this is exactly why they use personalization for users who provide such sensitive information, and non-personalized algorithms for users who refuse to provide such data, but still opt to use the app.

Another concerning issue is that tech companies who own such low-dimensional data have been accused of selling data to third parties (i.e. advertisers, insurance companies, etc.), and there is not full clarity as to how companies use our data [17, 27, 74, 101, 111, 118]. According to recent reports from consumer protection agencies [1, 6], this trend is not diminishing because of the way our data are shared and stored. Even though data might be anonymized before sharing it with third parties, it is not fully understood whether such low-dimensional data can still be used to infer sensitive attributes without user consent, specially for health-related information including food intake and activity levels. For example, a health insurance company can obtain anonymized food intake data through data brokers, and use a machine-learning model to infer sensitive attributes such as BMI (an indicator of the overall weight condition of a person), which could guide the decision to insure a person or not [2, 3]. Even though it is an unethical practice, these risks exist.

4.2 Background

Recently, given the appearance of frameworks to regulate the collection and use of personal data like the European General Data Protection Framework (GDPR) [140], there is a push for explicitly not collecting personal information from app users without a clear purpose [36, 65, 109]. Hence, two problems in the current operation of mobile food diaries are; (1) we do not know whether companies who collect low-dimensional data can infer sensitive attributes regarding users even when users do not provide such information; and (2) we do not know if any third parties who get access to health and food related data (even anonymized data) can infer sensitive attributes, which constitutes a privacy risk to users [6, 37, 100, 111]. But, similar to gender recognition from tweets [148, 149], or videos and images [98], recent research has demonstrated that high-resolution, fine grained mobile sensor information such as raw accelerometer and gyroscope traces can be used to infer sensitive attributes such as gender [67, 142] and age [99, 142], demonstrating the ability of mobile apps and recommendation engines to leverage these inferences in providing targeted content. As a counter-measure to such privacy risks, there are approaches [84–86] that try to keep application inference accuracy (e.g. activity recognition, step count) sufficiently high while lowering sensitive inference accuracy by using different techniques such as randomization, filtering, mapping and replacement of dataset features [86] in order to reduce the possibility of inferring sensitive attributes from mobile sensing data, thus protecting users when data is sent to the server, or shared with third parties [87, 110, 141]. There are two important aspects regarding these recent set of studies. First, they use fine-grained, rich, and high-resolution mobile sensor information traces to infer gender and age. Second, using privacy preservation techniques, even though gender inference accuracy can be contained closer to 50%, those techniques allow generating low-dimensional features such as step count, activity type, with accuracies over 85%. However, most modern mobile sensing applications do not send high-dimensional raw data to the cloud [4], but process data on-device to a certain level to generate low-dimensional features, which are then sent to cloud for storage and cross-platform syncing. This poses the question of whether transforming raw data traces for privacy preservation, as suggested in recent studies, is that useful, given the open issue of whether low-dimensional features generated from privacy-preserved, high-resolution data really remove the possibility of inferring sensitive attributes.

We use the term "sensitive inference" for inferences performed in mobile food diaries and mobile health applications to benefit users. These inferences might reveal private or health related information (e.g. gender, BMI, weight, height, etc.), and hence are sensitive in nature. We use the term "application inference" for inferences that are done in mobile apps that match the original purpose, i.e., to provide

Table 5. Feature groups are Demographic (D), Contextual (C), Food Category (F), and Activity (A). Type describes whether the feature is categorical (CA) or numerical (NU), and if it is categorical, how many categories are represented by the feature. The total number of features are 18 and 44 in the CH and MX datasets, respectively.

CH-Dataset				MX-Dataset			
Feature	**Description**	**Type**	**Group**	**Feature**	**Description**	**Type**	**Group**
gender	Man/Woman	CA(2)	D	gender	Man/Woman	CA(2)	D
bmi	Body Mass Index category (high/low)	CA(2)	D	bmi	Body Mass Index category (high/low)	CA(2)	D
meal_snack	Whether it is a meal or a snack	CA(2)	F	meal_snack	Whether it is a meal or a snack	CA(2)	F
sweet	Whether it is a sweet food or not	CA(2)	F	fatty	Whether food is fatty or non-fatty	CA(2)	F
dairy	Whether the food contains dairy or not	CA(2)	F	meat	Whether the food contains meat or not	CA(2)	F
time_since_meal	Time in minutes, since the last meal	NU	C	time_since_meal	Time in minutes, since the last meal	NU	C
time_in_min	Time of the day	NU	C	time_in_min	Time of the day	NU	C
where	Location of eating	CA(10)	C	where	Location of eating	CA(10)	C
withwhom	Social context of a eating (alone, friends, etc)	CA(4)	C	withwhom	Social context of a eating (alone, friends, etc)	CA(8)	C
whatelse	Concurrent activities while eating	CA(17)	C	whatelse	Concurrent activities while eating	CA(11)	C
steps_X_Y	Features derived using fitbit step counts	NU	A	charging or not	Whether the phone is charging when eating	CA(2)	C
	X = total, median, mean or std. deviation			battery_level	phone battery level when eating	NU	C
	Y = bef/aft to indicate before eating or after			screen_on/off	Number of screen on/off events	NU	C
				rog	radius of gyration during eating time window	NU	C
				app_X	whether X app was used or not	CA(2)	C
					X = facebook, instagram, whatsapp, etc.		
				mood, stress	mood and stress while eating	CA(5)	C
				acc_A_B	Derived using accelerometer sensor	NU	A
					B = bef/aft to indicate before eating or after		
					A = Used indicate the X,Y, or Z axis		

functionalities that ultimately benefit users. In the context of this study, we use three useful inferences done using low-dimensional data such as inferring if a user eats a meal or a snack, or a specific food type (sweet or dairy products) at a particular moment. These inferences are important in mobile intervention applications because excessive snacking are linked to overweight or obesity, while other unhealthy eating patterns are also associated to a variety of health issues [24, 30, 44, 73, 123, 157]. Hence, inferring such episodes could be vital in future mobile food diaries [22, 88].

4.3 Protecting Sensitive Attributes

For conducting experiments, we used the CH-Dataset and MX-Dataset mentioned in the previous section, for this section as well. Dataset features including sensitive and application inference related attributes are summarized in Table 5. First we show that sensitive inferences are possible using smartphone sensing data. Then, two sections named methodology and results will discuss the privacy preservation technique. The methodology and results sections contain two subsections each. First, in order to facilitate the process of transforming dataset features such that sensitive attributes are protected, we train a Multi-Task Neural Network (MT-NN) [33] (Step 1). Step 2 describes the procedure to use an Autoencoder (AE) [75] together with the trained MT-NN to transform features using a modified loss function such that using the output data from the AE, sensitive inferences cannot longer achieve high accuracies, while still enabling high accuracies for application inferences.

4.3.1 Inferring Sensitive Attributes using Mobile Food Diaries. In this section, we examine the feasibility of inferring sensitive attributes using the two low-dimensional datasets. We used support vector machines, neural networks, and random forest classifiers for this task. Due to space limitations, we only report results from random forest classifiers that were marginally higher than neural networks. For this experiment, we used Random Forest Classifiers (RF) with an ntree values of range 200-500 for different feature groups. We used 10-fold cross validation during training, and when preparing the dataset, we made sure that the classes are balanced by up-sampling the minority class. It should be noted that we followed a leave-k-participants-out strategy for all the experiments, where training, validation, and testing sets did not include data from the same user. We ended up with datasets with sizes 4200 in the CH-Dataset and 1000 in the MX-Dataset (corresponding to single eating events) for the experiment.

Results of this experiment are summarized in Table 6. In the CH-Dataset, when using sensor and self-reported contextual information alone (C), the classifiers achieved an accuracy of 72.51% using RF for gender inference. When we included BMI to contextual data (C+D), the accuracies were increased to 74.39%. Accuracy was even higher when using C+A feature group. However, when additional demographic information (BMI category) was also used to form the feature group C+A+D, gender inference accuracy increased to 91.39% with RF. Similar results were attained for gender inference in MX-Dataset as well. Moreover, in the BMI category inference task, we used gender as the feature in the D feature group. Results for BMI inference showed reasonably high accuracies in the range 74%-76% for both datasets, for C+A feature group. C+A+D feature group showed accuracies of 89.12% for the CH-Dataset and 81.29% for the MX-Dataset, again showing how knowing one sensitive attribute makes it easier to infer another sensitive attribute. Furthermore, since we are specifically interested in demonstrating the effects of smartphone sensing and self-reported data, when presenting accuracy values for sensitive inference and application inferences in later sections, we only present the accuracies obtained with the contextual and physical activity feature (C+A) for both sensitive and application inferences.

Table 6. Gender and BMI Inference accuracy from the random forest classifiers (RF) when using different feature groups.

	CH-Dataset		MX-Dataset	
Feature Groups	**Gender**	**BMI**	**Gender**	**BMI**
Baseline	50.00%	50.00%	50.00%	50.00%
A	65.13%	67.41%	66.73%	65.79%
C	72.51%	70.49%	68.91%	67.46%
C+D	74.39%	72.72%	74.39%	73.64%
C+A	**77.38%**	**74.75%**	**77.21%**	**76.39%**
C+A+D	91.39%	89.12%	80.63%	81.29%

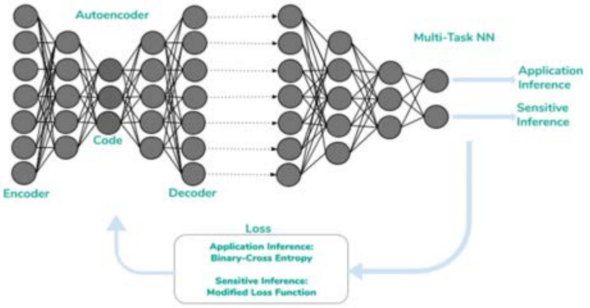
Autoencoder
Multi-Task NN
Application Inference
Sensitive Inference
Code
Encoder
Decoder
Loss
Application Inference: Binary-Cross Entropy
Sensitive Inference: Modified Loss Function

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

4.3.2 Methodology.

Step 1: Multi-task Neural Networks for Sensitive and Application Inferences. Most applications and third party services that use sparse, low-dimensional datasets such as the one studied here, use such data for application inferences. To show the conundrum between utility of application inferences and risks of sensitive inferences, we train a MT-NN, and show that the model is able to perform an application inference (e.g. meal vs. snack, sweet vs. non-sweet or dairy vs. non-dairy, etc.), and a sensitive inference (men vs. women or high-BMI vs. low-BMI) on the same dataset. We use examples of application inferences that mobile health apps could target, to illustrate the possibility of such joint

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

inferences. Example of such a joint inference is using a MT-NN to infer meal vs. snack and men vs. women in the CH-Dataset. Similarly, for each dataset, we considered six joint inference tasks (using two sensitive inferences and three application inferences), hence leading to a total of 12 inferences.

The MT-NN consisted of five layers, where the input layer had dense neurons equal to the number of input features. Intermediate layers had 32-64, 32-64 and 16-32 dense neurons, respectively depending on the inference task, whereas the two outputs corresponded to binary values representing the two inference tasks. Dropout was used for regularization in intermediate layers, relu was the activation function of intermediate layers, sigmoid activation was used for outputs, binary cross entropy was used to calculate loss for both inference tasks, and 10-fold cross validation was used. Even though the results hold for both C+A and C+A+D feature groups, we provide results only for the C+A feature group due to space limitations, and because that feature group represents a use-case where app servers have no sensitive information about users.

Step 2: An Autoencoder Based Architecture to limit Sensitive Inferences. We propose how deep learning techniques can be adjusted to suit a low-dimensional dataset, such that further privacy risks are reduced. Initially, we trained and tested the MT-NN as described in Step 1 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 + MT-NN based architecture is shown in Figure 5. We locked the weights of the MT-NN so that its weights do not get tuned during the training process of the AE, and then trained the AE using the training dataset.

