Towards a Federated Repository of Mobile Sensing Datasets for Pervasive Healthcare

Jesus Favela
CICESE
Ensenada, Mexico
+52 (646) 175-0500 x2309
favela@cicese.mx

Luis A. Castro
Sonora Institute of Technology
Ciudad Obregon, Mexico
+52 (644) 410-9000 x1523
luis.castro@acm.org

Layla Michan
CICESE
Ensenada, Mexico
+52 (646) 175-0500 x2300
layla.michan@gmail.com

ABSTRACT
Mobile sensing is becoming a popular approach to infer patterns of activities and behavior to determine how they affect health and wellbeing. This data-driven approach to discovery has the potential to become a major tool in the field of epidemiology, aimed at determining the causes of disease in populations. These studies have motivated the creation of datasets with information opportunistically gathered from sensors in mobile devices. The nature of this data gathering effort raises a number of issues, such as the heterogeneity of the devices and sensors used, which hamper information sharing and integration needed to conduct longitudinal studies and validate and construct over previous results as new data becomes available and algorithms are improved. This paper proposes the development of an open access federated repository of datasets for preservation and sharing. We propose a process that involves data curation and integration into a unified schema from which researchers can query and use diverse dataset for comprehensive studies which could, for instance, compare two populations sensed in different periods and with somewhat different conditions.

CCS Concepts
• Information systems → Information systems applications → Mobile information processing systems.

Keywords
Mobile sensing; Dataset repository; Data curation.

1. INTRODUCTION
Smartphones and wearable technology incorporate advanced computing and sensing capabilities that enable collecting data related to their users and their surroundings. The augmented capabilities of smartphones and wearables, and their ubiquity have contributed to an emerging field known as mobile sensing. This area aims at collecting and analyzing data from several devices scattered around a particular geographical area worn or carried by users. This emerging area is of particular interest for epidemiology researchers as it provides an unusual lens for better understanding human behavior such as disease outbreaks, disease onsets, medical care, and health status and outcomes.

Mobile sensing uses a data-driven paradigm of scientific discovery through data. Thus, the body of knowledge generated in this field is profoundly associated to the data captured by these devices. Facilitating the preservation and sharing of these datasets is paramount to the growth and maturity of the field.

Data collection itself is an important task for data-driven science but it may require the participation of hundreds of volunteers to contribute (and interpret) data through participatory sensing campaigns. This is referred in other fields to as citizen science or crowd-sourced science, which is scientific research conducted, in whole or in part, by amateur or nonprofessional scientists. At times, this turns out to be fundamental for science, as the amount of unstructured data to be analyzed can be overwhelming for scientists and current infrastructure for supporting analysis.

Even when analysis could potentially be carried out by civic scientists, in mobile sensing, the main concern at the moment is generating reliable datasets that can be used to push forward the boundaries of the area, which necessarily involves providing structure to the data. While current efforts have generated significant results, our vision is that scientists could tap into a distributed repository of datasets that have been curated and integrated to facilitate conducting new research such as generalizing previous findings when comparing with new data from a different population and creating longitudinal studies controlling for the conditions in which data was gathered over long periods of time.

This position paper proposes a federated repository of datasets of mobile sensing data for healthcare. Achieving this goal will require a collaborative effort to define a common evolving schema and data curation standards to guarantee the provenance as the datasets being incorporated into the repository. In the next section, we describe some of the current issues that need to be tackled for data sharing and integration.

2. RELATED WORK
Efforts to create repositories of datasets are not new, the machine learning community has made available numerous repositories, including some in pervasive healthcare, aimed at detecting activities of daily living [3]. Examples of these repositories include the UCI Machine Learning Repository [5] and Keel [1]. These datasets, however, are not consolidated. They are meant to be used independently to test new learning algorithms or to be used by students as a learning tool.

Closer to our aim are efforts such as re3data (Registry of Research Data Repositories, http://www.re3data.org), Dryad with data from papers (http://www.datadryad.org), DataOne (http://www.dataone.org), a repository on Earth and environmental data, a repository of datasets for ecological

PervasiveHealth 2016, May 16-19, Cancun, Mexico

DOI 10.4108/eai.16-5-2016.2263872

modeling and forecasting [2]. Finally, also Synapse (http://www.synapse.org), which is an open source software platform aimed at data scientists for research in human health, however the platform is not necessarily designed for mobile sensing.

