The Mobile Data Challenge: Big Data for Mobile Computing Research

Juha K. Laurila
Nokia Research Center
Lausanne, Switzerland
juha.k.laurila@nokia.com

Daniel Gatica-Perez
Idiap and EPFL, Switzerland
gatica@idiap.ch

Imad Aad
Nokia Research Center
Lausanne, Switzerland
imad.aad@nokia.com

Jan Blom
Nokia Research Center
Lausanne, Switzerland
jan.blom@nokia.com

Olivier Bornet
Idiap, Switzerland
bornet@idiap.ch

Trinh-Minh-Tri Do
Idiap, Switzerland
do@idiap.ch

Olivier Dousse
Nokia Research Center
Lausanne, Switzerland
olivier.dousse@nokia.com

Julien Eberle
Nokia Research Center
Lausanne, Switzerland
julien.eberle@nokia.com

Markus Miettinen
Nokia Research Center
Lausanne, Switzerland
markus.miettinen@nokia.com

ABSTRACT

This paper presents an overview of the Mobile Data Challenge (MDC), a large-scale research initiative aimed at generating innovations around smartphone-based research, as well as community-based evaluation of related mobile data analysis methodologies. First we review the Lausanne Data Collection Campaign (LDCC) – an initiative to collect unique, longitudinal smartphone data set for the basis of the MDC. Then, we introduce the Open and Dedicated Tracks of the MDC; describe the specific data sets used in each of them; and discuss some of the key aspects in order to generate privacy-respecting, challenging, and scientifically relevant mobile data resources for wider use of the research community. The concluding remarks will summarize the paper.

1. INTRODUCTION

Mobile phone technology has transformed the way we live, as phone adoption has increased rapidly across the globe [17]. This has widespread social implications. The phones themselves have become instruments for fast communication and collective participation. Further, different user groups, like teenagers, have started to use them in creative ways. At the same time, the number of sensors embedded in phones and the applications built around them have exploded. In the past few years smartphones remarkably started to carry sensors like GPS, accelerometer, gyroscope, microphone, camera and Bluetooth. Related application and service offering covers e.g. information search, entertainment or healthcare.

The ubiquity of mobile phones and the increasing wealth of the data generated from sensors and applications are giving rise to a new research domain across computing and social science. Researchers are beginning to examine issues in behavioral and social science from the Big Data perspective – by using large-scale mobile data as input to characterize and understand real-life phenomena, including individual traits, as well as human mobility, communication, and interaction patterns [11, 12, 9].

This new research, whose findings are clearly important to society at large, has been often conducted within corporations that historically have had access to these data types, including telecom operators [13] or Internet companies [6], or through granted data access to academics in highly restricted forms [12]. Some initiatives, like [1], have collected publicly available but in some extent limited data sets together. Clearly, government and corporate regulations for privacy and data protection play a fundamental and necessary role in protecting all sensitive aspects of mobile data. From the research perspective, this also implies that mobile data resources are scarce and often not ecologically valid to test scientific hypotheses related to real-life behavior.

The Mobile Data Challenge (MDC) by Nokia is motivated by our belief in the value of mobile computing research for the common good - i.e., of research that can result in a deeper scientific understanding of human and social phenomena, advanced mobile experiences and technological innovations. Guided by this principle, in January 2009 Nokia Research Center Lausanne and its Swiss academic partners Idiap and EPFL started an initiative to create large-scale mobile data research resources. This included the design and implementation of the Lausanne Data Collection Campaign (LDCC), an effort to collect a longitudinal smartphone data set from nearly 200 volunteers in the Lake Geneva region over one year of time. It also involved the definition of a number of research tasks with clearly specified experimental protocols. From the very beginning the intention was to share these research resources with the research community which required integration of holistic and proactive

approach on privacy according to the of privacy-by-design principles [2].

The MDC is the visible outcome of nearly three years of work in this direction. The Challenge provided researchers with an opportunity to analyze a relatively unexplored data set including rich mobility, communication, and interaction information. The MDC comprised of two research alternatives through an Open Research Track and a Dedicated Research Track. In the Open Track, researchers were given opportunity to approach the data set from an exploratory perspective, by proposing their own tasks according to their interests and background. The Dedicated Track gave researchers the possibility to take on up to three tasks to solve, related with prediction of mobility patterns, recognition of place categories, and estimation of demographic attributes. Each of these tasks had properly defined experimental protocols and standard evaluation measures to assess and rank all contributions.