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

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

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

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, denoted by $F_{ae}(.)$, where the resultant dataset from the autoencoder is $X_n^* = F_{ae}(X_n)$ such that $F_{sen}(X_n^*)$ accuracy is not high, hence preserving sensitive attributes about users, and $F_{app}(X_n^*)$ is high (closer to 100%), providing high inference accuracies for application inferences. 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 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 minimizing the loss for the training dataset. Finally, p is the probability of the outcome.

To make sure that AE learns its parameters 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/BMI (we use the value α=0.5 because it is desired accuracy for the binary

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

Table 7. The CH-Dataset: Accuracy for Application Inferences vs. Gender Inference and Application Inferences vs. BMI Category Inference using MT-NN and RF, before and after feature transformation using the AE. Results use C+A feature group

	Application Inference and Gender Inference						Application Inference and BMI Category Inference				
Task	Classification	MT-NN Before AE	MT-NN After AE	RF Before AE	RF After AE	Task	Classification	MT-NN Before AE	MT-NN After AE	RF Before AE	RF After AE
CH1	Meal vs. Snack	86%	81%	86%	85%	CH2	Meal vs. Snack	85%	84%	86%	82%
	Men vs. Women	67%	51%	77%	48%		High BMI vs. Low BMI	71%	48%	75%	53%
CH3	Sweet vs. Non-Sweet	83%	79%	82%	81%	CH4	Sweet vs. Non-Sweet	82%	79%	82%	80%
	Men vs. Women	69%	53%	77%	48%		High BMI vs. Low BMI	73%	45%	75%	52%
CH5	Dairy vs. Non-Dairy	78%	78%	73%	71%	CH6	Dairy vs. Non-Dairy	77%	76%	73%	72%
	Men vs. Women	76%	51%	77%	57%		High BMI vs. Low BMI	78%	54%	75%	44%

Table 8. The MX-Dataset: Accuracy for Application Inferences vs. Gender Inference and Application Inferences vs. BMI Category Inference using MT-NN and RF, before and after feature transformation using the AE. Results use C+A feature group

	Application Inference and Gender Inference						Application Inference and BMI Category Inference				
Task	Classification	MT-NN Before AE	MT-NN After AE	RF Before AE	RF After AE	Task	Classification	MT-NN Before AE	MT-NN After AE	RF Before AE	RF After AE
MX1	Meal vs. Snack	81%	78%	83%	79%	MX2	Meal vs. Snack	82%	79%	83%	79%
	Men vs. Women	77%	53%	76%	51%		High BMI vs. Low BMI	72%	49%	77%	54%
MX3	Fatty vs. Non-Fatty	80%	78%	82%	79%	MX4	Fatty vs. Non-Fatty	81%	80%	82%	81%
	Men vs. Women	79%	52%	76%	59%		High BMI vs. Low BMI	78%	51%	77%	60%
MX5	Meat vs. Non-Meat	84%	81%	85%	82%	MX6	Meat vs. Non-Meat	82%	78%	85%	79%
	Men vs. Women	79%	53%	76%	56%		High BMI vs. Low BMI	79%	53%	77%	59%

classification task to make sure that it has a lower accuracy [84]), 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/BMI classification is, it is not conveyed as it is to the AE due to the modified objective. This allows the AE to tune its weights such that resultant dataset after the feature transformation care less 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 data with transformed features using the AE.

As the final step, using the trained AE, we obtained a final dataset that is Privacy Preserved. We trained the RFs and NNs for sensitive and application inferences for both datasets using the final dataset. The intuition here is to check whether the modified dataset can provide good accuracies for application inferences, and lower the accuracies for sensitive inferences, even if a new model is trained.

4.3.3 Results.

Step 1: Multi-task Neural Networks can jointly infer sensitive attributes and eating events . Results from this experiment are shown under the column *MT-NN Before AE* in Table 7 and Table 8 for the CH-Dataset and MX-Dataset, respectively. In the CH-Dataset, the MT-NN achieved a meal vs. snack inference accuracy of 86%, sweet vs non-sweet inference accuracy of 83%, and dairy vs. non-dairy inference accuracy of 78%. These results are similar to results obtained using the RF (RF Before AE). Moreover, C+A feature groups provide significantly high accuracies for gender/BMI inference which highlights the need for privacy-preserving solutions for low-dimensional and sparse data from mobile

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

food journals. Similar results hold for the MX-Dataset where application inference accuracies using both RF and MT-NN were in the range 80%-85% and sensitive inference accuracies were in the range 72%-79% before using the AE based feature transformation.

Step 2: Our architecture limits sensitive inferences while providing utility for eating-related inferences. After training the AE to transform dataset features so that sensitive inferences are made difficult following the procedure given above, we measure both the application inference and sensitive inference accuracies for the transformed dataset using the newly trained RFs and MT-NNs. Table 7 and Table 8 show the results for the CH-Dataset and MX-Dataset respectively, using a comparison between accuracy results before and after the use of AE for MT-NN and RF for three inference pairs in both datasets. Application inference accuracies have been kept reasonably high for all three inference pairs in both datasets (the CH-Dataset: above 81% for MT-NN and 85% for RF in meal vs. snack and similar results hold for other two application inferences as well; the MX-Dataset: above 78% for meal vs. snack and similar results hold for other two inferences). At the same time, in the CH-Dataset, we were able to reduce the gender inference accuracy from 67% to 51% for MT-NN and from 77% to 48% for RF, and a similar trend can be seen for the BMI-category. A similar pattern in results can be seen for other two application inferences in the CH-Dataset, and for sensitive inferences in the MX-Dataset too. Hence the output from this procedure is still low-dimensional (similar to the original dataset), but also privacy preserving because the sensitive attributes can not be directly inferred with high accuracies from the resultant data even if a model is newly trained.

4.3.4 Generalization of our technique. In the results, we showed that our technique generalizes well to two datasets from mobile food diaries with passive sensing from two different countries. For both datasets, we attempted two sensitive inference tasks paired with three application inferences. Hence, we believe the above combination of datsets, sensitive inferences, and application inferences reasonably show the generalization potential of our technique. Moreover, it should also be noted that we were able to obtain similar results for other application inferences such as fruit vs. no-fruit and cereal vs. non-cereal too for both datasets, when used with both sensitive inferences gender and BMI category. However, the results are not included in the paper due to space limitations.

4.4 Discussion

Using Feature Transformation Techniques on High Dimensional vs. Low Dimensional Data. 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. For example, the step count of a person is derived by processing high-resolution accelerometer and gyroscope data 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 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. On the other hand, preserving sensitive attributes from low-dimensional

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

or low-resolution data might mitigate the privacy risk discussed here to a larger extent. Researchers and developers who use mobile sensing datasets should be aware of these findings, specially when they store or share data with other parties.

Data Before and After Feature Transformation. The feature transformation process proposed here 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 "with_family", "with_friends" and "with_date" before the transformation, and after the transformation resulted in numerical values. 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, the dataset after feature transformation would be uninterpretable unless the party using the transformed data had prior knowledge of the feature transformation process. This naturally protects the dataset from privacy risks from third parties who may gain access to the transformed data. For example, if a transformed dataset was shared with a third party by the data owner together with instructions regarding useful application inferences, it would be difficult for the third party to interpret data for other purposes. As another example, if the data was stored after feature transformation (i.e., in processed form) by the data owner, even if the data fell in the hands of a third party through hacking or a data breach, since the data was only interpretable for the original data owners, the dataset would become of less use for the third party. In other words, the technique we propose would create uninterpretable datasets for sharing and storage, increasing the likelihood that datasets are used only for required purposes, and not for anything else.

Dataset Diversity. A limitation of our study is the relative homogeneity of the participants who volunteered in the CH and MX datasets. The dataset used is from university students of two countries, hence, even though the participants are diverse in terms of eating routines, ethnicities, and behaviors, they are homogeneous in terms of age and occupation. While the results show evidence of sensitive inference using food diary entries, and that a feature transformation technique can preserve privacy, we believe that conducting a larger scale experiment more countries with people having different behavioral habits, ages, professions would shed more light into the results we present here. We hypothesize that even though using more diverse user populations might demonstrate varieties of eating behaviors, the technique we have proposed might still be useful.

Personalization, Privacy, and Utility. As researchers, we usually strive to enhance utility of applications and algorithms, and often use personalisation as a tool to increase utility. While this is important, an increasing body of work has also emphasized the importance of privacy preservation and the use of less sensitive data [17, 23, 43, 86, 89]. Personalization and privacy preservation are at the two opposite ends of the spectrum because personalisation has typically required more personal data to provide high utility, while privacy preservation aims at providing reasonable utility from the application, while preserving privacy of users from known risks. The trade-off between these goals are also reflected among 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 personalized services, there are other users who have concerns regarding application developers, and also regarding how their personal data would be used if they provided such information. As seen from the results, application inference utility slightly drops when privacy is preserved (after feature transformation). While we understand that personalisation of algorithms and services is an important research direction, 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 from foreseen and unforeseen circumstances. We believe that designing ubicomp technology for joint privacy and utility, and not only for personalisation, is important for the advancement of the field in a progressive and ethical manner. Recent literature further discusses why new privacy preservation techniques are needed by pointing out that simple anonymization techniques are no longer enough to preserve user privacy [17].

5 HANDLING HUMAN ANNOTATOR MISTAKES AND KNOWLEDGE DRIFT

5.1 Introduction

Smart personal assistants (PAs), smart environment systems and other AI applications need to recognize the context of the user in order to provide services. The context is defined by the location (e.g., the user's home), the activity (studying, working) and the social aspect (alone, with friends) [51]. For instance, the PA can suggest the user to take bus number 5 to get home by knowing that is visiting a specific museum and where the user's home is.

The context information is usually not available to the machine and needs to be inferred from a stream of sensors reading coming from the user's smartphone, such as GPS coordinates, acceleration values, nearby Bluetooth devices, WiFi networks. Personal devices like smartphones and smartwatches are always with the user, and thus, the PA can observe the users and their environment. The labels (aka classes or concepts) used to train a machine learning model to recognize the context are user specific (e.g., the user's home is not another user's home). Therefore, the machine interacts with the user to acquire them. However, the user often provides wrong labels that mislead the machine due to inattention or misunderstanding of the question [164].

These labels are structured in a hierarchy in which the labels are connected by *is-a* relations between them. For instance, if Bob is a friend of Ann, Ann's PA will maintain a hierarchy in which the label "Bob" is connected to "Friend", and the latter to "Person". Since the personal devices generate a continuous stream of data, acquired information can become obsolete and new information about the user becomes available. Hence, the label vocabulary and the relations between them are not static but evolve over time. The user visits new places, meets new persons and undertakes new activities over time.

We address these issues by proposing two methods. Section 5.2 presents Incremental Skeptical Gaussian Processes (ISGP) that handles the increasing number of classes and the noise in the user's annotation. ISGP recover the ground-truth labels from the user. Then, we introduce TRCKD in Section 5.3 to tackle the changes to the set of *is-a* relations among classes and to the set of classes. The machine adapts to changing world and user in order to provide useful, appropriate and correct suggestions.

5.2 Incremental classification in the wild: handling human annotator mistakes

AI systems deployed in real-world scenarios and interacting with end-user have to face the unreliability of the user when providing annotations. In WeNet, the users install the ILog mobile application [163] on their devices and answer the questions about their location, activities and social interaction administered regularly by the app. The answers are the target classes used to train a machine learning model that predicts the user's context. Unfortunately, the fraction of erroneous labels can be very high [138, 164] and they badly affect the performance of the classifier. Another challenge of this setting is that the interaction with the user occurs over time, and thus, the number of labels grows and must be efficiently integrated into the learned model.

We introduce a redesign of skeptical learning (SKL) [164], namely Incremental Skeptical Gaussian Processes, to overcome these issues. Skeptical learning is an interactive learning strategy in which the machine asks the user to review the label of the new example if it is confident that the example is mislabeled. The machine handles the noise by recovering the ground-truth class from the user, whereas other strategies implement robust models or discard examples [46].