Within the domain of healthcare, the Observational Health Data Sciences and Informatics (OHDSI) is an open-source initiative hosted at Columbia University with the purpose of standardizing observational data and analytics of healthcare data [7]. This is achieved through a common data model and a series of tools. OHDSI however uses patient registries rather than data obtained from mobile sensors.

The Open mHealth (http://www.openmhealth.org/) initiative recognizes the importance of mobile sensing data for healthcare research and is developing mechanisms to standardize data and integrate data streams to encourage the development of new pervasive healthcare apps. It also offers an Application Program Interface (API) to access data from a repository and to conduct data analytics. While the effort to standardize data from different sensors is encouraging, the platform is meant for applications for the end user, rather than as a research platform.

MD2K [4] is a Center for Excellence for Mobile Sensor Data-to-Knowledge funded by the National Institutes of Health (NIH) with the purpose of facilitating gathering, analyzing and interpreting health data generated by mobile and wearable sensors. The current effort is focused on two application domains with a long-term vision of acting as a repository of mobile data using Open mHealth standards.

3. ISSUES IN MOBILE SENSING FOR DATA SHARING AND INTEGRATION

Mobile sensing faces a number of issues related to data gathering and sharing, some of which are similar to those found in other areas, while others differ in numerous ways. We next describe some of the issues currently faced by this research community in order to achieve the goal of having a data infrastructure from which to continuously construct new knowledge and validate previous findings.

Heterogeneity in data gathering. While mobile phones have become the most popular means for gathering mobile data, there are important differences in the manner in which information is collected. The sensors available in each device might have different accuracies or the data might be sampled at different frequencies due to the restrictions of each sensing campaign. For instance, to avoid battery depletion, it might be decided to record location information once every five minutes, or due to privacy, to record it only Monday to Friday from 9 to 5pm. As mobile devices evolve and new and more powerful sensing and storage capacity becomes available we might face the issue of how to compare recordings from the same individual who over the years has used different recording devices (i.e. from a Fitbit device, to a smartphone, to a smartphone+smartwatch, to intelligent clothing). One relatively straightforward alternative is to turn to the lowest common denominator but it has the inconvenience of losing the advantage provided by richer data. Another option is to interpolate data in datasets with more sparse information and consider the uncertainty raised by this during data analysis.

Heterogeneity in sensor data. Due to the great diversity in the market of mobile phones and chipset vendors, sensor data generated by mobile devices can greatly vary. For instance, the wireless Received Signal Strength Indicator (RSSI), which is a measurement of the power present in a received radio signal, is commonly used for inferring indoor location. Depending on the chipset vendor, the RSSI may vary as they use different RSSI_Max value. This can lead to different measures or samples even in the same network. Mobile operating systems such as Android or Tizen tend to "hide" many of these details, but these challenges can be more acute if off-the-shelf sensors are used in a sensing campaign. Heterogeneity remains an issue since there are many actors involved, but for the purposes of this area, the peculiarities of the hardware used for creating/collecting the data should be made public.

Data pre-processing. Some sensors such as microphone, accelerometer or camera have the potential to generate considerable amounts of data. For instance, one minute of uncompressed audio at 44 kHz requires approximately 5MB of storage capacity. If continuous audio is recorded from several subjects for days or weeks it quickly becomes impractical to store all this information. Besides, specific studies might be interested in detecting only some events, for instance, when the individual is involved in a conversation, or detecting that she is outdoors. In addition, location or audio data might be processed to preserve the anonymity of the subject. The consent agreement might specify, for instance, that the audio recorded will be processed so that the content of the conversation or the identity of the speakers could not be inferred. Thus, it is a common practice to process the data gathered in a manner that some information is lost. If years later, a new study wants to use these findings, it needs to be aware of how the data was processed. As an example, speaker identification in a legacy study might have 93% accuracy, compared to 97% for a new study, with which we want to conduct an integrated analysis. Since the original data from the legacy study are not available, the new algorithms cannot be run on that dataset. The researcher needs to be made aware of these differences, and, if possible, be provided with tools to take them into account during the analysis.

