SUPPLEMENTARY MATERIAL

Supplementary material has technical details about the following components of the work:

- Mobile app
- Data collection procedure
- Sensor measurements
- Extracted features
- Label processing
- Classification methods
- Performance evaluation
- User personalization assessment
- Detailed results tables

Mobile app

For the purpose of data collection in a large scale we developed a mobile application called *ExtraSensory*, with versions for both iPhone and Android smartphones, and a companion application for the Pebble smartwatch that integrates with both. The app was used for supervised data collection, meaning it collects both sensor measurements and ground truth context labels. The app is scheduled every minute to automatically record measurements for 20 seconds from the sensors. Sensors are sampled in frequencies appropriate for their domain, and include motion-responsive sensors, location services, audio, environment sensors, as well as bits of information about the phone's state. When the watch is available (within Bluetooth range and paired with the phone) measurements from the watch are also collected by the app during the 20 second recording session. More details about the sensors are provided in "Sensor measurements". At the end of the 20 second recording session the measurements are bundled in a zip file and sent through the internet (if a WiFi network is available) to our lab's server, which runs a quick calculation and replies with an initial prediction of the activity (*e.g.* sitting, walking, running). All communication between the app and the server is secure and encrypted, and identified only by a unique universal identifier (UUID) that was randomly generated for each user.

In addition to collecting sensor measurements, the app's interface provides several mechanisms for the user to report labels about their context. This was a crucial part of the research design and we had to overcome a basic trade-off: on one hand we wanted to collect ground truth labels for as many examples (minutes) as possible and with much detail (combination of all the relevant context labels). On the other hand we did not want the subject to interact with the app every minute to report labels, both because it would be an extreme inconvenience for the subject and because it would impact the natural behavior of the subject and miss the point of collecting data in-the-wild. To balance this trade-off, we designed a flexible interface that helps minimize the user-app interaction time. The following label-reporting mechanisms were included:

- A history journal presents the user's activities chronologically as a calendar and enables the user to easily edit the context labels of time ranges in the past (up to one day in the past). The user can easily merge a sequence of minutes to a single "event" with the same context labels, or split a calendar event to describe a change in context. See Figure 2 (A). The real-time predictions from the server assist the user to recall when their activity changed — consecutive minutes with the same prediction from the server are merged to a single item on the history calendar.
- The user can also initiate active feedback by reporting labels describing their context in the immediate future (starting "now" for up to half an hour in the future). See Figure 2 (B).
- Every x minutes (by default, x is 10 minutes, but can be set by the user) the app presents a notification to remind the user to provide labels. If the the user has recently provided labels, the notification asks whether the user was still engaged in the same activities — allowing for a quick and easy response if the answer is "yes". See Figure 2 (C).
- The notifications also appear on the smartwatch, allowing for an easier response with a click of a button on the watch, without using the phone itself.
- When selecting labels from the menu, a side-bar index allows quick search of the relevant labels, either by categories (*e.g.* sports, work, company) or through a "frequently used labels" menu, which presents labels that the user has applied in the past. The category in which a label was presented in the menu does not matter, and a label can appear under different categories (*e.g.* "skateboarding" appears under "sports", "leisure" and "transportation") — the only reason for these categories is to make it easy for the user to find the relevant label quickly.

Data collection procedure

The study's research plan and consent form were approved by the university's institutional review board (IRB). Human subjects were recruited for the study via fliers across campus, university mailing lists and word of mouth. Every subject read and signed the consent form. The researchers installed the app on each subject's personal phone (to maximize authenticity of natural behavior). The subject then engaged in their usual behaviors for approximately a week, while keeping the app running in the background on their phone as much as was possible and convenient. The subject was asked to report as many labels as possible without interfering too much with their natural behavior. Subjects varied in their level of rigorousness with respect to providing labels: some wanted to be very precise (with specific detailed combinations of labels, and trying to keep minute-to-minute precision) and others tended to be less specific and to dedicate less effort. The subjects who used the watch, which we supplied them with, were told that it is fine to get it wet (wash hands, shower) but not submerge it (swimming). They were also asked to turn off the watch app whenever they removed the watch from their wrist and to turn it back on when they wore the watch — so we can generally assume that whenever watch measurements are available they were taken from the subject's wrist.

Using the app consumes the phone's battery more quickly than normal. To make up for this, the researchers provided participants with an external portable battery, which provides one extra charge during the day. The researchers also provided the subject with the Pebble smartwatch (56 of the subjects agreed to use the watch). The external battery and the smartwatch were returned at the end of the study. Each subject was compensated for their participation. The basic compensation was in the amount of $40, and an additional amount was calculated based on the amount of labeled data that the subject contributed (as an incentive to encourage reporting many labels). The total compensation per subject was between $40 and $75.

Technical difficulties

During the development of the iPhone app, there were releases of new iOS versions that caused the app to not work well and required us to change the code.

Since subjects used their personal phones, the app had to handle different devices, and in the Android case, different makers. For some of the Android users when we installed the app we noticed it didn't work well. In three cases the workaround was to install a slightly different version of the app that didn't use the gyroscope. After installing the changed app and making sure it works those users began collecting records (without gyroscope measurements).

On top of the dataset's 60 subjects, there were four more subjects that participated and received the basic compensation, but whose data was not included in the dataset. For two of them the app didn't work well on their devices. The other two were too stressed or otherwise occupied during participation, and produced too little and un-reliable labels, so we decided to discard their data.

Sensor measurements

Raw sensor measurements are provided in the publicly available dataset.

High frequency measurements:

Each sensor (and pseudo-sensor) in the following list was sampled at 40Hz during the $\sim$20 second recording session to produce a time series of $\sim$800 time points. The sampling relies on the design of the phone's hardware and operating system and the sampling rate was not guaranteed to be accurate (especially for the Android devices). For that reason the time reference of each sample in a time series was also recorded; the differences between consecutive time references were approximately 25 milliseconds.

- Accelerometer. Time series of 3-axis vectors of acceleration according to standard axes of phone devices.
- Gyroscope. Time series of 3-axis vectors of rotation rate around each of the phone's 3 axes.
- Magnetometer. Time series of 3-axis vectors of the magnetic field.
- Processed signals. Both iPhone and Android operating systems provide processed versions of the signals: The raw acceleration is split to the gravity acceleration (estimated direction of gravity at every moment, the magnitude is always 1G) and the user-generated acceleration (subtraction of the gravity signal from the raw acceleration). For the gyroscope the OS has a calibrated version that attempts to remove drift effects. For the magnetometer the OS has an unbiased version that subtracts the estimated bias of the magnetic field created by the device itself.

In this paper we used the raw acceleration signal (which includes the effects of gravity) and the calibrated version of the gyroscope signal. Acceleration is reported in units of G (gravitational acceleration on the surface of the Earth) on iPhone and in units of m/s^2 on Android. Before extracting features we converted the Android acceleration measurements to units of G.

Watch measurements: From the Pebble smartwatch we collected signals from the two available sensors—accelerometer and compass. Acceleration was sampled at 25Hz and describes a 3-axis vector of acceleration (in units of mG) relative to the watch's axes-system. The compass does not have a constant sampling rate; it was requested to provide an update of the heading whenever a change of more than one degree was detected. The compass takes some time to calibrate before providing measurements, so some examples that have watch acceleration measurements are missing compass measurements.