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

Algorithm 1 Pseudo-code of ISGP.

```
for $t = 1, 2, \ldots$ do
    receive $\mathbf{x}_t$
    predict $\hat{y}_t$ for $\mathbf{x}_t$
    if uncertain about $\hat{y}_t$ then
        request label, receive $\tilde{y}_t$
        if skeptical about $\tilde{y}_t$ then
            challenge user with $\hat{y}_t$, receive $y'_t$
        else
            $y'_t \leftarrow \tilde{y}_t$
        add $(\mathbf{x}_t, y'_t)$ to data set and update GPs
        add $\{y'_t\}$ to known classes
```

EXAMPLE. *Ann is riding her bicycle. Ann's PA asks her "What are you doing?", and she answers "Running". The PA is suspicious because the acceleration values are inconsistent with the running activity. So, the PA contradicts Ann and asks "Are you running or cycling?".*

The improvements of ISGP over SKL are:

- ISGP leverages on the uncertainty estimation of Gaussian Processes (GPs) [155]. Thus, the use of GPs uncertainty prevents the issues deriving from being overconfident. The issues are that the machine contradicts the user continuously, regardless of the user reliability, and fails to learns from informative examples that are far from the training set;
- ISGP uses the uncertainty of GPs to decide when to be suspicious and thus to contradict the user;
- apart from the parameters of the GPs, ISGP does not have hyperparameters that need to be fine-tuned;
- ISGP implements an incremental learning approach that improves scalability [83].

In the following, we describe in detail ISGP and report the experimental results on synthetic and real-world data collected with ILog [163].

Further details about this work can be found here: Andrea Bontempelli, Stefano Teso, Fausto Giunchiglia, and Andrea Passerini. 2020. Learning in the Wild with Incremental Skeptical Gaussian Processes. *In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20.*

5.2.1 Technical approach.

Preliminaries. The classifier receives in input a stream of sensor readings $\mathbf{x}_t \in \mathcal{R}^d$ with $t = 1, 2, \ldots$ and outputs the prediction $\hat{y}_t \in \mathcal{Y}$. For instance, the prediction can be the location or the activity of the user. The ground-truth labels $y_t \in \mathcal{Y}$ are asked to the user, who may provide incorrect label, i.e., $\tilde{y}_t \neq y_t$. New classes appear over time and thus $\mathcal{Y}_t \subseteq \mathcal{Y}$. Hence, at each iteration t, $\hat{y} \in \mathcal{Y}_{t-1}$, $y_t \in \mathcal{Y}$ and $\tilde{y}_t \in \mathcal{Y}_t$.

Method. ISGP is built on Gaussian Processes (GPs) [155], a non-parametric distribution over functions that is fully described by a mean function $\mu(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$. The latter can be implemented with any kernel function and describes the assumptions about the underlying function that GPs try to model. We rely on multi-class GPs implementation that supports incremental updates [83]. The idea is to have a GPs for each label $l \in \mathcal{Y}_t$, which generates a collection of binary classification problems. The examples annotated with the label l are considered positive examples and all the rest as negative.

ISGP is presented in Algorithm 1. The classifier receives a new example $\mathbf{x}_t$ and predict a label $\hat{y}_t$ (line 3). The method has to decide whether to request a label to the user based on is uncertainty about its own prediction (line 4). Intuitively, the machine is uncertain if either $\mu(\mathbf{x})$ is small or $\sigma(\mathbf{x})$ is large, given $\sigma(\mathbf{x}) = \sqrt{k(\mathbf{x}, \mathbf{x})}$. The choice is made by sampling a_t from a Bernoulli distribution with the parameter β_t defined as:

$$\beta_t = 1 - \Theta\left(\frac{\mu_{\hat{y}_t,t}(\mathbf{x}_t)}{\sigma_t(\mathbf{x}_t)}\right)$$

where Θ is the cdf of a standard normal distribution, and if $a_t = 1$, the user is queried. The variance increases when the incoming instances are far from the training examples seen so far, and thus, the user is queried, allowing the machine to acquire the label of informative instances. When the user and the machine disagree about the label, i.e., $\tilde{y}_t \neq \hat{y}_t$, the machine must decide whether to contradict the user. As for the active query, the choice is made by sampling from a Bernoulli distribution with γ_t defined as:

$$\gamma_t = \Theta\left(\frac{\mu_{\hat{y}_t,t}(\mathbf{x}_t) - \mu_{\tilde{y}_t,t}(\mathbf{x}_t)}{\sigma_t(\mathbf{x}_t)}\right)$$

The intuition is that the probability of contradiction is high if the difference between the uncertainty of the machine on the predicted labels and on the user's label is high. In other words, the machine is very confident about its prediction. The user replies with y', which may still be noisy, but the user is not challenged a second time since we assume a collaborative user. After that, the model is updated with $(\mathbf{x}, y')$ (line 10). Further details about the method are reported in [29].

5.2.2 Experiments. We evaluated ISGP on synthetic and real-world data to answer the following research questions:

Q1 Has ISGP better predictive performance than the original formulation of skeptical learning?
Q2 Does ISGP identify mislabeled examples?
Q3 Does ISGP scale better than the original formulation of skeptical learning?

We compared out method against three variants: **SRF**, the original formulation of skeptical learning based on random forest; $\mathbf{GP}_{always}$, a variation of ISGP that is always suspicious about the user's label and always asks feedback; $\mathbf{GP}_{never}$ a variation of ISGP that never contradicts the user.

Synthetic data set. The method was run on a synthetic data set. The simulated user provides a wrong label η% of the time. We experimented with $\eta = 10$ and $\eta = 40$. The GPs use a squared exponential kernel. The examples were presented in a stream by choosing the examples for sequential classes, i.e., all examples of one class, then all example of another class. This simulates the appearance of new classes over time.

Two trends are visible in the results. First, when SRF is asked to reach the same F_1 of ISGP, it requests a label on all incoming examples overwhelming the user. This trend occurs because the method needs to remain in the training phase, i.e., it asks the label to the user for more iterations. Second, by limiting

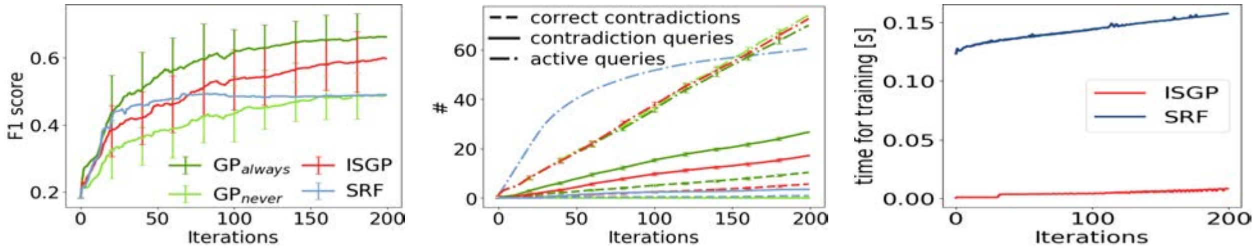

Fig. 6. Results on location prediction. **Left to right**: F_1 score, # of queries (cumulative), and run-time (not cumulative) as learning proceeds on the real-world data set (the training step is performed at each iteration).

the query budget, SRF becomes very confident in few iterations and stop querying the user. The consequences are that the following new classes are not acquired. The complete results can be viewed in [29].

Location prediction. We evaluated ISGP on a real-world data set introduced in [163] containing sensor readings coming from up to 30 sensors of the smartphones from 72 university students. The data were collected through the ILog mobile application [163], which asks every 30 minutes the location and the activity of the users, and with whom they are. In our experiment, the task is to predict the user's location (i.e., *Home*, *University* or *Others*), for which exists an oracle that provides the ground-truth annotation. The kernel of GPs is a combination of constant, rational quadratic, squared exponential and white noise kernel.

Figure 6 reports the experiment results where SRF is tuned to make the same number of queries of ISGP. The plots highlight that ISGP clearly outperforms SRF in terms of F_1 (leftmost plot). ISGP lies between the lower bound $\mathbf{GP}_{never}$ and the upper bound $\mathbf{GP}_{always}$. The F_1 of SRF reaches a plateau around iteration 70 due to the fact that the method becomes over-confident and queries less frequently the user (central plot), whereas the predictive performance of ISGP keeps increasing. The training time of SRF and ISGP on the real-world data set is shown in the rightmost plot of Figure 6. ISGP is clearly less expensive in terms of computational costs than SRF thanks to the incremental updates.

5.2.3 Conclusion. In this section, we introduced interactive classification in the wild, in which the learner receives a stream of examples and it can query a unreliable human annotator. We presented ISGP, a redesign of skeptical learning, that addresses the noise in the user's labels and the new classes that arrive over time. The empirical results show that ISGP avoid pathological cases where the user is always or never queried, and highlight the better predictive performance and the ability of identifying mislabeled examples.

5.3 Human-in-the-loop handling of knowledge drift

PAs are equipped with a concept hierarchy about the user and the world that is used to perform hierarchical classification. However, the set of concepts, the set of *is-a* relations between them and the their distribution change over time. We introduce knowledge drift (KD) to indicate this changes. The predictor needs to adapt to this changes to avoid wrong or irrelevant prediction.

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

Example. *Ann is working in her office with Bob. Ann's PA predicts "working", "office" and "Bob" as concepts describing her context. Moreover, the PA derives from the concept hierarchy that "Bob" is a "colleague" and a "person". If Bob resigns from his job, the is-a relation between "Bob" and "colleague" is removed.*

We identify four types of changes in the hierarchy: concept addition, concept removal, relation addition and relation removal. Concept addition occurs when a new concept is added, and a concept is removed when it becomes obsolete and no longer relevant. Relation addition and removal are the changes to the set of *is-a* relations. The main challenge is how to distinguish among the different types of drift. For instance, the addition of a relation between two concepts can be confused with a change in the distribution of the two concepts since they have similar effect on the data stream. If the machine fails to identify the kind of drift, it acquires a wrong hierarchy and this entails prediction errors on future instances. Moreover, errors can be propagated across the hierarchy and this leads to cascading prediction errors. Existing approaches for learning under concept drifts does not disambiguate between the different type of changes [47].

In human-in-the-loop applications like PAs, the user knows which type of drift occurred. For instance, considering our previous example, Ann is aware that she does not work with Bob anymore and that he is no longer a colleague. Based on this observation, we introduce a method to deal with KD, namely trckd (TRacking Knowledge Drift). It combines three steps: automated drift detection, interactive drift disambiguation and knowledge aware adaptation strategy. trckd is build on top of a multi-label *k*NN [132].

The contribution of this work are:

- the introduction of knowledge drift, a form of concept drift that occurs in hierarchical classification;
- the design of an approach for handling knowledge drift, namely trckd. The approach implements a completely new interactive drift disambiguation stage;
- empirical evaluation of trckd on three data sets that shows that it outperforms the competitors by asking few user queries.

In the rest of this section, we present trckd in details and we report the main results of the experiments. Additional details can be found in this publication under review: Andrea Bontempelli, Fausto Giunchiglia, Andrea Passerini, Stefano Teso. 2021. Human-in-the-loop Handling of Knowledge Drift. arXiv preprint https://arxiv.org/abs/2103.14874.

5.3.1 Technical approach.

Preliminaries. In hierarchical classification, the concepts (aka classes) are organized in a ground-truth hierarchy H, a direct acyclic graph where nodes are the concepts and the edges are the *is-a* relations between them. The instance $\mathbf{x}$ belongs to one or more concepts represented by the indicator vector $\mathbf{y}$ in which the ith element of $\mathbf{y}$ is 1, i.e., $y^i = 1$, if $\mathbf{x}$ belongs to the ith concepts in H and 0 otherwise. The machine observes a stream of examples $\mathbf{z}_t = (\mathbf{x}_t, \mathbf{y}_t)$ drawn from a ground-truth distribution $P_t(\mathbf{X}, \mathbf{Y})$ that is always consistent with the ground-truth hierarchy. This means that if the jth concept *is-a* specialization of the ith concept, then $y^j = 1$ implies $y^i = 1$ and conversely $y^i = 0$ implies $y^j = 0$. The goal is to learn a classifier that perform well on future instances. Learning in a dynamic enviroment means that both P_t and H_t can both change over time and they cannot be observed directly. We refer to this changes as knowledge drift.

