DiversityOne Open Challenge: Exploring People's Everyday Life Behavior with Mobile Data

Andrea Bontempelli
University of Trento
Trento, Italy
andrea.bontempelli@unitn.it

Matteo Busso
University of Trento
Trento, Italy
matteo.busso@unitn.it

Lakmal Meegahapola*
Nokia Bell Labs
Cambridge, United Kingdom
lakmal.meegahapola@nokia.com

Amalia de Götzen
Aalborg University
Copenhagen, Denmark
ago@create.aau.dk

Fausto Giunchiglia
University of Trento
Trento, Italy
fausto.giunchiglia@unitn.it

Daniel Gatica-Perez
Idiap Research Institute
Martigny, Switzerland
EPFL
Lausanne, Switzerland
gatica@idiap.ch

Abstract

Combining passive smartphone sensor data with self-reports enables the study of everyday life behavior. The *DiversityOne* dataset, recently released to the UbiComp community, overcomes the limitation of existing datasets by containing data from eight countries from both Global South and Global North, including data from 782 college students, combining the data from 26 smartphone sensors and 350K+ self-reports, and extensive demographic and psychosocial survey data from 18K college students. The richness of the dataset opens the way to investigating important problems in multiple disciplines. Existing studies that leverage the dataset have only scratched the surface of the potential research questions it can help answer. Thus, we are organizing this workshop to foster creative and multidisciplinary works on this dataset: from AI and ubiquitous and mobile computing to computational social science and design.

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing**; *Empirical studies in ubiquitous and mobile computing*; *Smartphones*; *Mobile computing*.

Keywords

datasets, diversity, social practices, mobile sensing, smartphone sensing, wellbeing, health, generalization

ACM Reference Format:
Andrea Bontempelli, Matteo Busso, Lakmal Meegahapola, Amalia de Götzen, Fausto Giunchiglia, and Daniel Gatica-Perez. 2025. *DiversityOne* Open Challenge: Exploring People's Everyday Life Behavior with Mobile Data. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '25), October 12–16, 2025, Espoo, Finland.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3714394.3750577

*Work done while at ETH Zurich

UbiComp Companion '25, Espoo, Finland

ACM ISBN 979-8-4007-1477-1/2025/10
https://doi.org/10.1145/3714394.3750577

1 Introduction and Background

DiversityOne is the result of the large-scale European project "WeNet - The Internet of US"[1] [14]. The dataset paper by Busso et al. [5], recently accepted at IMWUT, provides a detailed description of the dataset, including the data collection methodology and descriptive statistics. *DiversityOne* contains data from college students in eight countries: China (Jilin University, **JLU**), Denmark (Aalborg University, **AAU**), India (Amrita Vishwa Vidyapeetham, **AMRITA**), Italy (University of Trento, **UNITN**), Mexico (Instituto Potosino de Investigación Científica y Tecnológica, **IPICYT**), Mongolia (National University of Mongolia, **NUM**), Paraguay (Universidad Católica "Nuestra Señora de la Asunción", **UC**), and the United Kingdom (London School of Economics and Political Science, **LSE**). The dataset contains questionnaire answers from more than 18,000 students, 782 of whom agreed to participate in an intensive longitudinal survey of four weeks. The study used the `iLog` app [18] to collect data from 26 smartphone sensors such as accelerometer, gyroscope, and GPS, as well as derived information such as notification interactions, app usage, activities, and step counts. During this period, participants self-reported their activities, locations, social context, moods over the day, and daily reports on sleep quality and daily expectations. The combination of sensor data and self-reported annotation across eight universities worldwide fosters research in fields such as ubiquitous computing, mobile sensing, machine learning and computation social science. The dataset's size and variety also allow the design of new data-driven studies of human behavior.

Previous studies investigated various aspects of high-quality rich datasets including sensor data and self-reports. Some examples are: the use of social media [8], the quality of answers and mislabeling [3], the usefulness of self-reports towards understanding the user subjective perspective of the local context [20], the impact of Covid on the students' lives [6], cross-individual activity recognition [17], mood inference [12], diversity perceptions in a community [10], activity recognition [4], social context inference while eating [9],

[1] WeNet - The Internet of US https://doi.org/10.3030/823783

inferring mood-while-eating [2], and the generation of contextually rich data with other reference datasets [7].