This paper presents an overview of the Mobile Data Challenge intended both for participants of the MDC and a wider audience. Section 2 summarizes the LDCC data, the basis for the MDC. Section 3 describes the MDC tracks and tasks in detail. Section 4 provides details on the specific data sets used for the MDC. Section 5 summarizes the schedule we have followed to organize the Challenge. Finally, Section 6 offers some final remarks.

2. THE LAUSANNE DATA COLLECTION CAMPAIGN (LDCC)

LDCC aimed at designing and implementing a large-scale campaign to collect smartphone data in everyday life conditions, grounding the study on a European culture. The overall goal was to collect quasi-continuous measurements covering all sensory and other available information on a smartphone. This way we were able to capture phone users' daily activities unobtrusively, in a setting that implemented the privacy-by-design principles [2]. The collected data included a significant amount of behavioral information, including both personal and relational aspects. This enables investigation of a large number of research questions related to personal and social context - including mobility, phone usage, communication, and interaction. Only content, like image files or content of the messages, was excluded because content capturing was considered too intrusive for the longitudinal study based on volunteering participation with selfless drivers. Instead log-files with metadata were collected both for imaging and messaging applications. This section provides a summary on the LDCC implementation and captured data types. An initial paper introducing LDCC, its data types and statistics early 2010 appeared in [14]. Part of the material in this section has been adapted from it.

2.1 LDCC design

Nokia Research Center, Idiap, and EPFL partnered towards LDCC since January 2009. After the implementation and evaluation of the sensing architecture, and the recruitment of the initial pool of volunteers, the data collection started in October 2009. Over time, smartphones with data collection software were allocated to close to 200 volunteers in the Lake Geneva region. A viral approach was used to promote the campaign and recruit volunteers. This resulted in a great proportion of the members of the campaign population having social connections to other participants, as well as to the demographical representativeness. A key aspect of the success of LDCC was the enthusiastic participation of volunteers who agreed to participate and share their data mainly driven by selfless interest. The campaign concluded in March 2011.

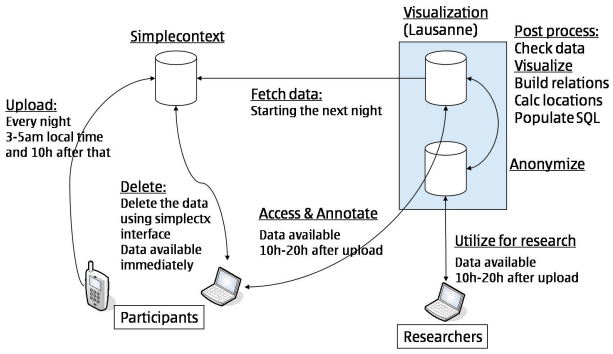

Figure 1: LDCC data flow, progressing from mobile data from volunteers to anonymized data for research [14]).

Data was collected using Nokia N95 phones and a client-server architecture that made the data collection invisible to the participants. A seamless implementation of the data recording process was a key to make a longitudinal study feasible in practice – many participants remained in the study for over a year. Another important target for the client software design was to reach an appropriate trade-off between quality of the collected data and phone energy consumption.

The collected data was first stored in the device and then uploaded automatically to a Simple Context server via WLAN. The server received the data, and built a database that could be accessed by the campaign participants. The Nokia Simple Context backend had been developed already earlier by the Nokia Research Center in Palo Alto. Additionally data visualization tool was developed which offered a "life diary" type of view for the campaign participants on their data. Simultaneously, an anonymized database was being populated, from which researchers were able to access the data for their purposes. Fig. 1 presents a block diagram of the collection architecture.

2.2 Data characteristics

The LDCC initiative produced a unique data set in terms of scale, temporal dimension, and variety of data types. The campaign population reached 185 participants (38% female, 62% male), and was concentrated on young individuals (the age range of 22-33 year-old accounts for roughly two thirds of the population.) A bird-eye's view on the LDCC in terms of data types appears in Table 1. As can be seen, data types related to location (GPS, WLAN), motion (accelerometer), proximity (Bluetooth), communication (phone call and SMS logs), multimedia (camera, media player), and application usage (user-downloaded applications in addition to system ones) and audio environment (optional) were recorded. The

Data type	Quantity
Calls (in/out/missed)	240,227
SMS (in/out/failed/pending)	175,832
Photos	37,151
Videos	2,940
Application events	8,096,870
Calendar entries	13,792
Phone book entries	45,928
Location points	26,152,673
Unique cell towers	99,166
Accelerometer samples	1,273,333
Bluetooth observations	38,259,550
Unique Bluetooth devices	498,593
WLAN observations	31,013,270
Unique WLAN access points	560,441
Audio samples	595,895

Table 1: LDCC main data types and amounts for each type.

numbers themselves reflect a combination of experimental design choices (e.g., every user had the same phone and data plan) and specific aspects of the volunteer population (e.g., many participants use public transportation).

