Reza Rawassizadeh and David Kotz *Department of Computer Science, Dartmouth*

Editors: Aruna Balasubramanian and Lin Zhong

DATASETS FOR MOBILE, WEARABLE AND IoT RESEARCH

The advent of affordable devices with sensors and communication capabilities has led to the proliferation of computing paradigms, such as the Internet of Things (IoT), mobile devices, and wearable technologies. For the sake of simplicity, we use the umbrella term "small devices" for these technologies. At the same time, in the past decade, the increasing availability of large datasets has shifted scientists' attention toward data science, and defined new trends in computation. Even some scientists call it an evolutionary shift that has changed the pace of scientific progress, i.e., the "fourth paradigm" [1].

Photo, istockphoto.com

Many applications benefit from analyzing the collected data from sensors of small devices. These applications use machine learning, data mining or other data analytical methods to "extract knowledge" from the collected data. Usually, a developer who builds a data analytical method for an application needs to test her approach on test data to improve the algorithm and its feature-extraction processes through several iterations of testing and optimizing the code. Small devices are usually battery powered, i.e., disconnected from the power line. Moreover, because of their size, they have a limited battery, and thus a limited CPU [2]. This often means that the developer must conduct the resource-intensive training processes outside the device, the target device only collecting sensor data, extracting features, and (later) executing the trained model. Research on new methods, new models, and new applications often must begin with a data set – for initial exploration of a new idea, for validation of a proposed algorithm, or for benchmarking against existing approaches.

However, the process of data collection is not trivial. Those collecting such data have to deal with technical challenges, subject recruitment, subject privacy, and organizational bureaucracy. Because of these challenges, there are currently few sensor-based datasets available for broad use by researchers. Fortunately, some researchers have made relevant datasets available for use by other research teams. Below, we describe some of the existing datasets that are available for public access. In particular, we introduce the structure of datasets and provide a list of available datasets to developers and researchers who are developing algorithms for small devices. We categorize dataset repositories into two main categories: "standalone" and "multi-dataset repositories" repositories.

STANDALONE REPOSITORIES

Some scientific institutions create useful static datasets and use their own web page to share them with the community; we call

them standalone repositories. Examples include well-known datasets that collect data through smartphones and their sensors, such as Reality Mining [3] (including Bluetooth devices in the proximity, geographical locations with labels, call and SMS logs, running applications on the smartphone, and so forth), Mobile Data Challenge for Nokia [4], and Device Analyzer [5].

Reality Mining[1] was one of the first efforts to use Nokia smartphones, equipped with the Symbian operating system, for collecting smartphone data from a group of volunteer students. This dataset was collected in 2005 with early versions of the smartphones. Now smartphones play a very important role in our life and our smartphone usage patterns have been changed a lot. Mobile Data Challenge (MDC)[2] is another smartphone dataset collected again by the old Nokia smartphones in 2009-10, from volunteers in Lausanne, Switzerland. The dataset includes data from calls, SMS, Bluetooth proximities, apps and media usage, battery status, and acoustic environment information and location.

Livelab [6] is an in-device logger framework that is designed for iPhones. Similar to other platforms, its dataset includes data from calls, SMS, web history, accelerometer data, battery state, and so forth. Its dataset has been collected in 2010 from iPhone devices. Device Analyzer is the largest available smartphone dataset that has more than 30,000 contributors to date.[3] It collected detailed hardware information, such as operating system information from storage, power, audio volume, and user interaction with the phone with timestamp of the picture, calls, SMS, and so forth. The Device Analyzer dataset has been announced in 2014 and is still ongoing. A newer dataset, PhoneLab [7] platform,[4] has been released in 2015. It enrolls volunteers with a Nexus 5 and 6 Android smartphones and allows researchers to deploy experiments across subsets of the volunteers' smartphones – even enabling researchers to modify the Android platform. Livelab and Phonelab did not get deployed into the market. Therefore, their datasets are from enrolled participants.

Another, somewhat different, dataset in this category is Insight for Wear [8][5], which is the largest wearable (smartwatch) dataset available for public access. At this writing, it has more than 1,000 installs and has collected more than 10 million records from those volunteer users.