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

Algorithm 2 Pseudo-code of TRCKD. Here, S_1 is the initial data set, w is the window size. $\mathbf{z}_t^i := (\mathbf{x}_t, y_t^i)$

```
 1: fit predictor on $S_1$
 2: for every concept $i$ in the machine's hierarchy do
 3:     $W^i_{\text{past}}$ gets first $w$ examples in $S_1$
 4: for $t = 1, 2, \ldots$ do
 5:     receive new example $\mathbf{z}_t = (\mathbf{x}_t, \mathbf{y}_t)$
 6:     for every concept $i$ in the machine's hierarchy do
 7:         add $\mathbf{z}_t^i$ to $W^i_{\text{curr}}$
 8:     if $\exists i$ : MMD value between $W^i_{\text{curr}}$ and $W^i_{\text{past}} \geq \tau$ then
 9:         present the detected KD to the user
10:         receive a description of KD
11:         adapt predictor, hierarchy and windows
```

Method. To tackle knowledge drift, we present TRCKD and the pseudo-code of the approach is shown in Algorithm 2. The underlying classifier is a multi-label kNN [132], however the method can be adapted to more complex models. The classifier is trained on an initial data set S_1 that is compatible with a given hierarchy. At each iteration $t = 1, 2, \ldots$, the machine receives a new example $\mathbf{z}_t$. The method performs three steps: automated detection, interactive disambiguation and adaptation of the classifier and machine's hierarchy.

Detection. TRCKD keeps two windows of examples for each concept i in the machine hierarchy: W^i_{curr} holds the w most recent examples and is updated at each iteration, W^i_{past} contains w older examples and is updated only after a drift affects the ith concept. Each concept i is predicted by kNN only using the examples in W^i_{curr}. Knowledge drift is detected if the examples in W^i_{curr} and W^i_{past} are drawn from different distributions. The difference is measured by the *maximum mean discrepancy* (MMD) [54]. Given $\mathbf{a}$ and $\mathbf{a}'$ two vector of samples drawn i.i.d. from the distribution P, and $\mathbf{b}$ and $\mathbf{b}'$ from Q, the MMD between P and Q is defined as:

$$\text{MMD}(P, Q) = \sqrt{\mathbb{E}[k(\mathbf{a}, \mathbf{a}')] - 2\mathbb{E}[k(\mathbf{a}, \mathbf{b})] + \mathbb{E}[k(\mathbf{b}, \mathbf{b}')]}.$$

If P and Q are the same, then $\text{MMD}(P, Q) = 0$. The kernel k over examples is application-specific. The kernel defined in TRCKD is composed of two parts using the tensor product: k_x over the instances and k_y over the labels. Hence, the kernel is defined as:

$$k(\mathbf{z}, \mathbf{z}') = k((\mathbf{x}, y), (\mathbf{x}', y')) = k_X(\mathbf{x}, \mathbf{x}') \cdot k_Y(y, y')$$

A drift is detected if MMD between W^i_{past} and W^i_{curr} is grater than a threshold τ (line 8).

Disambiguation. After detecting a drift, TRCKD presents to the user a visualization of the drift by showing part of the hierarchy affected by the drift. The user can select and modify the *is-a* relations or the concepts that drifted.

Adaptation. Upon receiving the user description, TRCKD applies a simple but effective knowledge-aware adaptation strategy to adapt the windows and the hierarchy accordingly. In case of concept drift on the ith concept, the content of W^i_{curr} replaces the content in W^i_{past} and $W^i_{\text{curr}} \leftarrow \varnothing$. The two windows are deleted if the concept is removed. For relation addition, the examples belonging to the child concept are copied to the ancestors' current windows. In this way, every time the child concept is predicted, the ancestors' concepts are also predicted. Conversely, for relation removal, the examples belonging to the child concepts are removed from the parent's window and the child concept is linked directly to its grand-parent.

5.3.2 Experiments. We investigated the following research questions:

Q1 Is knowledge-aware adaptation useful for handling knowledge drift?
Q2 Does interaction with an expert help adaptation?
Q3 Does TRCKD work in realistic, multi-drift settings?

We compared TRCKD agaist several competitors:

- **PAW-*k*NN**: punitive adaptive window *k*NN, a multi-label approach that has a single sliding window for all concepts and discards the examples that contribute to the prediction errors [120];
- **MW-*k*NN**: multi-windows *k*NN approach for multi-label classification [132];
- **TRCKD** $_{oracle}$: TRCKD that knows exactly when and which kind of drift occured;
- ***k*NN 1-window**: *k*NN with a single sliding window for all concepts that keeps the w most recent examples and adapt passively to drift;
- ***k*NN**: *k*NN with no detection and adaptation.

We ran the experiments on three data test. HSTAGGER is the hierarchical version of STAGGER, a synthetic data set with three categorical attributes describing geometric shapes (shape, color, size) [126]. The instances are labeled by drifting random formulas like "big and (green or red)". The hierarchy is created by selecting two concepts a part of a third one that acts as parent concept. EMNIST is a data set contains 28×28 images representing handwritten digits and letters. The hierarchy is created by grouping characters in higher level concepts like consonant and even number. 20NG contains newsgroups posts classified in twenty categories. The categories where grouped in super-topic like religion and politics. A random sequence of examples is sampled from the data sets to generate a stream.

Q1. We compared our knowledge-aware adaptation strategy to standard forgetting and passive adaptation. We introduced **TRCKD** $_{forget}$, TRCKD that knows when a drift occured but not the kind, and thus it adapts by forgetting all examples regardless of the type of drift. To removed unrelated effect due to wrong or delayed detections, we told all methods when KD occurs. The results shows that TRCKD$_{oracle}$ outperforms the competitor on all data sets and all kind of drift (the full plots are reported in [28]). The knowledge-aware adaptation is by far the best strategy compared to passive adaptation like MW-*k*NN and *k*NN 1-window, to standard forgetting and to PAW-*k*NN, which discards the examples contributing to the prediction errors.

Q2. To evaluate the impact of the interaction, we compared TRCKD against variants that retrieve different information from the expert supervisor. The results show that TRCKD $_{oracle}$ outperforms all alternatives in all cases except one, showing that interactive disambiguation is useful to guide the knowledge-aware adaptation. TRCKD performs better than the no-interaction strategies. In particular, it perform better than the variant that uses MMD for drift detection and the likelihood ratio test for disambiguation, and the variant that adopts a forgetting strategy on the concepts detected by MMD.

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

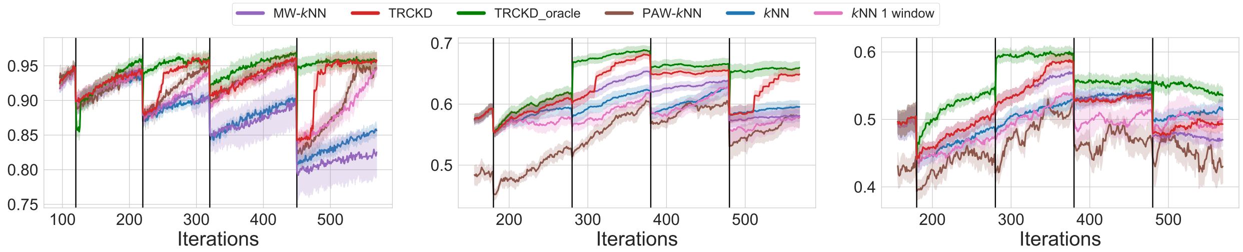

Fig. 7. Comparison of TRCKD versus competitors in terms of micro F_1 on multi-drift setting. The standard error is reported as a shaded area. **From left to right**: HSTAGGER, EMNIST and 20NG.

Q3. This final experiment aims to evaluate TRCKD in a realistic and more complex scenario in which multiple drifts occur in sequence, namely concept drift, relation addition, relation removal and concept removal. Figure 7 shows the results on the three data sets. TRCKD tends to outperforms all alternatives except the oracle. The plots confirms the advantages highlighted in the previous experiments and validates the benefit of knowledge-aware adaptation and interaction. The competitor PAW-*k*NN and MW-*k*NN lag behind due the limited reactivity to drifts.

5.3.3 Conclusion. In this section, we introduced knowledge drift, a problem that occurs in hierarchical classification and we presented TRCKD, an approach that tackles KD by combining automated drift detection and knowledge-aware adaptation with interactive disambiguation of the kind of drift. The empirical experiments validate the benefit of TRCKD in improving the predictive performance of the classifier Future works will focus on integrating knowledge-aware active learning because in practice the labels are not always available to the machine and needs to be acquired. Another interesting direction is to extend TRCKD to models other than *k*NN like neural networks.

6 FIRST ANALYSIS OF WENET PILOTS IN THE UK, DENMARK, MONGOLIA, AND PARAGUAY.

The first WeNet diversity pilot was conducted from November to December 2020 in five countries during the COVID pandemic. The data collection was organized in two phases. The first involved a large sample of university students from five universities, located in Denmark, Italy, Mongolia, Paraguay and the United Kingdom. The respondents had to fill a survey aimed at investigating their social practices and specific socio-demographic, cultural and psychological elements. In the second phase, a sub-sample of the respondents participated to a four-weeks data collection in which they were asked to fill in a self-reported time diary. This was done via a smartphone application, called iLog, which was also collecting data from thirty-four smartphone sensors, twenty-four hours a day. This dataset allows to investigate the diversity and daily routines of university students in a multi-layered perspective, both within and across countries, in a synchronic and diachronic way. Considering WP2's tasks, in this report, we only focus on the smartphone sensing data and time diary responses that were collected using iLog. The experiments were conducted during November and December of 2020. After data cleaning and processing, i-Log data was made available to the WeNet partners on April 30, 2021. However, due to technical issues, data from the Italian pilot were not made available. Hence, the preliminary analysis in this section is focused on the other four countries. We will include a detailed analysis in the final deliverable of WP2 (D2.3).

Time Use Diaries (TUDs) are meant to gather fine-grain data on how individuals spend their time. TUDs allow to measure the frequency and duration of human activities, behaviors and experiences offering a detailed view of social behavior. In a diary study, data are self-reported activity sequences in time episodes that can range from a few days to even a month or longer with a regular time interval. This type of data is usually collected via a self-completed time diary [43] that allows registering (at fixed time intervals) the sequence of an individual's activities. For each main activity in each interval, additional information is usually recorded, for instance about "where" and "with whom" this activity was done. So, in this section, we provide a preliminary analysis of time diaries, with some basic trends and an analysis of locations and routines, under three sections.

6.1 Trends from Time Diaries

For the iLog experiment, there were 379 participants from 5 countries including Denmark (27), United Kingdom (86), Mongolia (224), and Paraguay (42). The mean number of person days of data contribution (total of the number of days contributed by each person divided by the total number of users who contributed data) in the time diary is 16.92 (standard deviation = 10.79, median = 17, minimum = 1, maximum = 30). The mean compliance rate of time diaries calculated using compliance rates of all participants was 46.9% (standard deviation = 38.24%, minimum = 0.08%, maximum = 100%, median = 40.76%). To calculate the compliance rate, we divided the total number of time diaries in which all questions were answered (178361) and the total number of notifications sent to users (378761).