Location measurements: Both iPhone and Android provide location services (based on a combination of GPS, WiFi and cellular communications). The app samples location data in a non-constant rate, as the location service updates each time a movement is detected. This creates a time series of varying length (sometimes just a single time point in a recording session, sometimes more than 20 points) of location updates. Each update has a relative time reference and the estimated location measurements: latitude coordinate, longitude coordinate, altitude, speed, vertical accuracy and horizontal accuracy (these accuracies describe the range of reasonable error in location). Some of these values may be missing at times (*e.g.* when the phone is in a place with weak signals). In addition to the time series of location updates, the app calculates on the phone some basic heuristic location features: standard deviation of latitude values, standard deviation of longitude values, total change of latitude (last value minus first value), total change of longitude, average absolute latitude derivative and average absolute longitude derivative (as proxy to the speed of the user).

To protect our study subjects' privacy (collected examples with label "at home" that also include the exact location coordinates may reveal the subject's identity) the app has an option to select a location (typically their home) they would like to disguise. For the subjects that opted to use this option, whenever they were within 500 meters of their chosen location, the app would not send the latitude and longitude coordinates from the current recording (but it would send the other estimated location values such as altitude, speed, as well as the basic heuristic location features).

Low frequency measurements: These measurements were sampled just once in a recording session (approximately once per minute). Some of them describe the phone state (PS): app state (foreground/background), WiFi connectivity status, battery status (charging, discharging), battery level, or phone call status. Other low frequency measurements are taken from sensors built in to the phone, if available: proximity sensor, ambient light, temperature, humidity, air pressure.

Audio data: Audio was recorded from the phone's microphone in 22,050 Hz for the duration of the recording session (∼20 seconds). Audio was not available for recording when the phone was being used for a phone call. In order to maintain the privacy of the subjects, the raw audio recording was not sent to the server. Instead, standard audio processing features (Mel Frequency Cepstral Coefficients (MFCC)) were calculated on the phone itself and only the features were sent to the server. The MFCCs were calculated for half-overlapping windows of 2048 samples, based on 40 Mel scaled frequency bands and 13 cepstral coefficients (including the 0th coefficient). As part of the preprocessing of the recorded audio the raw audio signal was normalized to have maximal magnitude of 1 (dividing by the maximum absolute value of the sound wave). This normalizing factor is also sent as a measurement separately from the calculated MFCC features.

Extracted features

For the experiments in this work we focused on six sensors: accelerometer, gyroscope, watch accelerometer, location, audio and phone state. Other sensors' measurements are available in the public dataset. Every sensor measures different physical or virtual properties and has a different form of raw measurements. Therefore we designed specific features for each sensor. The published dataset includes files with these features pre-computed for all the users.

Accelerometer and Gyroscope (26 features each): Since in natural behavior the phone's position is not controlled we cannot assume it is oriented in a particular way, and it also may be changing its axes-system with respect to the ground (and with respect to the person). For that reason we had little reason to assume that any of the phone's axes will have a particular coherent correspondence to many behavioral patterns, and we extracted most of the features based on the overall magnitude of the signal. We calculated the vector magnitude signal as the euclidean norm of the 3-axis acceleration measurement at each point in time, *i.e.*, $a[t] = \sqrt{a_x[t]^2 + a_y[t]^2 + a_z[t]^2}$. We extracted the following features:

- Nine statistics of the magnitude signal: mean, standard deviation, third moment, fourth moment, 25^{th} percentile, 50^{th} percentile, 75^{th} percentile, value-entropy (entropy calculated from a histogram of quantization of the magnitude values to 20 bins) and time-entropy (entropy calculated from normalizing the magnitude signal and treating it as a probability distribution, which is designed to detect peakiness in time—sudden bursts of magnitude).
- Six spectral features of the magnitude signal: log energies in 5 sub-bands (0–0.5Hz, 0.5–1Hz, 1–3Hz, 3–5Hz, >5Hz) and spectral entropy.
- Two autocorrelation features from the magnitude signal. The average of the magnitude signal (DC component) was subtracted and the autocorrelation function was computed and normalized such that the autocorrelation value at lag 0 will be 1. The highest value after the main lobe was located. The corresponding period (in seconds) was calculated as the dominant periodicity and its normalized autocorrelation value was also extracted.
- Nine statistics of the 3-axis time series: the mean and standard deviation of each axis and the 3 inter-axis correlation coefficients.

Watch accelerometer (46 features): From the watch acceleration we extracted the same 26 features as from the phone accelerometer or gyroscope. Since the watch is positioned in a more controlled way than the phone (it is firmly fixed to the wrist), its axes have a strong meaning (*e.g.* motion along the x-axis of the watch describes a different kind of movement than motion along the z-axis of the watch). Hence we added 15 more axis-specific features—log energies in the same 5 sub-bands as before, this time calculated for each axis' signal separately. In addition, to account for the changes in watch orientation during the recording we calculated 5 relative-direction features in the following way: we first calculate the cosine-similarity between the acceleration directions of any two time points in the time series (value of 1 meaning same direction, value of -1 meaning opposite directions and value of 0 meaning orthogonal directions). Then we averaged these cosine similarity values in 5 different ranges of time-lag between the compared time points (0–0.5sec, 0.5–1sec, 1–5sec, 5–10sec, >10sec).

Location (17 features): In this work we used location features that were based only on relative locations, and not on absolute latitude/longitude coordinates. This was in order to avoid over-fitting to our location-homogeneous training set that will not generalize well to the outside world (*e.g.*, mistakenly learning that a specific location in the UCSD campus always corresponds to “at work”). Six features were calculated on the phone — this was in order to have some basic location features in cases where the subjects opted to hide their absolute location. These quick features included standard deviation of latitude, standard deviation of longitude, change in latitude (last value minus first value), change in longitude, average absolute value of derivative of latitude and average absolute value of derivative of longitude.

The transmitted location measurements were further processed to extract the following 11 features: number of updates (indicating how much the location changed during the 20 second recording), log of latitude-range (if latitudes were transmitted), log of longitude-range (if longitudes were transmitted), minimum altitude, maximum altitude, minimum speed, maximum speed, best (lowest) vertical accuracy, best (lowest) horizontal accuracy and diameter (maximum distance between two locations in the recording session, in meters).

Audio (26 features): From the time series of 13-dimensional MFCC vectors (typically around 400 time frames) we calculated the average and standard deviation of each of the 13 coefficients.

Phone State (34 features): For this work we used only the discrete phone state measurements. We represented them with a 26-dimensional one-hot representation—for each property we created a binary indicator for each of the possible values the property can take, plus one indicator denoting missing data. This representation is a redundant coding of the phone state, but it facilitates the use of simple, linear classifiers over this long binary vector representation. The keys were: app state (3 options: active, inactive, background), battery plugged (3 options: AC, USB, wireless), battery state (6 options: unknown, unplugged, not charging, discharging, charging, full), in a phone call (2 options: false, true), ringer mode (3 options: normal, silent no vibrate, silent with vibrate) and WiFi status (3 options: not reachable, reachable via WiFi, reachable via WWAN).