Due to space limitations, it is not possible to visualize multiple data types here. A compelling example, however, is presented in Fig. 2, which plots the raw location data of the LDCC on the map of Switzerland for the volunteer population after 1 week, and then after 1, 3, 6, 12, and 18 months of campaign. When seen in detail, the geographical coverage of the LDCC allows a reasonable tracing of the main routes on the map of Suisse Romande – French-speaking, western part of Switzerland – and gradually also of other regions of the country.

In addition to contributing phone data, participants of the LDCC also agreed to fill a small number of surveys during the data recording process. We would like to highlight two types of survey data which were important for the later development of the MDC - (1) a set of manual semantic labels for frequently and infrequently visited places for each user and (2) basic demographic attributes. The relevant places were first detected automatically with a method discussed in [15]. After that the campaign participants specified place categories from a fixed list of tags (home, work, leisure places, etc.). In sense of demographics, participants self-reported their attributes like gender, age group, marital status and job type etc.

2.3 Privacy

Privacy played an essential role in the design and implementation of the LDCC, given the nature and scale of the data shared by the participants of the initiative. In order to satisfy the ethical and legal requirements to collect data while protecting the privacy of the participants, the LDCC research team implemented an approach based on multiple strict measures. The approach can be summarized as follows (more details can be found in [14]):

1. Communication with volunteers about privacy. Following Nokia's general privacy policy, we obtained written consent from each individual participating the LDCC. We explicitly stated that data would be collected for research purposes. All participants were informed about their data rights, including the right to access their own collected data and to decide what to do with it (e.g. to delete data entries if they opted to do so). The participants had also opportunity to opt-out at any moment.

2. Data security. The data was recorded and stored using best industry practices in this domain.

3. Data anonymization. By design, the LDCC did not store any content information (e.g. no photo files or message content were recorded). The major portion of the collected data consisted of event logs, and when sensitive data beyond logs was collected, it was anonymized using state-of-the-art techniques and/or aggregated for research purposes [5]. Examples include the use of pseudonyms instead of identifiable data and the reduction of location accuracy around potentially sensitive locations. The researchers do have access only to the anonymized data.

4. Commitment of researchers to respect privacy. Privacy protection of such a rich data only by automatic anonymization techniques is not possible so that research value and richness of the data can be simultaneously maintained. In addition to technical means also agreement based countermeasures are necessary. Trusted researchers have been able to work with the LDCC data after agreeing in written form to respect the anonymity and privacy of the volunteering LDCC participants. This practically limited the access to the LDCC data to a small number of authorized partners and their affiliated researchers. After our initial experience with the LDCC, the next step was to outreach the mobile computing community at large, which motivated the creation of the Mobile Data Challenge, discussed in detail in the next sections.

3. MDC TRACKS

MDC's original intention was to be inclusive at a global scale. Other previously successful evaluation initiatives in computing, like those organized by NIST in several areas [16, 18] or the Netflix challenge [8, 7] focused on either one or at most a small number of tasks with objective evaluation protocols. This was also a guiding principle for MDC. On the other hand, the nature of mobile data is highly exploratory, so there was a clear benefit in encouraging and welcoming also open ideas.

Learning from these past experiences, we decided that MDC would feature both open and pre-defined options to participate. The *Open Track* was defined to receive self-defined ideas from the community. On the other hand, the concrete options were given in the *Dedicated Track*, which defined three classification/prediction tasks. These tasks covered several key aspects of mobility and mobile users.

The Open Track. This track enabled participants to propose their own Challenge task based on their own research interests and background. Examples proposed to the participants included the discovery of behavioral patterns through