Figure 8 and Figure 9 show the distribution of completed time diaries for each hour of the day for weekdays and weekends, respectively. Because the time diary notifications were sent uniformly throughout the day, ideally the distribution should be flat. However, the distribution for both weekdays and weekends show that participants responded to time diaries more often in the morning hours, suggesting that there is a higher tendency of responding after waking up. During the afternoon and

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

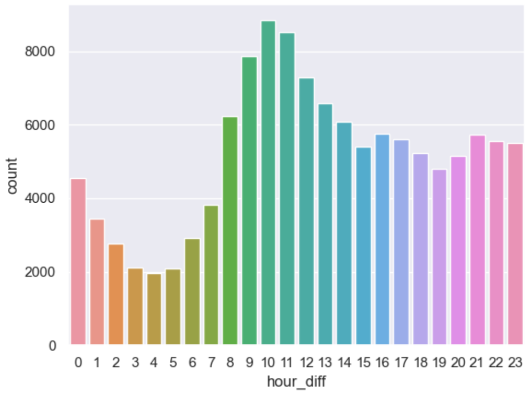

Fig. 8. Time Diary Completion Distribution Over Time of Day (hourly) on Weekdays.

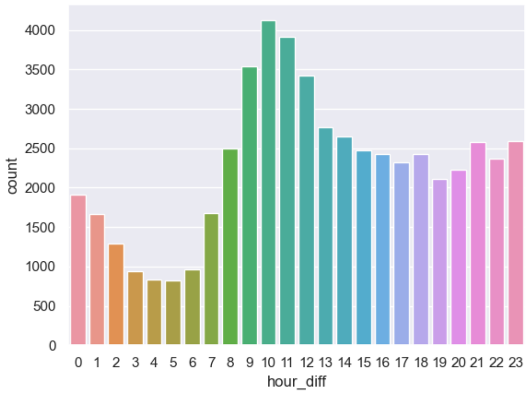

Fig. 9. Time Diary Completion Distribution Over Time of Day (hourly) on Weekends.

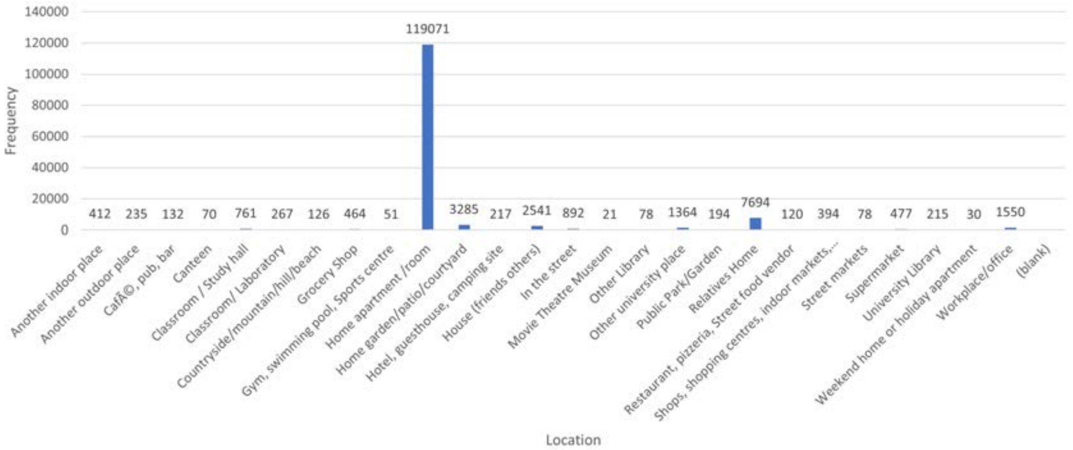

Fig. 10. Distribution of Locations in Self-Reports

evening, response rates are even. As expected, there is a dip in the number of responses after midnight. However, surprisingly, even during the early morning hours, there are a considerable number of responses for both weekdays and weekends. This could be because the participant cohort, who was composed of university students, was studying studying during these hours. Further, Figure 10 and Figure 11 show distributions of locations and activities as reported in self-reports.

6.2 Basic Locations and Routines

The dataset contained sensor logs including the records of the location coordinates of participants. By combining location coordinates and time diary responses about locations, we generated routine profiles of users. The steps to build the routines are:

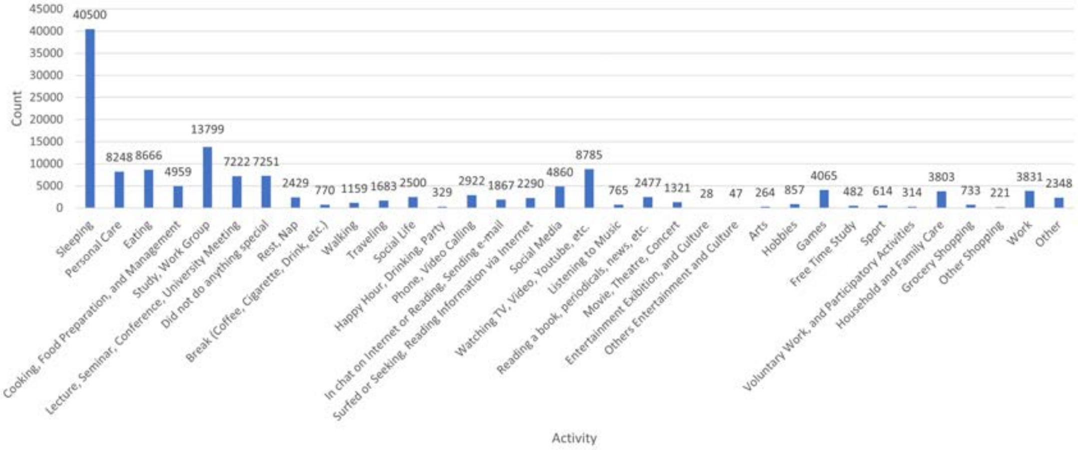

Fig. 11. Distribution of Activities in Self-Reports

- Extract the place of interests by using the same methodology from [113] (extract stay points and stay regions).
- Extract labels from the time diaries as ground truth.
- Use the geo-localization data to match with the answers from time dairies to attach latitude and longitude to each labels.
- Label the regions by combining localized time diaries with inferred regions.
- Use the geo-localization data through time to map the users behaviors with various time-slots.
- Regroup the time-based view of the routine by weekday; aggregate as the distribution of the labels from user's history.

When going through the above steps, data points from weekends were removed because those data points do not often reflect routine behaviors. Further, any user who did not even reported once were also removed for this analysis. Moreover, bar plots were generated by grouping by labels, and then summing the aggregated values to finally compute the percentage. The results are shown in Figure 12, Figure 13, Figure 14, and Figure 15.

The routines computed for each country show some differences in behaviors. The reports from the UK and Mongolia were done more often at home compared to Paraguay and Denmark. Figure 10 shows a distribution of locations in time diaries. Close to two thirds of self-reports were from home. The second largest category was relative's home. This is understandable since a majority of the people were working from home during the time period of data collection due to the COVID situation. Further, if we consider per user statistics, on average, a user reported from 4.97 locations (standard deviation = 3.41, minimum = 1, maximum = 17, median 4). If we look into this number further, the average number of locations reported by a person per day (only considering days on which at-least one location was reported) is 1.57 (median = 1, minimum = 1, maximum = 8, standard deviation = 0.94).

6.3 Descriptive Statistics related to Places in Time Diary and Stay Points/Regions

We observed that significant differences across countries can be inferred from the data. The differences between Paraguay and Mongolia are generally the highest. In table 9, the median number of visited

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

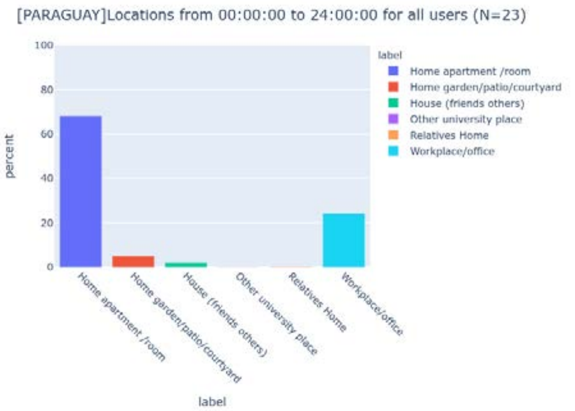

Fig. 12. Daily routines for Paraguay

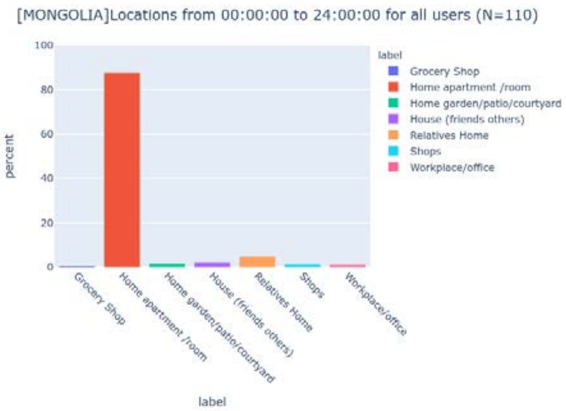

Fig. 13. Daily routines for Mongolia

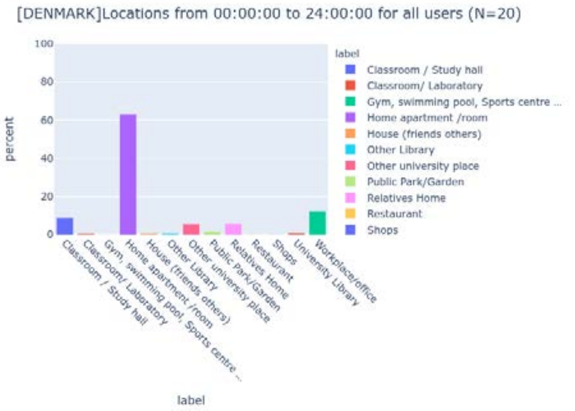

Fig. 14. Daily routines for Denmark

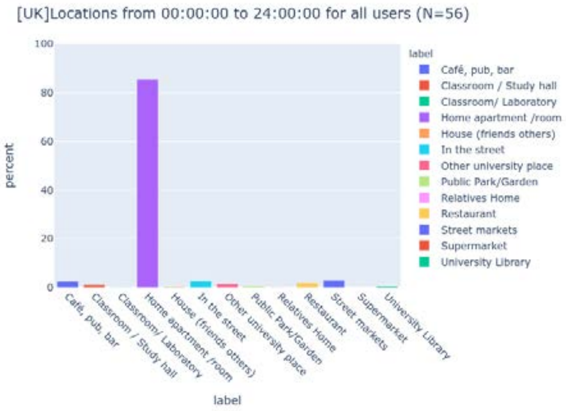

Fig. 15. Daily routines for United Kingdom

places in Paraguay is the highest (2.44). The lowest median is for Mongolia at only 1.28. Further, the results from table 10 and table 11 show that Mongolia users where significantly less mobile than the others. In regard to the median in table 10, the users from Paraguay were substantially more mobile than the ones from Mongolia. Surprisingly in table 14, it is Denmark that spent the least time at home. Except for the first quartile in which the users from Paraguay where almost not at home. Further, the median in table 13 highlight that there are missing data for half of the day for Denmark and Mongolia, and a third of the day for UK and Paraguay. This highlights a fundamental issue: mobile data in everyday life is noisy and can be missing, and this needs to be taken into consideration for further data analysis and algorithmic development and validation. Finally, Tables 13 and 15 show that it was hard to detect the locations of users from Denmark.