In addition, we added a set of features indicating time-of-day information. We used the timestamp of every example and (using San Diego local time) extracted the hour component (one out of 24 discrete values). In order to get a flexible, useful representation we defined 8 half-overlapping time ranges: midnight-6am, 3am-9am, 6am-midday, 9am-3pm, midday-6pm, 3pm-9pm, 6pm-midnight and 9pm-3am. Then we represented each example's hour with an 8-bit binary value, where 2 bins will be active for 1 relevant time range.

Label processing

Since the labels are obtained by subjects self-reporting their own behaviors, the reliability of annotation is not perfect. In some cases, this was the result of the subject reporting labels some time after the activity had occurred and misremembering the exact time. More common are cases where the subject neglected to report labels when relevant activities occurred (perhaps because the subject was distracted, did not have time to specify all the relevant labels, or was not aware of another relevant label in the vocabulary). As part of cleaning the data, we created adjusted versions for some labels using two methods: based on location data and based on other labels.

Location adjusted labels. We collected absolute location coordinates of the examples that had location measurements (selecting the location update with best horizontal accuracy from each example) and visualized them on a map. This made it easier to correct some labels which were clearly reported incorrectly. In examples without location data the original label was maintained.

- "At the beach". According to the few examples that reported being at the beach we marked areas that should be regarded as beach (and manually verified their validity by viewing them on a map). We then adjusted the label by applying "At the beach" to any example with a location within these areas.
- "At home". For each subject we identified the location of their home (by visualizing on a map all locations of examples where the subject reported being at home) and marked the coordinates of a visual centroid. This was only done when it was clear that we had indeed identified a location of a home. Three subjects reported being at home in two different houses, in which case we marked the two locations as locations of home. Two subjects never reported being at home but it was clear from their location to identify their location of residence. Some subjects had none or very few examples of "at home" with location data, so no home location was noted and their original reported labels were used. To define the adjusted version of the label "at home", whenever a subject's location was within 15 meters of their marked home location (or either of the two marked home locations), the adjusted value was set to "true"; whenever a subject's location was farther than 100 meters from all the subject's marked home locations the adjusted value was set to "false". In other cases (when the location was between 15 and 100 meters from a home location, or when there was no location data available) we retained the subject's originally reported value for "at home". This adjustment removed some obviously false reports of "at home" (*e.g.*, when the subject was clearly on a drive on a freeway). The adjustment manifested mostly by adding the missing label "at home" to many examples where the subject was clearly at home but failed to report it.
- "At main workplace". Similarly to home label we identified for each subject (if they used the original label "at work") the centroid location of their main workplace and created a new label — "at main workplace" — in a similar way. Some subjects reported being at work in different locations, so the original label "at work" is still valid for analysis and may have a different meaning than "at main workplace" (which was designed to capture behavioral patterns typical to the most common place that a person works in). This adjustment removed some examples where the label "At work" was probably incorrectly reported, but more importantly, it added the missing label in cases where the subject was clearly present at their most common workplace.

Labels corrected using other labels. We used reported values of other labels to adjust some labels. In a few cases it was clear that the reported labels were mistakes (because the combination of labels was unreasonable). In other cases a relevant label was simply not reported, even though it clearly should be relevant according to the other reported labels.

- "Walking". We identified a few cases where subjects reported walking together with labels related to driving. In cases where location data was available, it was clear on the map that the correct activity was the drive and not the walk. In the adjusted version of "walking" we changed the value to "false" whenever the subject reported "on a bus", "in a car", "drive (I'm the driver)", "drive (I'm a passenger)", "motorbike", "skateboarding" or "at the pool".
- "Running". The adjusted version was set to "false" for the same activities as in the adjusted "walking" label, plus in cases where the subject reported "playing baseball" or "playing frisbee". Although these cases are likely still valid (because the subject decided to report they were running during these playing activities), we decided to create the adjusted "running" version to represent a more coherent running activity (assuming that the playing activities involve a mixture of running, walking and standing intermittently). While the adjusted versions of "walking" and "running" may have a few misses (*e.g.*, some minutes during a baseball game when the subject was purely running), these misses don't harm the integrity of the multi-class experiments,

which used only examples that had positive labels of "running", "walking", "sitting", *etc.*.

- "Exercise". The adjusted version was set to "true" whenever the subject reported "exercising", "running", "bicycling", "lifting weights", "elliptical machine", "treadmill", "stationary bike" or "at the gym". This adjustment boosted the representation of the exercise behavior and also took advantage of reported specific activities without enough examples to be analyzed on their own.
- "Indoors". The adjusted version was set to "true" whenever the subject reported "indoors", "sleeping", "toilet", "bathing — bath", "bathing — shower", "in class", "at home", "at a bar", "at the gym" or "elevator". It is reasonable that many subjects simply did not bother to report being indoors every time they did an activity indoors.
- "Outside". The adjusted version was set to "true" whenever the subject reported "outside", "skateboarding", "playing baseball", "playing frisbee", "gardening", "raking leaves", "strolling", "hiking", "at the beach", "at sea" or "motorbike".
- "At a restaurant". In the adjusted version we changed the value to "false" whenever the subject reported "on a bus", "in a car", "drive (I'm the driver)", "drive (I'm a passenger)" or "motorbike".

Classification methods

Our system uses binary logistic regression classifiers (with a fitted intercept). Logistic regression provides a real-valued output, interpreted as the probability of the relevance of the label (value larger than 0.5 yielding a decision of "relevant"). For each context label we used an independent model. We first randomly partitioned the training examples into internal training and validation subsets, allocating one third of the training examples for the validation subset, while maintaining the same proportion of positive *vs.* negative examples in both subsets. We then used grid search to select the cost parameter C for logistic regression: for each value (out of $\{0.001, 0.01, 0.1, 1, 10, 100\}$) we trained a logistic regression model on the internal train subset and tested the model on the validation subset. We selected the value of C that resulted in highest F1 measure on the validation subset. We then re-trained a logistic regression model with the selected value on the entire training set. For the leave-one-user-out experiment with the EF system we simplified the procedure and only trained the logistic regression models with value of $C = 1$ (instead of performing grid search). The learned weights from LFL for a set of selected labels that are presented in Figure 4 (A) are taken from the first (of five) training set of the cross validation evaluation. To look at misclassifications and to produce the confusion matrices in Figure 4 (B–G) we used the multiclass (one-versus-rest) version of logistic regression, with a fixed cost value of $C = 1$. Each multiclass experiment used the set of examples annotated with exactly one label from the examined label subset and with data from all of the sensors of interest (so an experiment with only accelerometer and gyroscope sensors might have more examples than an experiment with accelerometer, gyroscope and watch accelerometer). These experiments were more fitting than binary classification in cases where missing labels are common. For example, labels describing the phone's position were not always consistently annotated. A binary classifier will use all negative examples to learn a decision boundary, including examples the subject forgot to label, which may skew the results if there are many missing labels.