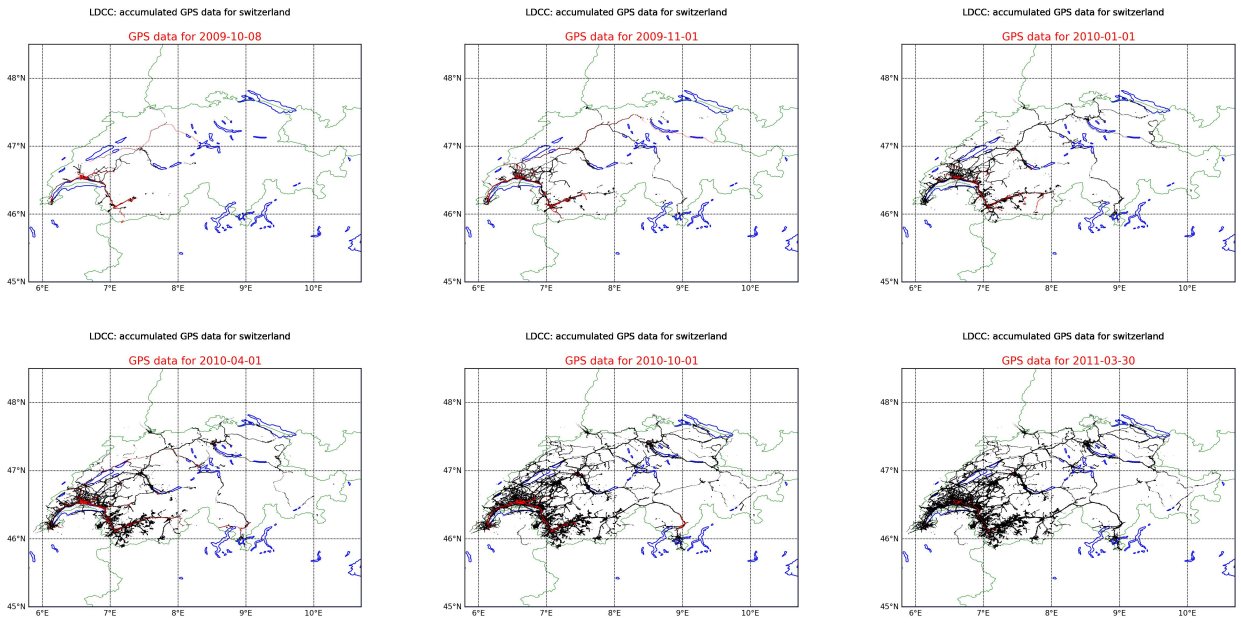

Figure 2: LDCC location data (in black) plotted at the country level (outlined in green) after 1 week, 1 month, 3 months, 6 months, 12 months, and 18 months of campaign. The data for each specific day is plotted in red.

statistical techniques, the development of efficient mobile data management methods, or the design of ways to visualize mobile Big Data.

The Dedicated Track. This Track gave the possibility of taking up to three concrete tasks to solve, with properly defined training and test sets, and evaluation measures used to assess and rank all the contributions. The participants of the Dedicated Track were allowed to define their own features and algorithms. The three tasks of this Track followed a two-stage schedule. In the first stage, the training set (including raw data, labels, and performance measures) was made available to the participants, who were expected to design their features and train their models using this data set. In the second stage, the test set was made available, in which labels were kept hidden. Participants were allowed to submit up to five runs of results, and the evaluation of all methods was meant to be conducted by the MDC organizers. The three tasks proposed were the following:

1. Semantic place prediction. Inferring the meaning of the most significant places that a user visits frequently is a relevant problem in mobile computing [13]. The goal of the task was to predict the semantic meaning of these places for a number of users of the MDC data. Each place was represented by a history of visits over a period of time, for which other contextual information sensed by the user's smartphone was available. Participants needed to extract relevant features for predicting the semantic labels. Good methods for this task were needed, given the particular type of input information (sequence of visits as opposed to geographic location). Importantly, it was decided that geo-location was not provided as a feature for this task for privacy reasons, as some of the place categories are privacy-sensitive (home, work, etc.). On the other hand, several other types of phone data were provided as features (see next section). Semantic place labels (manually provided by the LDCC users through surveys as discussed in Section 2) were given as part of the training set for this task.

2. Next place prediction. Predicting the location of phone users has a key relevance for context-aware and mobile recommendation systems [10]. The goal of this task was to predict the next destination of a user given the current context, by building user-specific models that learn from their mobility history, and then applying these models to the current context to predict where the users go next . In the training phase, the mobility history of each user was represented by a sequence of visits to specific places, and several types of phone data associated with these visits were made available (see next Section). Furthermore, in the testing phase, previously unseen data from the same set of users was provided, with the goal of predicting the next place for each user given their current place or a short history of places.