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

Table 9. Number of visited places (by day)

Country	nb_users	min	max	q1	median	q3	std
United Kingdom	57	1.00	4.00	1.50	2.00	2.41	0.74
Denmark	20	1.00	3.46	1.57	2.14	2.41	0.67
Mongolia	111	1.00	3.50	1.09	1.28	1.50	0.40
Paraguay	23	1.00	3.66	1.65	2.44	2.66	0.72

Table 10. Mean distance from home (in meters)

Country	nb_users	min	max	q1	median	q3	std
United Kingdom	53	1.97	3683514.85	237.13	834.09	3628.12	551739.57
Denmark	18	9.08	17425.11	670.97	1456.74	4456.56	4166.48
Mongolia	107	0.86	92188.65	15.19	54.81	531.89	17389.41
Paraguay	22	19.59	29706.87	1100.38	1995.05	3349.72	7395.93

Table 11. Max distance from home (in meters)

Country	nb_users	min	max	q1	median	q3	std
United Kingdom	53	7.91	11681537.17	4657.82	8846.46	16844.31	2125297.74
Denmark	18	133.58	180627.67	6100.24	7273.58	13701.59	39959.57
Mongolia	107	0.86	1132676.53	315.72	2712.03	8525.70	222782.27
Paraguay	22	4086.84	301753.01	8941.59	14016.90	26670.88	62469.01

Table 12. Number of hours spent at home (by day)

Country	nb_users	min	max	q1	median	q3	std
United Kingdom	57	0.00	22.50	3.40	11.55	16.46	6.97
Denmark	20	0.00	17.50	1.04	4.25	10.90	5.99
Mongolia	111	0.00	22.12	2.67	10.65	15.78	6.64
Paraguay	23	0.00	19.37	0.19	9.90	15.32	7.32

Table 13. Number of hours spent without geo-location data (by day)

Country	nb_users	min	max	q1	median	q3	std
United Kingdom	57	0.87	23.00	5.50	8.85	14.17	6.15
Denmark	20	3.02	23.02	6.68	11.70	17.35	6.32
Mongolia	111	1.80	23.25	7.10	11.30	17.37	5.91
Paraguay	23	1.16	21.32	5.02	8.00	11.62	5.60

Table 14. Number of hours spent at unknown locations (by day). Unknown locations are locations that are recorded but we are unable to link them with know regions infered from the time diaries

Country	nb_users	min	max	q1	median	q3	std
United Kingdom	57	0.00	7.84	0.50	0.97	1.67	1.74
Denmark	20	0.18	9.50	0.75	1.35	1.97	2.55
Mongolia	111	0.00	4.00	0.00	0.15	0.38	0.64
Paraguay	23	0.05	5.62	0.65	1.44	2.36	1.50

Table 15. Number of hours spent in a region that we do not know about (by day). Unknown regions are regions that we inferred from the behavior of the users, but they are not linked to any labels from the time diaries.

Country	nb_users	min	max	q1	median	q3	std
United Kingdom	57	0.00	6.57	0.00	0.14	0.40	1.24
Denmark	20	0.00	8.25	0.05	0.23	0.47	2.38
Mongolia	111	0.00	2.35	0.00	0.00	0.06	0.35
Paraguay	23	0.00	4.33	0.07	0.26	0.51	1.20

Future work. This preliminary analysis demonstrates that the WeNet pilot iLog data is both promising and challenging. In the next period, we first plan to conduct a comprehensive study that will use the country of residence as a diversity indicator. Based on mobile sensing features and machine learning models, we will address a number of research questions in both country-specific and country-agnostic settings:

(1) Food consumption, thus extending our previous work in WP2 [91, 93].
(2) Everyday life activities.
(3) Data quality and privacy aspects [92], linking our work with the work in WP9 [125].

In the second place, we will use the diversity analysis from WP1 to deepen the understanding of the multi-site pilots and the development of diversity-aware technology, by:

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

(1) Studying whether WeNet diversity indicators are reflected into observable behavioral differences (e.g. everyday activities.)
(2) Building and validating diversity-aware inference models using WeNet diversity indicators.
(3) Conducting bias analyses to identify potential issues with such models.

7 CONCLUSION

This deliverable describes individual learning methods developed in WeNet. The work included an analysis about identifying food consumption behaviors using mobile sensing and machine learning; inferring the social context of eating episodes using mobile sensing; privacy protection of mobile food diaries; handling human annotator mistakes and knowledge drift; and the initial analysis of multi-site WeNet mobile data, which represents the WeNet diversity pilot dataset collected in the project.

The work in WP2 has progressed at a reasonable pace, even though the progress was hindered to an extent due to the non-availability of data on time due to reasons such as COVID19 restrictions and other reasons. One key objective for the next period will be the application of these methodologies to the datasets to be collected in year 3 of the project, 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 project scenarios.

REFERENCES

[1] 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
[2] 2018. *The Guardian view on big data and insurance: knowing too much.* Retrieved Feb 13, 2020 from https://www.theguardian.com/commentisfree/2018/sep/27/the-guardian-view-on-big-data-and-insurance-knowing-too-much
[3] 2018. *Health Insurers Are Vacuuming Up Details About You — And It Could Raise Your Rates.* Retrieved Feb 13, 2020 from https://www.npr.org/sections/health-shots/2018/07/17/629441555/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates?t=1581596827223
[4] 2019. *FitBit Privacy Policy.* Retrieved Nov 11, 2019 from https://www.fitbit.com/eu/legal/privacy-policy
[5] 2019. *MyFitnessPal.* Retrieved April 28, 2020 from https://www.myfitnesspal.com/
[6] 2019. *OUT OF CONTROL: How consumers are exploited by the online advertising industry.* Retrieved Nov 11, 2020 from https://www.forbrukerradet.no/undersokelse/no-undersokelsekategori/report-out-of-control/
[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] 2020. *seaborn violinplot.* Retrieved May 28, 2020 from https://seaborn.pydata.org/generated/seaborn.violinplot.html
[12] 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.
[13] Nabil Alshurafa, Jayalakshmi Jain, Rawan Alharbi, Gleb Iakovlev, Bonnie Spring, and Angela Pfammatter. 2018. Is More Always Better? Discovering Incentivized MHealth Intervention Engagement Related to Health Behavior Trends. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 153 (Dec. 2018), 26 pages.

[14] Martin Atzmueller and Katy Hilgenberg. 2013. Towards Capturing Social Interactions with SDCF: An Extensible Framework for Mobile Sensing and Ubiquitous Data Collection. In *Proceedings of the 4th International Workshop on Modeling Social Media* (Paris, France) *(MSM '13)*. Association for Computing Machinery, New York, NY, USA, Article 6, 4 pages.

[15] 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.

[16] Gianni Barlacchi, Christos Perentis, Abhinav Mehrotra, Mirco Musolesi, and Bruno Lepri. 2017. Are you getting sick? Predicting influenza-like symptoms using human mobility behaviors. *EPJ Data Science* 6 (12 2017), 27.

[17] 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

[18] Adrian Bauman, Guansheng Ma, Frances Cuevas, Zainal Omar, Temo Waqanivalu, Philayrath Phongsavan, Kieren Keke, and Anjana Bhushan. 2011. Cross-national comparisons of socioeconomic differences in the prevalence of leisure-time and occupational physical activity, and active commuting in six Asia-Pacific countries. 65, 1 (2011), 35–43.

[19] Akram Bayat, Marc Pomplun, and Duc A. Tran. 2014. A Study on Human Activity Recognition Using Accelerometer Data from Smartphones. *Procedia Computer Science* 34 (2014), 450 – 457. The 9th International Conference on Future Networks and Communications (FNC'14)/The 11th International Conference on Mobile Systems and Pervasive Computing (MobiSPC'14)/Affiliated Workshops.

[20] 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.

[21] Ethan M. Berke, Tanzeem Choudhury, Shahid Ali, and Mashfiqui Rabbi. 2011. Objective Measurement of Sociability and Activity: Mobile Sensing in the Community. *The Annals of Family Medicine* 9, 4 (2011), 344–350.

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

[23] 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

[24] Joan-Isaac Biel, Nathalie Martin, David Labbe, and Daniel Gatica-Perez. 2018. Bites'N'Bits: Inferring Eating Behavior from Contextual Mobile Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 125 (Jan. 2018), 33 pages.

[25] 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.

[26] Peggy Bongers, Anita Jansen, Remco Havermans, Anne Roefs, and Chantal Nederkoorn. 2013. Happy eating. The underestimated role of overeating in a positive mood. *Appetite* 67 (2013), 74 – 80.

[27] Joseph Bonneau and Sören Preibusch. 2010. The Privacy Jungle:On the Market for Data Protection in Social Networks. In *Economics of Information Security and Privacy*, Tyler Moore, David Pym, and Christos Ioannidis (Eds.). Springer US, Boston, MA, 121–167.

[28] Andrea Bontempelli, Fausto Giunchiglia, Andrea Passerini, and Stefano Teso. 2021. Human-in-the-loop Handling of Knowledge Drift. *arXiv preprint arXiv:2103.14874* (2021).

[29] Andrea Bontempelli, Stefano Teso, Fausto Giunchiglia, and Andrea Passerini. 2020. Learning in the Wild with Incremental Skeptical Gaussian Processes. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20.*

[30] 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

[31] Hilde Bruch. 1964. Psychological Aspects of Overeating And Obesity. *Psychosomatics* 5, 5 (1964), 269 – 274.

[32] E. A. Carroll, M. Czerwinski, A. Roseway, A. Kapoor, P. Johns, K. Rowan, and M. C. Schraefel. 2013. Food and Mood: Just-in-Time Support for Emotional Eating. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction.* 252–257.

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

[33] Rich Caruana. 1997. Multitask Learning. *Machine Learning* 28, 1 (01 Jul 1997), 41–75. https://doi.org/10.1023/A:1007379606734

[34] 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.

[35] J Chua, Stephen Touyz, and AJ Hill. 2004. Negative mood-induced overeating in obese binge eaters: An experimental study. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity* 28 (05 2004), 606–10.

[36] Gillian Cleary. 2018. *Mobile Privacy: What Do Your Apps Know About You?* Retrieved Nov 06, 2019 from https://www.symantec.com/blogs/threat-intelligence/mobile-privacy-apps

[37] Rob Copeland. 2019. *Google's 'Project Nightingale' Gathers Personal Health Data on Millions of Americans.* Retrieved Jan 21, 2020 from https://www.wsj.com/articles/google-s-secret-project-nightingale-gathers-personal-health-data-on-millions-of-americans-11573496790

[38] 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. Social Influences on Eating.

[39] B. V. Dasarathy. 1997. Sensor fusion potential exploitation-innovative architectures and illustrative applications. *Proc. IEEE* 85, 1 (1997), 24–38.

[40] John B.F. de Wit, F. Marijn Stok, Derek J. Smolenski, Denise D.T. de Ridder, Emely de Vet, Tania Gaspar, Fiona Johnson, Lyliya Nureeva, and Aleksandra Luszczynska. 2015. Food Culture in the Home Environment: Family Meal Practices and Values Can Support Healthy Eating and Self-Regulation in Young People in Four European Countries. *Applied Psychology: Health and Well-Being* 7, 1 (2015), 22–40.

[41] Tamara Denning, Adrienne Andrew, Rohit Chaudhri, Carl Hartung, Jonathan Lester, Gaetano Borriello, and Glen Duncan. 2009. BALANCE: Towards a Usable Pervasive Wellness Application with Accurate Activity Inference. In *Proceedings of the 10th Workshop on Mobile Computing Systems and Applications* (Santa Cruz, California) *(HotMobile '09)*. Association for Computing Machinery, New York, NY, USA, Article 5, 6 pages.

[42] Robin Ian MacDonald Dunbar. 2017. Breaking Bread: the Functions of Social Eating. *Adaptive Human Behavior and Physiology* 3 (2017), 198 – 211.

[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] 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

[45] Alison E. Field, C. Barr Taylor, Angela Celio, and Graham A. Colditz. 2004. Comparison of self-report to interview assessment of bulimic behaviors among preadolescent and adolescent girls and boys. *International Journal of Eating Disorders* 35, 1 (2004), 86–92.

[46] Benoît Frénay and Michel Verleysen. 2014. Classification in the presence of label noise: a survey. *IEEE Trans. Neural Netw. Learn. Syst* (2014).

[47] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM Comput Surv* (2014).

[48] Purificacian Garcia-Segovia, Robert J. Harrington, and Han-Seok Seo. 2015. Influences of table setting and eating location on food acceptance and intake. *Food Quality and Preference* 39 (2015), 1 – 7.

[49] 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*.