Some other well-known datasets, hosted in their own domain, are not focused on smartphones or smartwatches but do include data from small devices: idiap activity recognition dataset,[6] Smarthome sensor data of the UMass Trace Repository[7], T-Drive with one-week trajectories of Beijing Taxis,[8] or GeoLife,[9] which includes GPS trajectory of 182 individuals over three years.

Most of these datasets are provided as a static dataset, but two of the aforementioned sources (Device Analyzer and Insight for Wear) are data *streams*. Data streams are continuously generated by data sources. The data collection process is (continuously) ongoing and the new data is continuously released as part of the shared dataset. Device Analyzer provides a client that can be used to download a recent snapshot of the dataset. Insight for Wear is a smartwatch application, continuously in use by hundreds of volunteer subjects; this dataset is also streamed and continuously growing. When accessed, researchers can download a static snapshot of the most-recent version of the dataset.

MULTI-DATASET REPOSITORIES

Multi-dataset repositories host many datasets, often with a unifying theme. These repositories may host datasets based on a structural criterion (for example, their data are all time series), technical criterion (for example, all about wireless networks), or application domain (for example, health and wellness). Here we describe four well-known repositories.

UCI Machine-Learning Repository [9]: One well-known repository is the University of California Irvine machine-learning repository.[10] It hosts datasets from different domains, typically in CSV format. It encourages contributors to label data and attribute types, and to specify the area (target domain) of the dataset. This repository has several advantages, including flexibility in the donation of the data, diversity of data, and its multi-disciplinary portfolio. Some examples of this repository are human activity recognition extracts from smartphones, daily living activities, energy performance of a building's components and areas, and sports-related behaviors, e.g. running or exercising on a fitness device.

UCR TimeSeries Archive [10]: University of California Riverside provides a repository hosting datasets that represent time series.[11] Many interesting small-device sensors, such as accelerometers, naturally produce time-series data. This repository poses some limitations on the format of the dataset. For instance, each dataset in this repository comes in two parts; train and test. Nevertheless, these format policies make it easy for users who are interested in using a time series and offer some useful features such as sanity check. It also hosts datasets such as electrocardiogram data, robots' surface detection data, and insects' movements that have been measured via ubiquitous devices.

PhysioBank [11]: An important domain of applications for wearable and mobile devices relates to health and wellness improvement. Although wearable or mobile devices could be called new technologies, health-monitoring and the use of wireless devices (Holter monitors) have a long history. PhysioBank[12] is a rich source of psychological health-related signals. For example, there are datasets including electroencephalogram (EEG), electrocardiogram (ECG), gait data of human in different stages, and several other forms of multivariate time series of medical data.

CRAWDAD [12]: The Community Resource for Archiving Wireless Data At Dartmouth (CRAWDAD)[13] is focused on hosting data collected from wireless networks and mobile devices. It hosts datasets about the mobility of small devices and about the wireless networks they use. CRAWDAD's wireless and network data may be useful for research on mobility modeling, localization, crowd estimation, WLAN protocol and traffic modeling, Radio Frequency signal modeling, Mobile Ad-hoc Network simulations, and so forth.

Table 1 provides a brief summary on multi-dataset repositories that host datasets from small devices.

COMMERCIAL AND OTHER DATASET REPOSITORIES

There are a few other commercial sources that sell datasets from small devices, such as opensensors.io or crowdsignals.io. For instance, the crowdsignals.io initiative

TABLE 1. Multi-Dataset Repositories that host data from small devices.

Repository Name	Hosting Policy	Format Limitation
UCI Machine Learning Repository	Multi-purpose datasets, without any specific characteristics or attributes.	Prefers a CVS format, with some customized metadata.
UCR Time Series Archive	Only time-series datasets.	Strict data formats, such as providing both training and testing instances.
PhysioBank	Health-related data, primarily physiological data like EEG and ECG.	Limited to CSV format.
CRAWDAD	Data about mobile devices and/or wireless networks; sometimes, data about location or mobility of people who use mobile devices.	No limitation on data format.

intends to collect and share smartphone and smartwatch data. They have used crowd funding to provide incentives to their developers and to users who intend to contribute their data. Other well-known repositories with general datasets, including KDnuggets,[14] DRYAD,[15] Datahub,[16] and re3data.[17] are not primarily focused on hosting ubiquitous or small-device datasets, but some of their datasets may include small devices as well.