Performance evaluation

In order to make a fair comparison among the different sensors, evaluation was done on the subset of examples with data from all six core sensors available ($\sim$177k examples from 51 subjects). In the training phase, however, a single-sensor classifier was allowed to use all examples available (*e.g.*, all examples in the dataset had phone state data, so the PS single-sensor classifier was trained with all examples). While the early fusion system benefited from the advantage of modeling correlations between features of different sensors, it was limited to being trained only on examples with all sensor data available. The late fusion systems, on the other hand, had the advantage of using single-sensor classifiers that were trained on many more examples.

Classifier performance was evaluated using 5-fold cross validation. The subjects were randomly partitioned once into 5 folds, while equalizing the proportion of iPhone *vs.* Android users between folds (To keep a fair evaluation it was important to partition the subjects, and not randomly partition the pool of examples, in order to avoid having examples from the same subject appear in both the training set and the test set). The cross validation procedure repeats the following for each fold: (1) hold out the selected fold to act as the test set (2) train a classifier on the remaining folds (3) apply the classifier to the held out test set. For each fold and for each label, we counted the numbers of true positives (TP. Examples that were correctly classified as positive), true negatives (TN. Examples that were correctly classified as negative), false positives (FP. Examples that were wrongfully classified as positive) and false negatives (FN. Examples that were wrongfully classified as negative). At the end of the 5-fold procedure we summed up the total numbers of TP, TN, FP and FN over the entire evaluation set and calculated the following performance metrics:

- *Accuracy* is the proportion of correctly classified examples out of all the examples. This metric is sensitive to imbalanced label proportion in the data.
- *True positive rate* (TPR, also called sensitivity or recall) is the proportion of positive examples that were correctly classified as positive: recall=TPR=TP/(TP+FN).
- *True negative rate* (TNR, also called specificity) is the proportion of negative examples that were correctly classified as negative: TNR=TN/(TN+FP).
- *Precision* (prec) is the proportion of correctly classified examples out of the examples that the classifier declared as positive: precision=TP/(TP+FP).
- *Balanced accuracy* is a measure that accounts for the tradeoff between true positive rate and true negative rate: BA=(TPR+TNR)/2.

- The *F1* measure is another such measure, which takes the harmonic mean of precision and recall: F1=(2*TPR*prec)/(TPR+prec).

While the balanced accuracy is easy to interpret (chance level is always 0.5 and perfect performance is 1) the F1 measure is very sensitive to how rare the positive examples are, so for each label a typical F1 value is different. The 5-fold subject partition is available with the dataset, and we encourage researchers using the dataset to evaluate new methods to use the same 5-fold partition, in order to promote fair comparisons.

Random chance. To assess the statistical significance of the performance scores we achieved, we evaluated a distribution of performance scores achieved by a random classifier. The random classifier declares "relevant" with probability 0.5 independently for each example and for each label. To estimate the distribution of scores that such a classifier obtains, we ran 100 simulations (each time the classifier randomly assigned binary predictions and the performance scores were calculated over the entire evaluation dataset). Chance level (expected value of the random classifier) of balanced accuracy is 0.5 for every label. For the F1 measure the chance level for each label is dependent on the proportion of positive and negative examples in the data. For each performance measure and for each label we estimated a value which we call *p99*, the 99^{th} percentile among the 100 scores achieved in the 100 simulations. Hence the probability of obtaining a score greater than p99 by chance is less than 1%. For average (over a set of labels) scores the p99 value was calculated similarly (computing the average score for each of the 100 simulations).

User personalization assessment

To assess the advantages of user personalization, we selected a single test subject that had provided relatively many examples and many labels. We partitioned this user's examples into the first half and second half of the examples (according to their recording timestamps), to simulate an adaptation training period (the first half) and a deployment period (the second half). We used early fusion (EF) classifiers to combine the features from all 6 sensors. The *universal model* was the one used in previous experiments, taken from the fold where the test user was part of the cross validation test set (so the universal model was trained on 48 other users). The *individual model* was trained only on data from the test user, taken from the first half of the subject's examples. The *adapted model* was a combination of both the universal and individual models using the LFA method (*i.e.,* averaging the probability outputs of both models). All three models were tested on the same set of unseen test examples (the second half of the subject's examples). For some labels, an insufficient number of examples to train an individual classifier resulted in a trivial classifier (always declaring the same answer). In those cases the performance was reported as chance level (BA of 0.5 and F1 of 0).

DETAILED RESULTS TABLES

5-fold cross validation evaluation

	n_e	n_s	p99	**Acc**	**Gyro**	**WAcc**	**Loc**	**Aud**	**PS**	**EF**	**LFA**	**LFL**
Lying down	54359	47	0.50	0.72	0.69	0.81	0.66	0.79	0.85	0.87	0.86	**0.88**
Sitting	82904	50	0.50	0.63	0.61	0.68	0.61	0.65	0.69	**0.76**	0.75	0.75
Walking	11892	50	0.51	0.77	0.80	0.75	0.66	0.63	0.70	0.80	0.80	**0.80**
Running	675	19	0.52	0.69	0.66	**0.80**	0.56	0.48	0.58	0.62	0.69	0.71
Bicycling	3523	22	0.51	0.81	0.81	0.84	0.81	0.77	0.83	0.87	0.87	**0.87**
Sleeping	42920	40	0.50	0.75	0.70	0.81	0.62	0.79	0.87	0.88	0.88	**0.89**
Lab work	2898	8	0.51	0.71	0.62	0.65	0.84	0.71	0.81	0.84	**0.87**	0.85
In class	2872	13	0.51	0.60	0.63	0.57	0.74	0.76	0.67	0.70	0.77	**0.80**
In a meeting	2904	34	0.51	0.60	0.57	0.62	0.63	0.79	0.73	0.80	0.79	**0.82**
At main workplace	20382	26	0.50	0.57	0.49	0.63	0.76	0.65	0.78	0.80	0.80	**0.81**
Indoors	107944	51	0.50	0.66	0.66	0.67	0.63	0.71	0.72	0.75	0.75	**0.76**
Outside	7629	36	0.51	0.70	0.73	0.70	0.66	0.66	0.73	0.74	0.77	**0.78**
In a car	3635	24	0.51	0.79	0.65	0.71	0.81	0.77	0.84	0.85	0.86	**0.86**
On a bus	1185	24	0.52	0.73	0.69	0.67	0.75	0.74	0.82	0.77	**0.84**	0.83
Drive (I'm the driver)	5034	24	0.51	0.79	0.61	0.75	0.82	0.74	0.83	0.84	0.86	**0.87**
Drive (I'm a passenger)	1655	19	0.51	0.76	0.71	0.64	0.79	0.76	0.81	0.84	0.84	**0.85**
At home	83977	50	0.50	0.65	0.63	0.66	0.63	0.71	0.70	0.75	0.77	**0.78**
At a restaurant	1320	16	0.52	0.62	0.67	0.68	0.58	**0.85**	0.77	0.76	0.83	0.81
Phone in pocket	15301	31	0.50	0.69	0.75	0.67	0.61	0.64	0.72	0.77	**0.77**	0.77
Exercise	5384	36	0.51	0.73	0.73	0.77	0.71	0.70	0.77	0.81	0.80	**0.81**
Cooking	2257	33	0.51	0.52	0.53	0.68	0.57	0.62	0.68	0.71	0.71	**0.72**
Shopping	896	18	0.52	0.70	0.70	0.69	0.54	0.59	0.79	0.69	0.76	**0.80**
Strolling	434	8	0.53	0.67	0.74	0.72	0.67	0.64	0.75	0.66	**0.77**	0.74
Drinking (alcohol)	864	10	0.52	0.71	0.69	0.50	0.56	0.80	0.74	0.70	**0.82**	0.81
Bathing - shower	1186	27	0.52	0.53	0.55	**0.73**	0.47	0.63	0.47	0.64	0.67	0.70
average			0.50	0.68	0.66	0.70	0.67	0.70	0.75	0.77	0.80	0.80