[50] Luke Gemming, Aiden Doherty, Jennifer Utter, Emma Shields, and Cliona Ni Mhurchu. 2015. The use of a wearable camera to capture and categorise the environmental and social context of self-identified eating episodes. *Appetite* 92 (2015), 118 – 125.

[51] Fausto Giunchiglia, Enrico Bignotti, and Mattia Zeni. 2017. Personal context modelling and annotation. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*.

[52] Holly C. Gooding, Carly Milliren, Christina M. Shay, Tracy K. Richmond, Alison E. Field, and Matthew W. Gillman. 2016. Achieving Cardiovascular Health in Young Adulthoodâ€"Which Adolescent Factors Matter? *Journal of Adolescent*

Health 58, 1 (2016), 119 – 121.

[53] Greenland Sander, Senn Stephen J., Rothman Kenneth J., Carlin John B., Poole Charles, Goodman Steven N., and Altman Douglas G. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* 31, 4 (2016), 337–350.

[54] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alex Smola. 2012. A kernel two-sample test. *JMLR* (2012).

[55] Regina Guthold, Melanie J. Cowan, Christine S. Autenrieth, Laura Kann, and Leanne M. Riley. 2010. Physical Activity and Sedentary Behavior Among Schoolchildren: AÂ 34-Country Comparison. *The Journal of Pediatrics* 157, 1 (2010), 43 – 49.e1.

[56] Juliet Haarman, Roelof de Vries, Emiel Harmsen, Hermie Hermens, and Dirk Heylen. 2020. Sensory Interactive Table (SIT) — Development of a Measurement Instrument to Support Healthy Eating in a Social Dining Setting. *Sensors* 20 (05 2020), 2636.

[57] Gabriella Harari, Sandrine Mueller, Clemens Stachl, Rui Wang, Weichen Wang, Markus Buehner, Peter Rentfrow, Andrew Campbell, and Samuel Gosling. 2019. Sensing Sociability: Individual Differences in Young Adults' Conversation, Calling, Texting, and App Use Behaviors in Daily Life. *Journal of Personality and Social Psychology* 119 (05 2019).

[58] Gabriella M. Harari, Samuel D. Gosling, Rui Wang, Fanglin Chen, Zhenyu Chen, and Andrew T. Campbell. 2017. Patterns of behavior change in students over an academic term: A preliminary study of activity and sociability behaviors using smartphone sensing methods. *Computers in Human Behavior* 67 (2017), 129 – 138.

[59] Samuli Hemminki, Petteri Nurmi, and Sasu Tarkoma. 2014. Gravity and Linear Acceleration Estimation on Mobile Devices. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services* (London, United Kingdom) *(MOBIQUITOUS '14)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL, 50–59.

[60] C.P. Herman, J. Polivy, and T. Leone. 2005. 6 - The psychology of overeating. In *Food, Diet and Obesity*, David J. Mela (Ed.). Woodhead Publishing, 115 – 136.

[61] C. P. Herman and J. Polivy. 2003. Dieting as an exercise in behavioral economics. *Time and decision: Economic and psychological perspectives on intertemporal choice* (2003).

[62] Marion M. Hetherington. 2007. Cues to overeat: psychological factors influencing overconsumption. *Proceedings of the Nutrition Society* 66, 1 (2007), 113–123.

[63] Marion M. Hetherington, Annie S. Anderson, Geraldine N.M. Norton, and Lisa Newson. 2006. Situational effects on meal intake: A comparison of eating alone and eating with others. *Physiology |& Behavior* 88, 4 (2006), 498–505.

[64] Suzanne Higgs and Jason Thomas. 2016. Social influences on eating. *Current Opinion in Behavioral Sciences* 9 (2016), 1 – 6. Diet, behavior and brain function.

[65] Lena Holzer. 2018. *Reporton third gender markeror no gender marker options.* Retrieved Nov 11, 2019 from https://www.ilga-europe.org/sites/default/files/non-binary_gender_registration_models_in_europe_0.pdf

[66] Ya-Li Huang, Won O. Song, Rachel A. Schemmel, and Sharon M. Hoerr. 1994. What do college students eat? Food selection and meal pattern. *Nutrition Research* 14, 8 (1994), 1143 – 1153.

[67] 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

[68] Margaret M. Jastran, Carole A. Bisogni, Jeffery Sobal, Christine Blake, and Carol M. Devine. 2009. Eating routines. Embedded, value based, modifiable, and reflective. *Appetite* 52, 1 (2009), 127 – 136.

[69] Jisu Jung, Lyndal Wellard-Cole, Colin Cai, Irena Koprinska, Kalina Yacef, Margaret Allman-Farinelli, and Judy Kay. 2020. Foundations for Systematic Evaluation and Benchmarking of a Mobile Food Logger in a Large-Scale Nutrition Study. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 47 (June 2020), 25 pages.

[70] Thivya Kandappu, Abhinav Mehrotra, Archan Misra, Mirco Musolesi, Shih-Fen Cheng, and Lakmal Meegahapola. 2020. PokeME: Applying Context-Driven Notifications to Increase Worker Engagement in Mobile Crowd-Sourcing. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (Vancouver BC, Canada) *(CHIIR '20)*. 3–12.

[71] Gregory S. Keenan, Louise Childs, Peter J. Rogers, Marion M. Hetherington, and Jeffrey M. Brunstrom. 2018. The portion size effect: Women demonstrate an awareness of eating more than intended when served larger than normal

portions. *Appetite* 126 (2018), 54 – 60.

[72] Tae Kim. 2015. T test as a parametric statistic. *Korean Journal of Anesthesiology* 68 (11 2015), 540.

[73] 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

[74] D. Kotz, C. A. Gunter, S. Kumar, and J. P. Weiner. 2016. Privacy and Security in Mobile Health: A Research Agenda. *Computer* 49, 6 (June 2016), 22–30. https://doi.org/10.1109/MC.2016.185

[75] Mark A. Kramer. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal* 37, 2 (1991), 233–243. https://doi.org/10.1002/aic.690370209 arXiv:https://aiche.onlinelibrary.wiley.com/doi/pdf/10.1002/aic.690370209

[76] D Kromhout, A Keys, C Aravanis, R Buzina, F Fidanza, S Giampaoli, A Jansen, A Menotti, S Nedeljkovic, and M Pekkarinen. 1989. Food consumption patterns in the 1960s in seven countries. *The American Journal of Clinical Nutrition* 49, 5 (05 1989), 889–894.

[77] Daniël Lakens. 2013. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 4: 863. *Frontiers in psychology* 4 (11 2013), 863.

[78] Dong Kyu Lee. 2016. Alternatives to P value: confidence interval and effect size. In *Korean journal of anesthesiology*.

[79] 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.

[80] Soo Ling Lim, Peter Bentley, Natalie Kanakam, Fuyuki Ishikawa, and Shinichi Honiden. 2014. Investigating Country Differences in Mobile App User Behavior and Challenges for Software Engineering. *IEEE Transactions on Software Engineering* 41 (09 2014).

[81] Emma V. Long, Lenny R. Vartanian, C. Peter Herman, and Janet Polivy. 2020. What does it mean to overeat? *Eating Behaviors* 37 (2020), 101390.

[82] 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.

[83] Alexander Lütz, Erik Rodner, and Joachim Denzler. 2013. I want to know moreâ€”efficient multi-class incremental learning using Gaussian processes. *Pattern recognition and image analysis* 23, 3 (2013), 402–407.

[84] 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

[85] 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

[86] 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

[87] Vijini Mallawaarachchi, Lakmal Meegahapola, Roshan Madhushanka, Eranga Heshan, Dulani Meedeniya, and Sampath Jayarathna. 2020. Change Detection and Notification of Web Pages: A Survey. *ACM Comput. Surv.* 53, 1, Article 15 (Feb. 2020), 35 pages. https://doi.org/10.1145/3369876

[88] 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

[89] Lakmal Meegahapola, Noel Athaide, Kasthuri Jayarajah, Shili Xiang, and Archan Misra. 2019. Inferring Accurate Bus Trajectories from Noisy Estimated Arrival Time Records. *CoRR* abs/1907.08483 (2019). arXiv:1907.08483 http://arxiv.org/abs/1907.08483

[90] L. Meegahapola and D. Gatica-Perez. 2021. Smartphone Sensing for the Well-Being of Young Adults: A Review. *IEEE Access* 9 (2021), 3374–3399.

[91] Lakmal Meegahapola, Salvador Ruiz-Correa, and Daniel Gatica-Perez. 2020. Alone or With Others? Understanding Eating Episodes of College Students with Mobile Sensing. In *19th International Conference on Mobile and Ubiquitous Multimedia* (Essen, Germany) *(MUM 2020)*. Association for Computing Machinery, New York, NY, USA, 162–166.

[92] Lakmal Meegahapola, Salvador Ruiz-Correa, and Daniel Gatica-Perez. 2020. Protecting Mobile Food Diaries from Getting too Personal. In *Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia* (Essen, Germany) *(MUM '20)*. Association for Computing Machinery, New York, NY, USA.

[93] Lakmal Meegahapola, Salvador Ruiz-Correa, Viridiana del Carmen Robledo-Valero, Emilio Ernesto Hernandez-Huerfano, Leonardo Alvarez-Rivera, Ronald Chenu-Abente, and Daniel Gatica-Perez. 2021. One More Bite? Inferring Food Consumption Level of College Students Using Smartphone Sensing and Self-Reports. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 26 (March 2021), 28 pages.

[94] David Mela. 2005. *Food, Diet, and Obesity.* Elsevier. https://www.elsevier.com/books/food-diet-and-obesity/mela/978-1-85573-958-1

[95] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. 2019. A Survey on Food Computing. *ACM Comput. Surv.* 52, 5, Article 92 (Sept. 2019), 36 pages.

[96] Mark Mirtchouk, Dana McGuire, Andrea Deierlein, and Samantha Kleinberg. 2019. Automated Estimation of Food Type from Body-worn Audio and Motion Sensors in Free-Living Environments. *Proceedings of machine learning research* 106 (08 2019), 641–662.

[97] Mobius. 2019. *11 surprising mobile health statistics.* Retrieved April 28, 2020 from https://www.mobius.md/blog/2019/03/11-mobile-health-statistics/

[98] Choon Boon Ng, Yong Haur Tay, and Bok-Min Goi. 2012. Recognizing Human Gender in Computer Vision: A Survey. In *PRICAI 2012: Trends in Artificial Intelligence*, Patricia Anthony, Mitsuru Ishizuka, and Dickson Lukose (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 335–346.

[99] 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

[100] University of Groningen. 2020. *Sensitive data and medical confidentiality.* Retrieved Jan 21, 2020 from https://www.futurelearn.com/courses/protecting-health-data/0/steps/39608

[101] Lukasz Olejnik, Tran Minh-Dung, and Claude Castelluccia. 2013. Selling Off Privacy at Auction. (Dec. 2013). https://hal.inria.fr/hal-00915249 working paper or preprint.

[102] Guy S. Parcel, Lana D. Muraskin, and Carolina M. Endert. 1988. Community education: Study group report. *Journal of Adolescent Health Care* 9, 6, Supplement (1988), S41 – S45.

[103] K.A. Patel and D.G. Schlundt. 2001. Impact of moods and social context on eating behavior. *Appetite* 36, 2 (2001), 111 – 118.

[104] Hannah Payne, Cameron Lister, Joshua West, and Jay Bernhardt. 2015. Behavioral Functionality of Mobile Apps in Health Interventions: A Systematic Review of the Literature. *JMIR mHealth and uHealth* 3 (02 2015), e20.

[105] Pekka and Antti Kouvo. 2007. LINKED OR DIVIDED BY THE WEB?: Internet use and sociability in four European countries. *Information, Communication & Society* 10, 2 (2007), 219–241.

[106] Iryna Pentina, Lixuan Zhang, Hatem Bata, and Ying Chen. 2016. Exploring privacy paradox in information-sensitive mobile app adoption: A cross-cultural comparison. *Computers in Human Behavior* 65 (2016), 409 – 419.