Some of these findings advocate for training localized models, for instance, over multi-country models [12] or user clusters [17], in order to capture local nuances. Highly subjective and complex inferences must consider the individual target user and not only the population-level data. For this reason, hybrid models combine population-level with user-specific data to personalize the model in complex daily activities recognition and social context recognition tasks [1, 11]. Meegahapola et al. [13] worked on domain adaptation by applying an unsupervised adaptation model on mood and social context tasks. This solution mitigates distribution shift by integrating diverse streams, e.g., activity, app usage and device usage, allowing to capture diverse contextual and activity patterns. Other studies investigated how to identify mistakes in participants' self-reports to improve the quality of the data [19].

In this workshop, we want to encourage participants to explore the potentiality of this dataset beyond the work already done. The richness of the data allows novel and multidisciplinary analysis that can answer intriguing research questions. Thus, this workshop will facilitate the exchange of ideas and approaches across different research fields, thereby fostering new collaborations. In the rest of this work, we detail the workshop objectives and structure.

2 Workshop objectives

The primary objective of the workshop is to explore the potential of the dataset given its size and multifaceted diversity, in terms of, e.g., sensors used, human feedback, and geographical diversity. The workshop topics are kept quite open to ensure contributions from a broader community of researchers. A non-exhaustive list of topics includes:

- studies exploiting the richness in terms of size and type of data of *DiversityOne*;
- studies focusing on a diversity-aware comparison of human behaviors across cultures or profiles;
- studies focusing on the design and documentation of the dataset collection;
- studies focusing on the design affordances of the dataset;
- proposals for new types of data-driven studies;
- machine learning algorithms to improve the datasets (e.g., data-centric AI);
- smartphone sensing for behavior modeling.

All the studies above are welcome to focus on algorithms, Artificial Intelligence, user studies, computational social science, data-centric design, user-centered design and other areas. This challenge does not advocate the use of the datasets as a benchmark [15, 16], which implies using the dataset to evaluate the model without considering the data collection and evaluation context, and claiming general model capabilities beyond the scope of the dataset. Indeed, the dataset emphasis on Global South and the diverse cultural and socio-demographic factors of the study participants facilitates multidisciplinary analysis and fosters the exploration of human behavior using novel and creative approaches. The characteristics of the dataset stimulate solutions that study the diversity among people and their similarities. The workshop is an exciting venue for fostering the exchange of ideas and interaction among researchers from diverse fields and backgrounds. The dataset has already been used by the UbiComp community, and it is a valuable research resource in ubiquitous computing, human-computer interaction, mobile computing and machine learning. The combination of passive sensor data with self-reports and questionnaires about personality traits and habits can support the inference of various aspects of everyday life, including mood, activities, social context, and eating habits.

3 Promoting diversity and participation

We welcome contributions from participants from diverse scientific backgrounds, geographical regions and cultures. Diversity is one aspect that characterizes *DiversityOne* dataset, and we would like it to be also reflected in the participation. We believe that a successful workshop should involve participation from both the Global North and the Global South to foster diverse interpretations and approaches to this open challenge. To promote participation, we will discuss with the workshop chairs the possibility of supporting video/remote presentations for authors unable to travel to the conference, offering reduced rates for them and/or special funds for students from the Global South, as implemented in other conferences.

4 Organizers' background

The organizers of this workshop also contributed in various ways to the dataset. Their research background is varied and covers the main workshop topics. The expertise is in sociology, service design, machine learning, ubiquitous computing, mobile sensing, and knowledge representation. Below, we briefly present each organizer.

Andrea Bontempelli is a postdoctoral researcher at the University of Trento working on interactive machine learning and concept drift with a focus on sensor data streams and noisy data. He also has expertise in research data management.

Matteo Busso holds a Master's in Sociology and a PhD in Computer Science from the University of Trento (Italy), where he focused on integrating sociological methods into ubiquitous computing experiments. His research aims to gather high-quality data on individual diversity for social interactions. He teaches AI to non-experts and contributes to designing and implementing a research infrastructure for generating and sharing diversity-aware data.