Privacy. Due to their nature, many personal data are considered sensitive which can include location, behavior, or personal health records. Other data, on the other hand, can be disclosed without much concern. In the case of personal data, some people may be willing to share them only if authorized people access them (and analyze them). For instance, individuals can be willing to share health records with authorized health care professionals, but not with certain government agencies. For this, mechanisms such as proper anonymization of data must be paramount, releasing data with certain restrictions (data license for commercial or research use), or providing some sort of private/public keys to access data.

Ethics. Making sure that participants understand the prolonged validity of the data they are contributing is paramount, as the data can be used in future studies. Consent forms should be handled in such a way that participants are fully aware of the study they are participating in. A recent example of how remote informed consents are handled is through Apple's Research Kit (http://researchkit.org/), in which participants sign through the mobile phone.

Data annotation. Due to the naturalistic conditions in which mobile sensing studies are conducted, annotating data and establishing ground truth can require significant effort. Mobile sensing studies often rely on user participation to annotate the data, which might compromise its reliability. In addition, the annotation might be incomplete if the user is not always able or willing to provide the data.

Limited data sharing. While several research teams in mobile sensing have made their datasets available to other researches, this is not the norm, and each group uses different formats and metadata. This to a large degree reflects the novelty of the field but also the fact that the effort to make the data

available (adding metadata, put in adequate format, anonymization, etc.), might not be compensated.

These issues, along with others, have motivated the proposal for creating a repository of datasets of mobile sensing data, as explained next.

4. OPENSENSEDATA: A VISION OF A FEDERATED REPOSITORY OF MOBILE SENSING DATA

The vision of OpenSenseData is to make datasets of mobile sensing for healthcare available to researchers in a manner that facilitate their use through a distributed repository with a unified API to access the data. The goals of OpenSensaData are:

- Archiving and preserving data so that published results can be extended and validated.
- Facilitating data sharing for creating new knowledge.
- Providing a standard schema, thesaurus, and ontologies to facilitate data integration and understanding.
- Acting as a registry of published datasets to attest quality and confer proper credit to authors with pertinent credits and adequate licensees.

4.1 A scenario for OpenSenseData

The following scenario aims at explaining our vision for OpenSenseData:

An epidemiology researcher has preliminary evidence linking the regularity in the time at which individuals eat their meals in mid-life to medication adherence when they become older adults. She decides to consult OpenSenseData for datasets that could be used to obtain further evidence to strengthen her hypothesis. She queries the repository for datasets that have information on the time at which subjects eat with duration of at least 20 days. The search results in 47 datasets. She uses the provenance criteria in the repository to eliminate datasets from which the meal schedule data was derived and subject to error (for instance using audio recognition) and studies for which the subjects can no longer be contacted. This results in 19 datasets with a total 708 informants who will now be 65 or older. Using the tools provided by OpenSenseData, the researchers sends a short medication adherence survey to all participants of the selected 19 studies and waits for the results to complete her analysis. The researcher has no direct access to the contact information or other personal data of the informants. She submits an application to the repository to register the data collected as a new dataset, for which she will get credit.

4.2 Lifecycle of a research dataset

Figure 1 illustrates the lifecycle of a research dataset. To make this vision a reality, a number of technical and organizational issues need to be addressed in each of these phases.

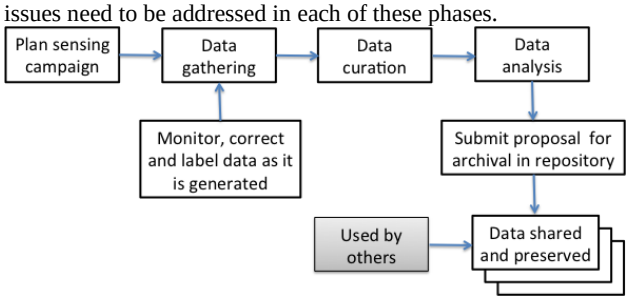

Figure 1 Lifecycle of a research dataset

Plan sensing campaign. From the moment the research is conceived, considerations for data sharing need to be taken into account. A preserved and shared dataset should be considered a research outcome in itself. This would involve selecting data and metadata standards and clearly defining the dataset pipeline, from information gathering to submission to a repository.

Data gathering and monitoring. Mobile sensing software has the capacity to monitor data as they are created so that the research team can label data or make corrective actions. For instance, it could detect that a subject may not be wearing her smartwatch during the day and thus label those data as not corresponding to daily use, or event contact the user to encourage het to wear the device. Tools need to be created to facilitate the semi-automatic monitoring of the data as they are created.