[107] Janet Polivy and C. Peter Herman. 2020. Overeating in Restrained and Unrestrained Eaters. *Frontiers in Nutrition* 7 (2020), 30.

[108] Janet Polivy, C. Peter Herman, and Rajbir Deo. 2010. Getting a bigger slice of the pie. Effects on eating and emotion in restrained and unrestrained eaters. *Appetite* 55, 3 (2010), 426 – 430.

[109] Mollie Powles. 2018. *Personalization Versus Privacy: Making Sense of the Privacy Paradox.* Retrieved Jan 21, 2020 from https://blog.hubspot.com/marketing/personalization-versus-privacy

[110] privacyinternational. 2019. *No Body's Business But Mine: How Menstruation Apps Are Sharing Your Data.* Retrieved Nov 08, 2019 from https://privacyinternational.org/long-read/3196/no-bodys-business-mine-how-menstruation-apps-are-sharing-your-data

[111] privacyrights. 2017. Mobile Health and Fitness Apps: What Are the Privacy Risks? https://pdfs.semanticscholar.org/c52c/67541657fd71022771edaed75148999b3c00.pdf

[112] V. M. Quick and C. Byrd-Bredbenner. 2013. Disturbed eating behaviours and associated psychographic characteristics of college students. *Journal of Human Nutrition and Dietetics* 26, s1 (2013), 53–63.

[113] J. Blom R. Montoliu and D. Gatica-Perez. 2013. Discovering Places of Interest in Everyday Life from Smartphone Data. *Multimedia Tools and Applications* (2013). https://doi.org/10.1007/s11042-011-0982-z

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

[114] 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.

[115] Tauhidur Rahman, Mary Czerwinski, Ran Gilad-Bachrach, and Paul Johns. 2016. Predicting "About-to-Eat" Moments for Just-in-Time Eating Intervention. In *Proceedings of the 6th International Conference on Digital Health Conference* (Montréal, Québec, Canada) *(DH '16)*. Association for Computing Machinery, New York, NY, USA, 141–150.

[116] Nairan Ramirez-Esparza, Matthias R. Mehl, Javier Alvarez-Bermudez, and James W. Pennebaker. 2009. Are Mexicans more or less sociable than Americans? Insights from a naturalistic observation study. *Journal of Research in Personality* 43, 1 (2009), 1 – 7.

[117] Marnie E. Rice and Grant T. Harris. 2005. Comparing Effect Sizes in Follow-Up Studies: ROC Area, Cohen's d, and r. *Law and Human Behavior* 29, 5 (01 Oct 2005), 615–620.

[118] Christopher Riederer, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, and Pablo Rodriguez. 2011. For Sale: Your Data: By: You. In *Proceedings of the 10th ACM Workshop on Hot Topics in Networks* (Cambridge, Massachusetts) *(HotNets-X)*. Association for Computing Machinery, New York, NY, USA, Article 13, 6 pages. https://doi.org/10.1145/2070562.2070575

[119] Natti Ronel and Galit Libman. 2003. Eating Disorders and Recovery: Lessons from Overeaters Anonymous. *Clinical Social Work Journal* 31 (06 2003), 155–171.

[120] Martha Roseberry, Bartosz Krawczyk, and Alberto Cano. 2019. Multi-label punitive kNN with self-adjusting memory for drifting data streams. *ACM Transactions on Knowledge Discovery from Data (TKDD)* (2019).

[121] Helen K. Ruddock and Charlotte A. Hardman. 2018. Guilty pleasures: The effect of perceived overeating on food addiction attributions and snack choice. *Appetite* 121 (2018), 9 – 17.

[122] Alan Russell, Craig Hart, Clyde Robinson, and Susanne Olsen. 2003. Children's sociable and aggressive behaviour with peers: A comparison of the US and Australia, and contributions of temperament and parenting styles. *International Journal of Behavioral Development* 27, 1 (2003), 74–86.

[123] Krushnapriya Sahoo, Bishnupriya Sahoo, Ashok Choudhury, Nighat Sofi, Raman Kumar, and Ajeet Bhadoria. 2015. Childhood obesity: causes and consequences. *Journal of Family Medicine and Primary Care* 4 (04 2015), 187–92. https://doi.org/10.4103/2249-4863.154628

[124] 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.

[125] Laura Schelenz, Ivano Bison, Matteo Busso, Amalia de Götzen, Daniel Gatica-Perez, Fausto Giunchiglia, Lakmal Buddika Meegahapola, and Salvador Ruiz-Correa. 2021. The Theory, Practice, and Ethical Challenges of Designing a Diversity-Aware Platform for Social Relations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 11. https://doi.org/10.1145/3461702.3462595

[126] Jeffrey C Schlimmer and Richard H Granger. 1986. Incremental learning from noisy data. *Machine learning* (1986).

[127] 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, CHE, 103–112.

[128] 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.

[129] R. Sharma, V. I. Pavlovic, and T. S. Huang. 1998. Toward multimodal human-computer interface. *Proc. IEEE* 86, 5 (1998), 853–869.

[130] Rachel Shelton, Lorna Mcneill, Elaine Puleo, Kathleen Wolin, Karen Emmons, and Gary Bennett. 2011. The Association Between Social Factors and Physical Activity Among Low-Income Adults Living in Public Housing. *American journal of public health* 101 (02 2011), 2102–10.

[131] Christine Sheppard-Sawyer, Richard McNally, and Jennifer Fischer. 2000. Film-induced sadness as a trigger for disinhibited eating. *The International journal of eating disorders* 28 (10 2000), 215–20.

[132] Eleftherios Spyromitros-Xioufis, Myra Spiliopoulou, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. Dealing with concept drift and class imbalance in multi-label stream classification. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

[133] Emma J Stinson, Susanne B. Votruba, Colleen A Venti, Marisol Perez, Jonathan Krakoff, and Marci E. Gluck. 2018. Food insecurity is associated with maladaptive eating behaviors and objectively measured overeating. *Obesity (Silver Spring, Md.)* 26 (2018), 1841 – 1848.

[134] J Graham Thomas, Sapna Doshi, Ross D. Crosby, and Michael R Lowe. 2011. Ecological momentary assessment of obesogenic eating behavior: combining person-specific and environmental predictors. *Obesity* 19 8 (2011), 1574–9.

[135] Edison Thomaz, Irfan Essa, and Gregory D. Abowd. 2015. A Practical Approach for Recognizing Eating Moments with Wrist-Mounted Inertial Sensing. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 1029–1040.

[136] Edison Thomaz, Irfan Essa, and Gregory D. Abowd. 2015. A Practical Approach for Recognizing Eating Moments with Wrist-Mounted Inertial Sensing. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. 1029–1040.

[137] 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.

[138] Roger Tourangeau, Lance J Rips, and Kenneth Rasinski. 2000. *The psychology of survey response.*

[139] Bengisu Tulu, Carolina Ruiz, Joshua Allard, Joseph Acheson, Andrew Busch, Andrew Roskusku, Gage Heeringa, Victor Jaskula, Jessica Oleski, and Sherry Pagoto. 2017. SlipBuddy: A Mobile Health Intervention to Prevent Overeating. (01 2017).

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

[141] Narseo Vallina-Rodriguez and Srikanth Sundaresan. 2019. *7 in 10 smartphone apps share your data with third-party services.* Retrieved Nov 08, 2019 from http://theconversation.com/7-in-10-smartphone-apps-share-your-data-with-third-party-services-72404

[142] 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

[143] Tatjana van Strien, C. Peter Herman, and Marieke W. Verheijden. 2009. Eating style, overeating, and overweight in a representative Dutch sample. Does external eating play a role? *Appetite* 52, 2 (2009), 380 – 387.

[144] Tatjana van Strien, C. Peter Herman, and Marieke W. Verheijden. 2012. Eating style, overeating and weight gain. A prospective 2-year follow-up study in a representative Dutch sample. *Appetite* 59, 3 (2012), 782 – 789.

[145] Lenny Vartanian, Natalie Reily, Samantha Spanos, C Herman, and Janet Polivy. 2017. Self-reported overeating and attributions for food intake. *Psychology & health* 32 (01 2017), 1–10.

[146] Lenny R. Vartanian, Natalie M. Reily, Samantha Spanos, Lucy C. McGuirk, C. Peter Herman, and Janet Polivy. 2017. Hunger, taste, and normative cues in predictions about food intake. *Appetite* 116 (2017), 511 – 517.

[147] Carine A Vereecken, Joanna Todd, Chris Roberts, Caroline Mulvihill, and Lea Maes. 2006. Television viewing behaviour and associations with food habits in different countries. *Public Health Nutrition* 9, 2 (2006), 244–250.

[148] Marco Vicente, Fernando Batista, and João Paulo Carvalho. 2015. Twitter gender classification using user unstructured information. *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (2015), 1–7.

[149] Marco Vicente, Fernando Batista, and Joao P. Carvalho. 2019. *Gender Detection of Twitter Users Based on Multiple Information Sources.* Springer International Publishing, Cham, 39–54. https://doi.org/10.1007/978-3-030-01632-6_3

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

[151] Birgitte Wammes, Boudewijn Breedveld, Stef Kremers, and Johannes Brug. 2006. The 'balance intervention' for promoting caloric compensatory behaviours in response to overeating: a formative evaluation. *Health Education Research* 21, 4 (04 2006), 527–537.

[152] Staphanie Ward, Mathieu Baelanger, Denise Donovan, and Natalie Carrier. 2016. Relationship between eating behaviors and physical activity of preschoolers and their peers: A systematic review. *International Journal of Behavioral Nutrition*

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

and Physical Activity 13 (12 2016).

[153] K.R. Westerterp. 2005. 4 - Physical activity and obesity. In *Food, Diet and Obesity*, David J. Mela (Ed.). Woodhead Publishing, 76 – 85.

[154] World Health Organization (WHO). 2020. *Obesity and overweight.* Retrieved April 28, 2020 from https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[155] Christopher KI Williams and Carl Edward Rasmussen. 2006. *Gaussian processes for machine learning.*

[156] Donald A. Williamson, David H. Gleaves, and Olga J. Lawson. 1991. Biased perception of overeating in bulimia nervosa and compulsive binge eaters. *Journal of Psychopathology and Behavioral Assessment* 13 (1991), 257–268.

[157] Richard J. Wurtman and Judith J. Wurtman. 1995. Brain Serotonin, Carbohydrate-Craving, Obesity and Depression. *Obesity Research* 3, S4 (1995), 477S–480S. https://doi.org/10.1002/j.1550-8528.1995.tb00215.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1550-8528.1995.tb00215.x

[158] Koji Yatani. 2016. *Effect Sizes and Power Analysis in HCI.* Springer International Publishing, Cham, 87–110.

[159] Luke Yates and Alan Warde. 2017. Eating together and eating alone: meal arrangements in British households. *The British Journal of Sociology* 68, 1 (2017), 97–118.

[160] Yang Yue, Tian Lan, Anthony G.O. Yeh, and Qing-Quan Li. 2014. Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. *Travel Behaviour and Society* 1, 2 (2014), 69 – 78.

[161] Tok Chen Yun, Siti Rohaiza Ahmad, and David Koh Soo Quee. 2018. Dietary Habits and Lifestyle Practices among University Students in Universiti Brunei Darussalam. *The Malaysian Journal of Medical Sciences : MJMS* 25 (2018), 56 – 66.

[162] 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.

[163] 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.*

[164] Mattia Zeni, Wanyi Zhang, Enrico Bignotti, Andrea Passerini, and Fausto Giunchiglia. 2019. Fixing mislabeling by human annotators leveraging conflict resolution and prior knowledge. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (2019).

[165] Lydia Zepeda and David Deal. 2008. Think before you eat: photographic food diaries as intervention tools to change dietary decision making and attitudes. *International Journal of Consumer Studies* 32, 6 (2008), 692–698.