TABLE S1

5-fold evaluation performance (BA) of the different classifiers on each label. Part 1 of the labels. For each label n_e is the number of examples and n_s is the number of subjects in the testing (possibly more examples participated in the training). p99 marks the 99^{th} percentile of random scores — a score above the p99 value has less than 0.01 probability to be achieved randomly. For each label the score of the highest performing classifier is marked in bold.

	n_e	n_s	p99	**Acc**	**Gyro**	**WAcc**	**Loc**	**Aud**	**PS**	**EF**	**LFA**	**LFL**
Cleaning	1839	22	0.51	0.63	0.64	**0.71**	0.41	0.60	0.51	0.60	0.70	0.68
Laundry	473	12	0.52	0.65	0.66	**0.66**	0.38	0.53	0.65	0.58	0.63	0.63
Washing dishes	851	17	0.52	0.40	0.52	0.70	0.58	0.60	0.57	0.65	0.70	**0.70**
Watching TV	9412	28	0.51	0.61	0.54	0.56	0.56	0.64	0.67	0.65	**0.70**	0.68
Surfing the internet	11641	28	0.50	0.56	0.55	0.60	0.54	0.60	0.57	0.59	0.63	**0.63**
At a party	404	3	0.53	0.74	0.71	0.49	0.54	**0.81**	0.56	0.54	0.76	0.75
At a bar	520	4	0.53	0.45	0.66	0.53	0.60	0.49	**0.93**	0.50	0.61	0.66
At the beach	122	5	0.55	0.62	0.48	0.47	**0.72**	0.58	0.70	0.50	0.71	0.70
Singing	384	6	0.53	0.57	0.64	0.46	0.61	**0.68**	0.59	0.50	0.65	0.53
Talking	18976	44	0.50	0.60	0.61	0.60	0.54	0.65	0.65	0.65	0.67	**0.67**
Computer work	23692	38	0.50	0.57	0.56	0.62	0.63	0.61	0.68	0.68	**0.71**	0.70
Eating	10169	49	0.51	0.59	0.58	0.60	0.51	0.61	0.62	**0.66**	0.65	0.65
Toilet	1646	33	0.51	0.57	0.51	0.58	0.57	0.64	0.59	0.65	0.66	**0.66**
Grooming	1847	25	0.51	0.44	0.49	0.62	0.59	0.63	0.58	0.60	**0.63**	0.63
Dressing	1308	27	0.52	0.51	0.52	0.64	0.54	0.65	0.61	0.64	**0.67**	0.67
At the gym	906	6	0.52	0.50	0.55	0.58	0.57	0.65	**0.70**	0.54	0.64	0.61
Stairs - going up	399	17	0.53	0.70	**0.73**	0.65	0.55	0.55	0.51	0.58	0.69	0.67
Stairs - going down	390	15	0.53	0.71	**0.73**	0.66	0.55	0.55	0.51	0.58	0.71	0.66
Elevator	124	8	0.55	0.72	**0.76**	0.44	0.54	0.71	0.51	0.49	0.73	0.73
Standing	22766	51	0.50	0.60	0.59	0.67	0.54	0.59	0.63	**0.68**	0.67	0.68
At school	25840	39	0.50	0.59	0.59	0.59	0.66	0.64	0.68	0.70	**0.70**	0.70
Phone in hand	8595	37	0.51	0.65	**0.68**	0.56	0.59	0.59	0.61	0.64	0.67	0.66
Phone in bag	5589	22	0.51	0.59	0.56	0.55	0.59	0.64	0.69	0.67	0.68	**0.69**
Phone on table	70611	43	0.50	0.60	0.61	0.56	0.53	0.55	0.61	0.61	0.62	**0.62**
With co-workers	4139	17	0.51	0.57	0.57	0.61	0.58	0.68	0.67	0.69	0.71	**0.72**
With friends	12865	25	0.50	0.55	0.58	0.53	0.54	0.60	0.60	0.55	**0.61**	0.58
average			0.50	0.59	0.60	0.59	0.56	0.62	0.62	0.60	0.67	0.66

TABLE S2

5-fold evaluation performance (BA) of the different classifiers on each label. Part 2 of the labels. For each label n_e is the number of examples and n_s is the number of subjects in the testing (possibly more examples participated in the training). p99 marks the 99^{th} percentile of random scores — a score above the p99 value has less than 0.01 probability to be achieved randomly. For each label the score of the highest performing classifier is marked in bold.

	n_e	n_s	p99	**Acc**	**Gyro**	**WAcc**	**Loc**	**Aud**	**PS**	**EF**	**LFA**	**LFL**
Lying down	54359	47	0.38	0.61	0.59	0.71	0.55	0.69	0.78	0.81	0.79	**0.82**
Sitting	82904	50	0.49	0.58	0.58	0.67	0.59	0.62	0.72	**0.75**	0.75	0.74
Walking	11892	50	0.12	0.38	0.38	0.31	0.22	0.19	0.22	0.38	**0.39**	0.38
Running	675	19	0.01	0.03	0.03	**0.04**	0.01	0.01	0.01	0.03	0.04	0.04
Bicycling	3523	22	0.04	0.19	0.16	0.23	0.18	0.12	0.17	**0.31**	0.25	0.26
Sleeping	42920	40	0.33	0.57	0.53	0.65	0.44	0.63	0.75	0.79	0.77	**0.81**
Lab work	2898	8	0.03	0.08	0.06	0.06	0.11	0.09	0.11	**0.21**	0.16	0.19
In class	2872	13	0.03	0.05	0.06	0.04	0.07	0.10	0.06	0.13	0.12	**0.14**
In a meeting	2904	34	0.03	0.05	0.04	0.05	0.05	0.11	0.07	**0.17**	0.10	0.14
At main workplace	20382	26	0.19	0.23	0.18	0.28	0.41	0.31	0.42	0.49	0.47	**0.50**
Indoors	107944	51	0.55	0.74	0.73	0.70	0.68	0.75	0.71	0.78	**0.79**	0.78
Outside	7629	36	0.08	0.20	0.20	0.18	0.15	0.15	0.20	0.23	**0.26**	0.25
In a car	3635	24	0.04	0.15	0.07	0.10	**0.27**	0.13	0.16	0.23	0.22	0.23
On a bus	1185	24	0.01	0.04	0.03	0.03	0.07	0.04	0.06	0.07	**0.07**	0.06
Drive (I'm the driver)	5034	24	0.06	0.21	0.09	0.15	**0.37**	0.16	0.23	0.31	0.31	0.31
Drive (I'm a passenger)	1655	19	0.02	0.07	0.04	0.04	**0.15**	0.07	0.08	0.14	0.12	0.12
At home	83977	50	0.49	0.66	0.65	0.64	0.63	0.70	0.67	0.74	0.76	**0.77**
At a restaurant	1320	16	0.02	0.02	0.03	0.03	0.02	0.08	0.05	**0.11**	0.07	0.09
Phone in pocket	15301	31	0.15	0.28	0.33	0.26	0.21	0.25	0.28	**0.38**	0.36	0.37
Exercise	5384	36	0.06	0.21	0.18	0.24	0.14	0.13	0.19	**0.27**	0.26	0.25
Cooking	2257	33	0.03	0.03	0.03	0.05	0.03	0.04	0.06	**0.09**	0.07	0.08
Shopping	896	18	0.01	0.03	0.03	0.02	0.01	0.02	0.04	**0.04**	0.04	0.04
Strolling	434	8	0.01	0.02	0.02	0.01	0.01	0.01	0.02	0.03	**0.03**	0.03
Drinking (alcohol)	864	10	0.01	0.03	0.02	0.01	0.01	0.04	0.03	**0.07**	0.07	0.06
Bathing - shower	1186	27	0.01	0.01	0.02	0.04	0.01	0.02	0.01	0.04	0.04	**0.05**
average			0.13	0.22	0.20	0.22	0.22	0.22	0.24	0.30	0.29	0.30