3. Demographic attribute prediction. The goal of this task was to infer basic demographic groups of users based on behavioral information collected from the phones. As discussed in Section 2, some of the voluntarily-provided demographic information in the LDCC included self-reported gender, age group, marital status, job type, and number of people in the household. This information was provided for training, and kept hidden for testing. Three subtasks, namely gender, marital status, and job prediction were formulated as classification problems, for which classification accuracy was used as evaluation measure. The two remaining attributes corre-

sponded to regression problems, for which the root mean square error (RMSE) was used as evaluation measure. Each subtask contributed equally to the final score which was defined as the average of relative improvements over baseline performance.

4. MDC DATA

This section presents an overview of the MDC datasets and the corresponding preparation procedures. We first describe the division of the original LDCC data that was needed in order to address the different MDC tasks. We then summarize the data types that were made available. We finalize by discussing the procedures related to privacy and data security.

4.1 Division of the dataset

The datasets provided to the participants of the MDC consist of slices of the full LDCC dataset. Slicing the data was needed in order to create separate training and test sets for the tasks in the Dedicated Track, but was also useful to assign the richest and cleanest parts of the LDCC dataset to the right type of challenge. Four data slices were created for the MDC:

Set A: Common training set for the three dedicated tasks.

Set B: Test set for demographic attribute and semantic place label prediction tasks.

Set C: Test set for location prediction task.

Open set. Set for all open track entries.

The overall structure of the datasets is given in Figure 3. The rationale behind this structure was the following. First, the participants of the LDCC were separated in three groups, according to the quality of their data according to different aspects. The 80 users with the highest-quality location traces were assigned to sets A and C. Set A contains the full data for these users except the 50 last days of traces, whereas set C contains the 50 last days for which location data is available for testing.

In order to maximize the use of our available data, we reused Set A as a training set for the two other dedicated tasks. A set of 34 further users was selected as a test set for these tasks and appeared as Set B. In this way, models trained on the users of Set A can be applied to the users of their most visited locations.

Demographic data and semantic labels, as explained in Section 2, were collected through surveys that the LDCC participants were asked to fill in. Since all steps of the LDCC participation were fully voluntary, a number of users chose not to complete surveys, or filled them only partially. However, this information was crucial for two of the three dedicated challenges. Therefore, the participants for whom complete questionnaire data was not available were assigned to the last set, which was used for the Open Track. In total, 38 users were assigned to this dataset.

Overall, with this data split, a total of 152 LDCC participants were included in the MDC datasets.

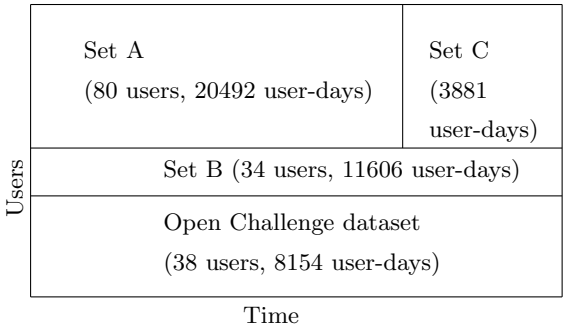

Figure 3: Division of the MDC dataset into four challenge subsets. For each set, the total number of user-days with data is also shown.

4.2 Data types

For both Open and Dedicated Tracks, most data types were released in a raw format except a few data types that needed to be anonymized. There are two main differences between the Open Track data and the Dedicated Track data. First, the physical location (based on GPS coordinates) was available in the Open Track but not in the Dedicated Track. Instead, we released a preprocessed version of the location data in the form of sequences of visited places for the Dedicated Track. This allowed to study performance of algorithms in location privacy-sensitive manner. The second main difference was in the availability of relational data between users. This included both direct contacts (e.g., when a user calls another user) and indirect contacts (e.g., if two users observe the same WLAN access point at the same time then they are in proximity). We decided to keep this data in the Open Track but removed it in the Dedicated Track since it could have potentially revealed the ground truth to be predicted. In the anonymization algorithm, a common encryption password was used for the users selected to the Open Track data sets. On the other hand, we used a different password for each user in the Dedicated Track.

Common data types. Each data type corresponds to a table in which each row represents a record such as a phone call or an observation of a WLAN access point. User IDs and timestamps are the basic information for each record. Specific information of each data type is detailed in Table 2.

Data types for Open Track only. Geo-location information was only available in the Open Track. In addition to GPS data, we also used WLAN data for inferring user location. The location of WLAN access points was computed by matching WLAN traces with GPS traces during the data collection campaign. The description of geo-location data is reported in Table 3.