SUMMARY

In short, there are a variety of public datasets that can benefit the Ubicomp and Mobile Computing community. Usually, finding a dataset to test or evaluate an algorithm is time consuming, so we hope that our summary of available datasets and their characteristics will assist those developers and researchers.

[1] http://realitycommons.media.mit.edu
[2] https://www.idiap.ch/dataset/mdc
[3] https://deviceanalyzer.cl.cam.ac.uk
[4] https://phone-lab.org
[5] http://insight4wear.com
[6] http://www.ife.ee.ethz.ch/research/groups/Dataset
[7] http://traces.cs.umass.edu
[8] http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13
[9] https://www.microsoft.com/en-us/download/details.aspx?id=52367
[10] http://archive.ics.uci.edu/ml
[11] http://www.cs.ucr.edu/~eamonn/time_series_data
[12] https://www.physionet.org/physiobank
[13] http://crawdad.org
[14] http://www.kdnuggets.com/datasets/index.html
[15] http://datadryad.org
[16] http://datahub.io
[17] http://www.re3data.org

Indeed, since the benefits of publicly available datasets are so apparent, we encourage our colleagues in both academic and industry, both research and development, to create and share datasets about small devices and their users. With appropriate safeguards for the anonymity of any human volunteers in the dataset, these data will be a tremendous resource for science. Many of the above repositories provide an excellent and convenient venue for sharing your data. ■

Acknowledgement

We appreciate Tristan Henderson and Aruna Balasubramanian's useful hints, feedback and dataset links that significantly enriched the quality of this paper.

Reza Rawassizadeh is a research associate with the Department of Computer Science at Dartmouth College. He received his PhD from University of Vienna in 2012. His research interests include wearable computing, Internet of Things and data mining. For more information, visit http://www.cs.dartmouth.edu/~reza/

David Kotz is the Champion International Professor in the Department of Computer Science at Dartmouth College. His research interests include security and privacy, pervasive computing for health care, and wireless networks. After receiving his A.B. in Computer Science and Physics from Dartmouth in 1986, he completed his Ph.D in Computer Science from Duke University in 1991 and returned to Dartmouth to join the faculty. He is an IEEE Fellow, a Senior Member of the ACM, a 2008 Fulbright Fellow to India, and an elected member of Phi Beta Kappa. For more information, visit http://www.cs.dartmouth.edu/~dfk/.

REFERENCES

[1] Hey, T., Tansley, S., & Tolle, K. M. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery (Vol. 1). Redmond, WA: Microsoft Research.
[2] Rawassizadeh, R., Price, B. A., & Petre, M. (2015). Wearables: has the age of smartwatches finally arrived? *Communications of the ACM*, 58(1), 45-47.
[3] Eagle, N., & Pentland, A. S. (2006). Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4), 255-268.
[4] Laurila, J. K., Gatica-Perez, D., Aad, I., Bornet, O., Do, T. M. T., Dousse, O., ... & Miettinen, M. (2012). The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing* (No. EPFL-CONF- 192489).
[5] Wagner, D. T., Rice, A., & Beresford, A. R. (2014). Device Analyzer: Large-scale mobile data collection. *ACM SIGMETRICS Performance Evaluation Review*, 41(4), 53-56.
[6] Shepard, C., Rahmati, A., Tossell, C., Zhong, L., & Kortum, P. (2011). LiveLab: Measuring Wireless Networks and Smartphone Users in the Field. *ACM SIGMETRICS Performance Evaluation Review*, 38(3), 15-20.
[7] Shi, J., Santos, E., & Challen, G. (2016). Why and How to Use Phonelab. *GetMobile: Mobile Computing and Communications*, 19(4), 32-38.
[8] Rawassizadeh, R., Tomitsch, M., Nourizadeh, M., Momeni, E., Peery, A., Ulanova, L., & Pazzani, M. (2015). Energy-Efficient Integration of Continuous Context Sensing and Prediction into Smartwatches. *Sensors*, 15(9), 22616-22645.
[9] Lichman, M. (2015). UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml
[10] Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A., & Batista, G. (2016). The UCR time series classification archive.
[11] Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G. & Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, 101(23), e215-e220.
[12] Kotz, D., & Henderson, T. (2005). CRAWDAD: A Community Resource for Archiving Wireless Data At Dartmouth. *IEEE Pervasive Computing*, 4(4), 12-14.