TABLE S3

5-fold evaluation performance (F1) of the different classifiers on each label. Part 1 of the labels. For each label n_e is the number of examples and n_s is the number of subjects in the testing (possibly more examples participated in the training). p99 marks the 99^{th} percentile of random scores — a score above the p99 value has less than 0.01 probability to be achieved randomly. For each label the score of the highest performing classifier is marked in bold.

	n_e	n_s	p99	**Acc**	**Gyro**	**WAcc**	**Loc**	**Aud**	**PS**	**EF**	**LFA**	**LFL**
Cleaning	1839	22	0.02	0.05	0.05	0.05	0.01	0.03	0.02	0.05	**0.06**	0.05
Laundry	473	12	0.01	0.01	0.01	0.01	0.00	0.01	0.01	**0.02**	0.01	0.02
Washing dishes	851	17	0.01	0.01	0.01	0.03	0.01	0.02	0.01	0.03	0.03	**0.04**
Watching TV	9412	28	0.10	0.14	0.11	0.12	0.12	0.17	0.18	0.21	0.22	**0.22**
Surfing the internet	11641	28	0.12	0.15	0.14	0.17	0.13	0.17	0.15	0.18	0.19	**0.20**
At a party	404	3	0.01	0.01	0.01	0.00	0.01	0.03	0.01	**0.04**	0.03	0.04
At a bar	520	4	0.01	0.00	0.01	0.01	0.01	0.00	**0.09**	0.00	0.03	0.06
At the beach	122	5	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	**0.02**	0.02
Singing	384	6	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.00	**0.01**	0.01
Talking	18976	44	0.18	0.25	0.26	0.24	0.21	0.29	0.27	0.29	0.30	**0.30**
Computer work	23692	38	0.21	0.26	0.25	0.30	0.31	0.30	0.35	0.38	0.39	**0.39**
Eating	10169	49	0.11	0.15	0.14	0.15	0.11	0.16	0.15	**0.19**	0.18	0.17
Toilet	1646	33	0.02	0.03	0.02	0.03	0.02	0.03	0.03	**0.04**	0.04	0.04
Grooming	1847	25	0.02	0.02	0.02	0.03	0.03	0.04	0.03	0.05	0.04	**0.05**
Dressing	1308	27	0.02	0.02	0.02	0.03	0.02	0.03	0.03	**0.04**	0.04	0.04
At the gym	906	6	0.01	0.01	0.01	0.02	0.01	0.02	0.02	0.03	0.03	**0.03**
Stairs - going up	399	17	0.01	**0.02**	0.02	0.01	0.01	0.01	0.00	0.01	0.02	0.01
Stairs - going down	390	15	0.00	**0.02**	0.02	0.01	0.01	0.01	0.00	0.01	0.02	0.01
Elevator	124	8	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01	**0.01**
Standing	22766	51	0.21	0.28	0.27	0.35	0.23	0.27	0.29	**0.36**	0.34	0.35
At school	25840	39	0.23	0.30	0.29	0.30	0.38	0.34	0.39	0.41	0.41	**0.41**
Phone in hand	8595	37	0.09	0.17	**0.17**	0.11	0.13	0.13	0.13	0.16	0.17	0.16
Phone in bag	5589	22	0.06	0.10	0.08	0.07	0.09	0.11	0.12	**0.15**	0.14	0.14
Phone on table	70611	43	0.45	0.58	0.58	0.51	0.50	0.51	0.56	0.56	**0.59**	0.58
With co-workers	4139	17	0.05	0.06	0.06	0.07	0.06	0.11	0.08	0.13	0.12	**0.13**
With friends	12865	25	0.13	0.15	0.17	0.14	0.15	0.18	0.18	0.15	**0.19**	0.18
average			0.08	0.11	0.11	0.11	0.10	0.11	0.12	0.13	0.14	0.14

TABLE S4

5-fold evaluation performance (F1) of the different classifiers on each label. Part 2 of the labels. For each label n_e is the number of examples and n_s is the number of subjects in the testing (possibly more examples participated in the training). p99 marks the 99^{th} percentile of random scores — a score above the p99 value has less than 0.01 probability to be achieved randomly. For each label the score of the highest performing classifier is marked in bold.