Location data in Dedicated Track. Physical location was not disclosed in the Dedicated Track. For each user in the dedicated track data, the raw location data (based on GPS and WLAN) was first transformed into a symbolic space which captures most of the mobility information and excludes actual geographic coordinates. This was done by first detecting visited places and then mapping the sequence of coordinates into the corresponding sequence of place visits

data type	description
accel.csv	user ID, time, motion measure, and accelerometer samples.
application.csv	user ID, time, event, unique identifier of the application, and name of the application.
bluetooth.csv	user ID, time, first 3 bytes of MAC address, anonymized MAC address, anonymized name of the Bluetooth device.
calendar.csv	user ID, time, entry ID, status (tentative/confirmed), entry start time, anonymized title, anonymized location, entry type (appointment/event), entry class (public/private), last modification time of the entry.
callog.csv	user ID, call time, call type (voice call/show message), SMS status (delivered, failed, etc.), direction (incoming, outgoing, missed call), international and region prefix of phone number, anonymized phone number, indicator if number is in phone book, call duration.
contacts.csv	user ID, creation time, anonymized name, international and region prefix of phone number, last modification time.
gsm.csv	user ID, time, country code and network code, anonymized cell id, anonymized location area code, signal strength.
mediaplay.csv	user ID, time, album name, artist, track, track title, track location, player state, track duration.
media.csv	user ID, record time, media file time, anonymized media file name, file size.
process.csv	user ID, record time, path name of running process.
sys.csv	user ID, time, current profile (normal, silent, etc.), battery level, charging state, free drive space, elapsed inactive time, ringing type (normal, ascending, etc.), free ram amount.
wlan.csv	user ID, time, first 3 bytes of MAC address, anonymized MAC address of WLAN device, anonymized SSID, signal level, channel, encryption type, operational mode.

Table 2: Common data types of Open and Dedicated Tracks (in alphabetical order).

data type	fields
wlan_loc.csv	user ID, time, first 3 bytes of MAC address, anonymized MAC address, longitude, latitude.
gps.csv	user ID, record time, time from GPS satellite, geo-location (altitude, longitude, latitude), speed, heading, accuracy and DOP, time since GPS system started.

Table 3: Specific data types for the Open Track.

(represented by a place ID). Places are user-specific and are ordered by the time of the first visit (therefore, the visit sequence starts with place ID=1). Each place corresponds to a circle with 100-meter radius. Although the absolute coordinates of places were not provided, a coarse distance matrix between places was computed for each user and provided for the MDC participants of this track.

4.3 Data anonymization

Various anonymization techniques were applied to the MDC data: truncation for location data, and hashing of phone numbers, names (such as contacts, WLAN network identifiers, Bluetooth device identifiers), and MAC addresses. This process is summarized in this subsection.

4.3.1 Anonymizing location data

The detailed locations can indirectly provide personally identifiable information, therefore risking privacy of the LDCC participants. A location that is regularly used at night, for instance, could indicate the participant's address, which could then be reversed using public directories to find out the participant's identity. While all researchers participating in the MDC committed in writing to respect the privacy of the LDCC participants, i.e. not trying to reverse-engineer any private data (see Section 5), we also took specific measures in terms of data processing.

Anonymizing location data for Open Track. In order to provide enough privacy protection, while simultaneously keeping the data useful, we applied k-anonymity by truncating the location data (longitude, latitude) so that the resulting location rectangle, or anonymity-rectangle, contains enough inhabitants. For instance, in city centers anonymity-rectangles tend to be very small, while in rural areas anonymity-rectangles can be kilometers wide. This step required a considerable amount of manual work that included visualizing the most visited places of the LDCC participants in order to correctly set the size of the anonymity-rectangles. Once set, those anonymity-rectangles were applied to all data from all users.

The data for the Open Track included also the WLAN based location information which was passed through the anonymity-rectangle filtering similarly as defined above.

Location data for Dedicated Track. As discussed earlier, geo-location data was not used for the Dedicated Track. For the next place prediction task, locations were labeled with one out of ten possible semantic categories, intrinsically removing all personally identifiable information. We used the following categories: home; home of a friend, relative or colleague; workplace/school; place related to transportation; workplace/school of a friend, relative or colleague; place for outdoor sports; place for indoor sports; restaurant or bar; shop or shopping center; holiday resort or vacation spot.

4.3.2 Anonymizing MAC addresses, phone numbers, and text entries

Finally, hashing was applied to a variety of text entries appearing in the MDC data, including Bluetooth names, WLAN network identifiers (SSID), calendar titles and event

locations, first names and last names in the contact lists and media filenames (such as pictures).