Data curation. Data curation is a labor-intensive task aimed at assuring the quality of the data before analysis. It involves tasks such as filling missing data, using data standards, add proper metadata to label the data, and, when appropriate, labeling uncertain data. This is a necessary step to guarantee that the data is trustworthy to preserve and share. It requires the development of metadata schemas and data standards for mobile sensing. It also involves enriching data with semantic notations and linked data. For this we can taka advantage of the ontologies developed in the domain of healthcare, such as MeSH or OpenClinical.

Data analysis. Information about how the data were derived and processed needs to be identified, to establish data provenance for posterior analysis. If raw data is not stored, the process to generate derived data needs to be clearly documented and identified for proper reference. Thus, research based on this derived data can refer to the original study (i.e. "from the average walking speed per subject derive in [ref to dataset], we estimated….".

Data preservation. Published results can be linked to data preserved in a repository, so that it can be validated and compared. As new tools become available to facilitate linking papers to their data, it might become mandatory to make the dataset public in a repository. Several issues hamper researchers from making their data sets available [6], if publishing a dataset is made as valuable as publishing a paper, researcher might be more motivated to do so. This will require data compatibility and interoperability and that the dataset is revised to guarantee quality and value. Thus, the repository might also work as a dataset registry. The research could choose one repository for storage (for instance in their own institution), but needs to go through the same clearing process and be searchable using the unified schema.

Data signature. Digital data can be easily manipulated. In scenarios where data are to be shared and distributed, certain users can be interested in making sure that the data they are sharing remain unaltered. Researchers, on the other hand, may want to make sure that the data they are analyzing is authentic. Ideally, data should be digitally signed for authenticity corroboration. That is, making sure that the data really come from the alleged source (e.g., adult in early sixties with diabetes, living in Mexico) and have not been modified after multiple distributions. Similar mechanisms exist in information security, where messages are authenticated to preserve data integrity from origin to destination, and no third parties have modified the message.

Using a dataset. A federated repository needs to provide a unified schema for users to query their content. This should include information about content, but also structural, descriptive, administrative, and metadata on how the information was derived. Semi-automatic tools can be used to facilitate the integration of a new dataset within the repository. This task can be performed by the researcher who submits the dataset for publication with the assistance of a data curator or dataset editor. Eventually these

repositories might work as a testbed for activity and behavior recognition algorithms.

5. CONCLUSIONS

In this paper, we propose the development of a federated repository of datasets with a process that involves data curation and integration into a unified schema from which researchers can query and use diverse dataset for comprehensive studies. This repository, we which have named OpenSenseData, could be used by researchers, for instance, to compare two populations sensed in different periods and with somewhat different conditions.

We believe that some of the issues that we have outlined merit careful consideration if the research in the area is to be pushed one step forward. Some of these issues have been previously discussed in other communities, but in mobile sensing for healthcare we are to still define an agenda that can help shape and lead the efforts of the area.

6. REFERENCES

[1] Alcalá, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., & Herrera, F. (2010). Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic and Soft Computing, 17(255-287), 11.

[2] Boose, E, A Ellison, L Osterweil, L Clarke, R Podorozhny, J Hadley, A Wise, and D Foster (2007) Ensuring reliable datasets for environmental models and forecasts. Ecological Informatics, 2(3):237–247

[3] Calatroni, A., Roggen, D., & Tröster, G. (2011). Collection and curation of a large reference dataset for activity recognition. In Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on (pp. 30-35). IEEE.

[4] Kumar, S., Abowd, G. D., Abraham, W. T., al'Absi, M., Beck, J. G., Chau, D. H., & Ganesan, D. (2015). Center of excellence for mobile sensor Data-to-Knowledge (MD2K). Journal of the American Medical Informatics Association, 22(6), 1137-1142.

[5] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[6] Nelson B. Data sharing: Empty archives. Nature. 2009 Sep 10;461(7261):160-3. doi: 10.1038/461160a.

[7] Voss, EA., Makadia, R., Matcho, A., Ma, Q., Knoll, C., Schuemie, M., DeFalco, FJ., Londhe, A., Zhu, V., Ryan, PB. (2015). Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. Journal of the American Medical Informatics Association. 22(3):553-64. doi: 10.1093/jamia/ocu023