Leave-one-user-out evaluation

	n_e	n_s	p99	**Acc**	**Gyro**	**WAcc**	**Loc**	**Aud**	**PS**	**EF**	**LFA**	**LFL**
Lying down	54359	47	0.50	0.73	0.69	0.81	0.65	0.79	0.84	0.87	0.86	**0.88**
Sitting	82904	50	0.50	0.63	0.61	0.68	0.61	0.65	0.69	**0.76**	0.75	0.75
Walking	11892	50	0.51	0.77	0.80	0.75	0.66	0.63	0.71	0.80	0.80	**0.81**
Running	675	19	0.52	0.69	0.69	**0.80**	0.56	0.50	0.57	0.67	0.72	0.76
Bicycling	3523	22	0.51	0.81	0.81	0.85	0.80	0.76	0.80	**0.87**	0.86	0.87
Sleeping	42920	40	0.50	0.75	0.70	0.81	0.62	0.80	0.85	0.88	0.88	**0.89**
Lab work	2898	8	0.51	0.69	0.61	0.65	0.82	0.70	0.84	0.84	**0.84**	0.84
In class	2872	13	0.51	0.61	0.63	0.58	0.74	0.77	0.72	0.74	0.79	**0.81**
In a meeting	2904	34	0.51	0.62	0.59	0.62	0.66	0.78	0.73	0.81	0.80	**0.82**
At main workplace	20382	26	0.50	0.55	0.49	0.64	0.76	0.65	0.80	0.80	0.81	**0.82**
Indoors	107944	51	0.50	0.67	0.66	0.68	0.63	0.70	0.72	**0.76**	0.75	0.76
Outside	7629	36	0.51	0.72	0.74	0.70	0.66	0.67	0.71	0.75	0.78	**0.79**
In a car	3635	24	0.51	0.79	0.66	0.71	0.82	0.76	0.83	0.85	0.86	**0.87**
On a bus	1185	24	0.52	0.74	0.69	0.68	0.72	0.72	0.80	0.78	**0.83**	0.82
Drive (I'm the driver)	5034	24	0.51	0.80	0.62	0.75	0.83	0.75	0.84	0.84	0.86	**0.87**
Drive (I'm a passenger)	1655	19	0.51	0.76	0.70	0.64	0.80	0.77	0.82	**0.84**	0.83	0.84
At home	83977	50	0.50	0.65	0.63	0.66	0.62	0.72	0.72	0.76	0.77	**0.77**
At a restaurant	1320	16	0.52	0.62	0.68	0.69	0.57	**0.84**	0.74	0.79	0.84	0.84
Phone in pocket	15301	31	0.50	0.69	0.75	0.67	0.61	0.64	0.71	0.77	0.77	**0.77**
Exercise	5384	36	0.51	0.74	0.73	0.77	0.71	0.70	0.75	0.81	0.81	**0.81**
Cooking	2257	33	0.51	0.52	0.55	0.67	0.57	0.62	0.68	0.71	0.72	**0.72**
Shopping	896	18	0.52	0.71	0.69	0.68	0.53	0.57	**0.79**	0.67	0.75	0.78
Strolling	434	8	0.53	0.64	0.73	0.70	0.63	0.62	0.71	0.67	0.74	**0.75**
Drinking (alcohol)	864	10	0.52	0.72	0.70	0.54	0.56	0.79	0.54	0.68	**0.79**	0.77
Bathing - shower	1186	27	0.52	0.50	0.54	**0.74**	0.48	0.63	0.48	0.64	0.69	0.72
average			0.50	0.68	0.67	0.70	0.66	0.70	0.74	0.78	0.80	0.81

TABLE S5

Leave-one-user-out evaluation performance (BA) of the different classifiers on each label. Part 1 of the labels. For each label n_e is the number of examples and n_s is the number of subjects in the testing (possibly more examples participated in the training). p99 marks the 99th percentile of random scores — a score above the p99 value has less than 0.01 probability to be achieved randomly. For each label the score of the highest performing classifier is marked in bold.

	n_e	n_s	p99	**Acc**	**Gyro**	**WAcc**	**Loc**	**Aud**	**PS**	**EF**	**LFA**	**LFL**
Cleaning	1839	22	0.51	0.62	0.63	**0.73**	0.42	0.62	0.49	0.70	0.71	0.70
Laundry	473	12	0.52	0.67	0.65	0.66	0.35	0.52	**0.78**	0.60	0.68	0.70
Washing dishes	851	17	0.52	0.36	0.48	**0.69**	0.54	0.61	0.54	0.66	0.67	0.68
Watching TV	9412	28	0.51	0.61	0.54	0.57	0.57	0.67	0.66	0.69	**0.72**	0.71
Surfing the internet	11641	28	0.50	0.55	0.58	0.59	0.56	0.60	0.57	0.61	**0.62**	0.62
At a party	404	3	0.53	0.73	0.71	0.48	0.70	**0.84**	0.67	0.52	0.79	0.76
At a bar	520	4	0.53	0.53	0.69	0.50	0.64	0.62	**0.88**	0.52	0.71	0.68
At the beach	122	5	0.55	0.66	0.51	0.52	**0.71**	0.58	0.69	0.57	0.71	0.71
Singing	384	6	0.53	0.56	0.62	0.46	**0.70**	0.68	0.60	0.48	0.68	0.53
Talking	18976	44	0.50	0.61	0.61	0.61	0.55	0.66	0.64	0.66	0.68	**0.68**
Computer work	23692	38	0.50	0.59	0.57	0.62	0.65	0.59	0.67	0.69	**0.71**	0.69
Eating	10169	49	0.51	0.59	0.58	0.60	0.53	0.61	0.63	0.66	**0.66**	0.66
Toilet	1646	33	0.51	0.57	0.52	0.57	0.57	0.63	0.56	0.65	**0.65**	0.65
Grooming	1847	25	0.51	0.46	0.53	0.62	0.60	0.65	0.53	0.63	0.64	**0.66**
Dressing	1308	27	0.52	0.51	0.54	0.66	0.53	0.67	0.55	0.66	0.67	**0.68**
At the gym	906	6	0.52	0.55	0.56	0.67	0.51	0.67	**0.70**	0.58	0.67	0.67
Stairs - going up	399	17	0.53	0.68	**0.76**	0.65	0.57	0.57	0.48	0.59	0.69	0.66
Stairs - going down	390	15	0.53	0.70	**0.75**	0.66	0.54	0.55	0.48	0.57	0.69	0.63
Elevator	124	8	0.55	0.68	0.70	0.56	0.57	**0.70**	0.54	0.50	0.62	0.61
Standing	22766	51	0.50	0.60	0.59	0.67	0.54	0.59	0.62	**0.68**	0.66	0.67
At school	25840	39	0.50	0.60	0.59	0.59	0.68	0.66	0.70	**0.72**	0.71	0.71
Phone in hand	8595	37	0.51	0.66	**0.68**	0.56	0.58	0.58	0.59	0.63	0.67	0.66
Phone in bag	5589	22	0.51	0.60	0.56	0.56	0.59	0.69	0.69	0.72	0.71	**0.73**
Phone on table	70611	43	0.50	0.60	0.61	0.56	0.52	0.56	0.61	0.61	**0.63**	0.62
With co-workers	4139	17	0.51	0.55	0.57	0.61	0.61	0.68	0.71	0.69	0.73	**0.74**
With friends	12865	25	0.50	0.56	0.57	0.54	0.55	0.62	0.59	0.58	**0.63**	0.61
average			0.50	0.59	0.60	0.60	0.57	0.63	0.62	0.62	0.68	0.67

TABLE S6

Leave-one-user-out evaluation performance (BA) of the different classifiers on each label. Part 2 of the labels. For each label n_e is the number of examples and n_s is the number of subjects in the testing (possibly more examples participated in the training). p99 marks the 99th percentile of random scores — a score above the p99 value has less than 0.01 probability to be achieved randomly. For each label the score of the highest performing classifier is marked in bold.