For anonymization of the WLAN and Bluetooth MAC addresses, we split them into two parts. First, the MAC prefix, also known as the "Organizationally Unique Identifier (OUI)" [3], was kept in clear text. Second, the rest of the MAC address was anonymized by hashing, after concatenating it with secret key, and the userID for dedicated challenges.

$hash(token) = sha256(token)$, where,
$token = (seckey1||information||seckey2)$, for open challenges,
$token = (userID||seckey1||information||seckey2)$, for dedicated challenges

Note that, for the dedicated challenges, this anonymization method results in the same MAC address appearing differently in different user data sets.

Phone numbers appearing in the call logs and in contact lists were also split in two parts. First, the number prefix, which contains the country and region/mobile operator codes, was left as clear text. Then, the rest of the phone number was hashed as described above. Also the cell ID and the location area code (LAC) of the cellular networks were anonymized using the hashing technique as described above.

4.4 Watermarking

The release of the MDC data set to a large community of researchers motivated an additional step in which each distributed copy of the data set was watermarked individually in order to identify it if necessary. The watermarking process introduced negligible alteration of the data that did not interfere with the results.

5. MDC SCHEDULE

The plans to organize MDC started in summer 2011. We targeted to organize the final MDC workshop within one year. We decided to keep the challenge open for all the researchers with purely academic affiliation. The prospective participants of the Open Track had to submit a short proposal with their concrete plan, and the participants of the Dedicated Track had to agree to participate at least one task. While the MDC was by nature open, a series of important steps were established for participant registration. Importantly, this included signature of the Terms and Conditions agreement. In that manner each researcher explicitly committed to treat the data only for research purposes, and to use the data in an ethical and privacy-respective manner (for instance, reverse engineering any portion of the MDC data to infer sensitive personal information was strictly forbidden).

The MDC registration process was launched in early November 2011 and closed in mid-December 2011. The challenge was received enthusiastically by the research community. In early January 2012, the MDC data was released to more than 500 individual participants as individually watermarked copies for more than 400 challenge tasks. The participants were affiliated with hundreds of different universities and research institutes, with a worldwide geographical distribution (Asia 23%, USA 22%, Europe 51%, other regions 4%). Many leading universities in the field participated in the two tracks of the challenge.

A total of 108 challenge submissions were received on April 15, 2012, corresponding to 59 entries for the Dedicated Track and 49 entries for the Open Track. All submitted contributions were evaluated by a Technical Program Committee (TPC), composed of senior members of the mobile and pervasive computing communities. The TPC members did not participate in the MDC themselves to minimize possible conflicts of interest. The complete list of TPC members can be found in the MDC proceedings' front matter [4].

The criteria to select entries for each Track were different. On one hand, the Open Track entries were evaluated according to a set of standard scientific criteria, including the novelty and quality of each contribution, and the paper presentation. All entries in the Open Track were reviewed at least by two members of the TPC. On the other hand, all entries in the Dedicated Track were evaluated using the objective performance as the only criterion to decide on acceptance to the Workshop. Entries of all three dedicated tasks were compared against a standard baseline method. In addition to this, also all Dedicated Challenge papers were also subject to review by the TPC in order to verify basic principles of originality, technical novelty, experimental correctness, and clarity. Papers corresponding to entries whose performance did not outperform the baseline were reviewed by one member of the TPC. All other papers were reviewed at least by two members of the TPC. The TPC evaluated all papers without knowledge of the performance obtained on the test set, and reviewed them based on their originality, quality, and clarity. While the reviews did not play any role in the acceptance decision for the Dedicated Track, they helped to detect a few problems, and in every case they were passed on to the authors. In particular, the TPC reviews served as guidelines to the authors of accepted entries to improve the presentation of their approach and achieved results. Final acceptance for all entries was decided during a face-to-face meeting involving all MDC co-chairs, in which all papers were discussed and in some cases additional reviews were appointed. In some cases, a shepherd was assigned to accepted entries to ensure that the key comments from the reviewers were implemented. As a result of the reviewing process, 22 entries to the Open Track and 18 entries to the Dedicated Track were accepted, resulting in an overall acceptance rate of 37%. For the Dedicated Track, we decided not to reveal the teams' absolute performance scores and relative ranking before the MDC Workshop.

Finally, as a way of acknowledging the top MDC contributions, a number of awards will be given, based on the entries' performance for the Dedicated Track, and following the recommendations of the Award Committee appointed for the Open Track. The Award Committee list can be found in the workshop proceedings' front matter [4].