	n_e	n_s	p99	**Acc**	**Gyro**	**WAcc**	**Loc**	**Aud**	**PS**	**EF**	**LFA**	**LFL**
Lying down	54359	47	0.38	0.62	0.59	0.71	0.54	0.69	0.76	0.81	0.79	**0.82**
Sitting	82904	50	0.49	0.58	0.58	0.68	0.58	0.62	0.71	**0.75**	0.75	0.74
Walking	11892	50	0.12	0.38	0.37	0.32	0.21	0.19	0.22	0.38	**0.39**	0.39
Running	675	19	0.01	0.03	0.02	0.04	0.01	0.01	0.01	0.04	0.04	**0.04**
Bicycling	3523	22	0.04	0.19	0.15	0.23	0.16	0.12	0.15	**0.30**	0.24	0.26
Sleeping	42920	40	0.33	0.56	0.53	0.65	0.44	0.64	0.74	0.79	0.76	**0.80**
Lab work	2898	8	0.03	0.07	0.06	0.06	0.11	0.08	0.11	**0.21**	0.15	0.18
In class	2872	13	0.03	0.05	0.06	0.04	0.08	0.11	0.07	0.13	0.12	**0.14**
In a meeting	2904	34	0.03	0.06	0.04	0.05	0.06	0.11	0.07	**0.17**	0.11	0.15
At main workplace	20382	26	0.19	0.22	0.19	0.29	0.41	0.31	0.43	0.49	0.48	**0.52**
Indoors	107944	51	0.55	0.75	0.73	0.71	0.68	0.75	0.71	**0.79**	0.79	0.79
Outside	7629	36	0.08	0.21	0.20	0.18	0.16	0.16	0.20	0.23	**0.26**	0.25
In a car	3635	24	0.04	0.15	0.08	0.10	**0.27**	0.13	0.16	0.23	0.22	0.23
On a bus	1185	24	0.01	0.04	0.03	0.03	0.05	0.04	0.05	0.07	**0.07**	0.07
Drive (I'm the driver)	5034	24	0.06	0.21	0.09	0.16	**0.38**	0.15	0.21	0.31	0.31	0.31
Drive (I'm a passenger)	1655	19	0.02	0.07	0.04	0.04	**0.15**	0.07	0.08	0.13	0.12	0.11
At home	83977	50	0.49	0.67	0.65	0.65	0.63	0.71	0.69	0.75	0.76	**0.76**
At a restaurant	1320	16	0.02	0.03	0.03	0.03	0.02	0.07	0.04	**0.11**	0.07	0.10
Phone in pocket	15301	31	0.15	0.29	0.34	0.26	0.22	0.25	0.27	**0.38**	0.37	0.37
Exercise	5384	36	0.06	0.21	0.16	0.22	0.14	0.13	0.15	0.26	**0.26**	0.24
Cooking	2257	33	0.03	0.03	0.03	0.05	0.03	0.04	0.05	**0.08**	0.07	0.07
Shopping	896	18	0.01	0.03	0.02	0.02	0.01	0.01	**0.04**	0.04	0.04	0.04
Strolling	434	8	0.01	0.01	0.02	0.01	0.01	0.01	0.02	0.02	**0.02**	0.02
Drinking (alcohol)	864	10	0.01	0.03	0.02	0.01	0.01	0.04	0.01	0.05	**0.06**	0.05
Bathing - shower	1186	27	0.01	0.01	0.02	0.04	0.01	0.02	0.01	0.04	0.04	**0.05**
average			0.13	0.22	0.20	0.22	0.22	0.22	0.24	0.30	0.29	0.30

TABLE S7

Leave-one-user-out evaluation performance (F1) of the different classifiers on each label. Part 1 of the labels. For each label n_e is the number of examples and n_s is the number of subjects in the testing (possibly more examples participated in the training). p99 marks the 99th percentile of random scores — a score above the p99 value has less than 0.01 probability to be achieved randomly. For each label the score of the highest performing classifier is marked in bold.

	n_e	n_s	p99	**Acc**	**Gyro**	**WAcc**	**Loc**	**Aud**	**PS**	**EF**	**LFA**	**LFL**
Cleaning	1839	22	0.02	0.04	0.04	0.06	0.01	0.03	0.02	**0.07**	0.06	0.06
Laundry	473	12	0.01	0.01	0.01	0.01	0.00	0.01	0.02	0.02	0.02	**0.02**
Washing dishes	851	17	0.01	0.00	0.01	0.02	0.01	0.02	0.01	**0.03**	0.03	0.03
Watching TV	9412	28	0.10	0.14	0.11	0.12	0.12	0.19	0.17	0.23	0.22	**0.24**
Surfing the internet	11641	28	0.12	0.14	0.16	0.16	0.14	0.17	0.15	**0.20**	0.19	0.19
At a party	404	3	0.01	0.01	0.01	0.00	0.01	0.03	0.01	0.02	0.03	**0.04**
At a bar	520	4	0.01	0.01	0.01	0.01	0.01	0.02	**0.06**	0.01	0.04	0.05
At the beach	122	5	0.00	0.00	0.00	0.00	0.01	0.00	0.01	**0.05**	0.02	0.02
Singing	384	6	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.00	**0.01**	0.01
Talking	18976	44	0.18	0.25	0.25	0.25	0.21	0.30	0.26	0.30	0.30	**0.30**
Computer work	23692	38	0.21	0.28	0.26	0.30	0.32	0.28	0.34	0.39	**0.39**	0.39
Eating	10169	49	0.11	0.15	0.14	0.15	0.11	0.16	0.15	**0.18**	0.18	0.18
Toilet	1646	33	0.02	0.03	0.02	0.03	0.02	0.03	0.02	**0.04**	0.04	0.04
Grooming	1847	25	0.02	0.02	0.02	0.04	0.03	0.04	0.02	0.05	0.04	**0.05**
Dressing	1308	27	0.02	0.02	0.02	0.03	0.02	0.03	0.02	**0.04**	0.04	0.04
At the gym	906	6	0.01	0.01	0.01	0.02	0.01	0.02	0.03	**0.04**	0.03	0.04
Stairs - going up	399	17	0.01	0.01	**0.02**	0.01	0.01	0.01	0.00	0.01	0.02	0.01
Stairs - going down	390	15	0.00	0.01	**0.02**	0.01	0.01	0.01	0.00	0.01	0.02	0.01
Elevator	124	8	0.00	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Standing	22766	51	0.21	0.28	0.27	0.35	0.23	0.27	0.29	**0.35**	0.34	0.35
At school	25840	39	0.23	0.30	0.30	0.29	0.42	0.37	0.40	**0.44**	0.42	0.42
Phone in hand	8595	37	0.09	0.17	**0.17**	0.11	0.13	0.12	0.12	0.16	0.17	0.16
Phone in bag	5589	22	0.06	0.11	0.08	0.08	0.08	0.13	0.11	**0.16**	0.15	0.16
Phone on table	70611	43	0.45	**0.59**	0.58	0.51	0.48	0.51	0.56	0.55	0.59	0.58
With co-workers	4139	17	0.05	0.06	0.06	0.07	0.07	0.11	0.09	0.13	0.12	**0.14**
With friends	12865	25	0.13	0.16	0.17	0.14	0.15	0.20	0.17	0.18	**0.20**	0.20
average			0.08	0.11	0.11	0.11	0.10	0.12	0.12	0.14	0.14	0.14

TABLE S8

Leave-one-user-out evaluation performance (F1) of the different classifiers on each label. Part 2 of the labels. For each label n_e is the number of examples and n_s is the number of subjects in the testing (possibly more examples participated in the training). p99 marks the 99th percentile of random scores — a score above the p99 value has less than 0.01 probability to be achieved randomly. For each label the score of the highest performing classifier is marked in bold.