6. CONCLUSIONS

This paper described a systematic flow of research over 3.5 years at the time of writing, targeting to create and provide unique longitudinal smartphone data set for wider use by the research community. In this paper we gave motivation for

this initiative and summarized the key aspects of the Lausanne Data Collection Campaign (LDCC) in which the rich smartphone data was collected from around 200 individuals over more than a year. We also described in further details the Mobile Data Challenge (MDC) by Nokia which was a data analytics contest making this data widely available to the research community. The data collection campaign was running in 2009-2011 whereas the challenge was organized in 2011-2012.

Collecting such data requires extensive effort and underlying investments, which often means that collected data sets are available for researchers only in the limited manner. This has recently generated some discussion about the basic principles of science in connection with the Big Data driven research. Verification of some claimed scientific findings can namely be challenging if access to the underlying data is very limited. Protecting privacy of individuals behind the data is obviously the key reason for access and usage limitations of Big Data.

With the examples described in this paper we demonstrated that open data sharing with the research community and therefore wider open innovation momentum around the same commonly available data set is possible. Achieving that requires proactive and holistic approach on privacy throughout the whole research flow. Privacy protection requires extremely careful considerations due to multimodality of the rich smartphone data. In this paper we described the needed countermeasures both when the smartphone data was originally collected and when it was later released to the research community as a part of the MDC. In practice this required both technical countermeasures and agreement based privacy protection. In that manner it was possible to achieve appropriate balance between the necessary privacy protection but simultaneously still maintaining the richness of the data for research purposes.

Already so far the Mobile Data Challenge has produced interesting findings and multidisciplinary scientific advances. The contributions to the MDC addressed various interesting angles from the perspective of mobile computing research, like investigations on predictability of human behavior patterns or opportunities to share/capture data based on human mobility, visualization techniques for complex data as well as correlation between human behavior and external environmental variables (like weather). The materials presented in the MDC workshop are available in [4]. Momentum around the interesting research resources described in this paper is expected to continue and expand also after the Mobile Data Challenge itself. Our intention is to maintain summary at least of the most important LDCC and MDC originated research outcomes in [4] also in the future.

7. ACKNOWLEDGMENTS

We sincerely thank all the volunteers in the LDCC initiative for their participation and contributed data, and all the researchers who responded to our open invitation to sign up and participate in the MDC. We also thank Niko Kiukkonen for his key role with the arrangements of the Lausanne Data Collection Campaign. David Racz's pioneering work with the Nokoscope and Simple Context data collection systems is also gratefully acknowledged. Additionally we thank Emma Dorée, Antti Rouhesmaa and various other people from Nokia organization for their contributions to the arrangements of the Mobile Data Challenge.

8. REFERENCES

[1] `http://crawdad.cs.dartmouth.edu/`.
[2] `http://privacybydesign.ca/`.
[3] `http://en.wikipedia.org/wiki/MAC_address`.
[4] `http://research.nokia.com/mdc`.
[5] I. Aad and V. Niemi. NRC Data Collection and the Privacy by Design Principles. In *PhoneSense*, 2011.
[6] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proc. World Wide Web Conf. (WWW)*, Apr. 2010.
[7] R. Bell, J. Bennett, Y. Koren, and C. Volinsky. The million dollar programming prize. *Spectrum, IEEE*, 46(5):28–33, 2009.
[8] J. Bennett and S. Lanning. The netflix prize. In *Proceedings of KDD Cup and Workshop*, volume 2007, page 35, 2007.
[9] G. Chittaranjan, J. Blom, and D. Gatica-Perez. Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, published online Dec. 2011.
[10] T. Do and D. Gatica-Perez. Contextual conditional models for smartphone-based human mobility prediction. In *Proc. ACM Int. Conf. on Ubiquitous Computing*, Pittsburgh, Sep. 2012.
[11] N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.
[12] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.
[13] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Identifying important places in people's lives from cellular network data. In *Proc. Int. Conf. on Pervasive Computing*, San Francisco, Jun. 2011.
[14] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. In *Proc. Int. Conf. on Pervasive Services*, Berlin, Jul. 2010.
[15] R. Montoliu and D. Gatica-Perez. Discovering human places of interest from multimodal mobile phone data. In *Proc. Int. Conf. on Mobile and Ubiquitous Multimedia*, Limassol, Dec. 2010.
[16] M. Przybocki and A. Martin. Nist speaker recognition evaluation-1997. In *Proceedings of RLA2C*, pages 120–123, 1998.
[17] P. Ross. Top 11 technologies of the decade. *Spectrum, IEEE*, 48(1):27–63, 2011.
[18] A. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 321–330. ACM, 2006.