Lakmal Meegahapola is a Research Scientist at Nokia Bell Labs in Cambridge, UK. Previously, he was a postdoctoral researcher at ETH Zurich. He obtained his PhD from EPFL in 2024 and has also worked at Google Research, University of Cambridge, and Singapore Management University. His work lies at the intersection of mobile and wearable sensing, digital health, and responsible AI, where he develops safe and robust AI/ML models and LLM pipelines for multimodal time series data. He is on the editorial board of ACM IMWUT, IEEE Pervasive Computing, and also is an Associate Chair of ACM CHI.

Amalia de Götzen is an Associate Professor at Aalborg University in Copenhagen and a member of the Service Design Lab. Her work as a researcher and educator lies at the intersection of interaction design and service design, with a focus on digital civics.

Fausto Giunchiglia is a professor of Computer Science at the University of Trento working on end-to-end data-centric AI, from data

collection to data analysis, with a special interest on the study of human behavior.

Daniel Gatica-Perez leads the Social Computing Group at Idiap and is a professor at EPFL in Switzerland. He has worked extensively on smartphone sensing research. He co-organized the Nokia Mobile Data Challenge in 2012, served as General Co-Chair of ACM Ubicomp/ISWC in 2015, and serves as Associate Editor of PACM IMWUT since the journal's foundation.

5 Program Committee

The Program Committee includes members with affiliations in various countries, including Uganda, Mexico, Cyprus, Italy, Mongolia, China, India and Paraguay, and diverse academic backgrounds to cover the various topics of the submitted papers. The final list is published on the workshop website.

6 Pre-Workshop plan

6.1 Event promotion

We will promote it across diverse research communities and geographical areas to ensure the participation of researchers with diverse expertise and provenance. Thanks to the diverse provenance and background of the organizers, the promotion will go beyond the traditional IMWUT community to reach a broader set of disciplines through topic-specific mailing lists, universities and institution partners in WeNet, and social media. This step is crucial for receiving multidisciplinary submissions.

6.2 Getting the dataset

6.2.1 Dataset discovery and understanding. The UniTN data catalog serves as the primary entry point for navigating dataset metadata and documentation. Researchers can request access to basic datasets from a specific pilot site (e.g., time diaries collected at UniTN) or a combination from multiple pilot sites. To facilitate the data selection and download, bundles of datasets have been created that group data commonly used together for main research purposes. For example, the motion bundle includes all motion sensor data relevant to activity recognition studies, while another bundle, combining questionnaires, time diaries, and location data, is tailored for studying social interactions. Any combination of bundles and single sensors can be downloaded. The catalog lists both datasets containing one single sensor and bundles, and supports the search by data type and location. The full list of bundles and sensors is reported in the dataset paper [5]. The full dataset size is approximately 94 GB in Parquet format, a space-efficient format supported by most data analysis tools.

Important Links

Dataset Catalog: https://livepeople.disi.unitn.it/
***DiversityOne* Website:** https://datascientia.disi.unitn.it/projects/diversityone/
Workshop Website: https://datascientiafoundation.github.io/diversityone-2025/

6.2.2 Request procedure. The dataset request, to be submitted to UniTN, has been designed together with legal and privacy experts to ensure full compliance with the GDPR. The procedure is articulated as follows.

(1) The participants navigate the dataset catalog and identify the data of interest.
(2) Through the web form available on the catalog website, participants register with their institutional email and fill it with the datasets or bundle identifiers, provide contact details of the other researchers and institutions involved in the work, and send a research proposal.
(3) The key requirements to obtain a copy of the data are the affiliation with a research institution, either private or public, the coherence between the requested data and the research proposal, and the acceptance of the Terms and License Agreement, also based on Art. 28 of the GDPR. Key licensing terms include: *(i)* datasets are used exclusively for research purposes; *(ii)* redistribution of the datasets is prohibited; *(iii)* datasets cannot be publicly shared (e.g., on a website); and *(iv)* any attempt to reverse engineer any portion of the data or to re-identify the participants is strictly forbidden and could constitute unlawful processing of personal data.
(4) The organizers evaluate the request, and if approved, participants receive an email with instructions for downloading the dataset. The evaluation procedure and requested data composition will take a few days.

6.3 Work period and paper submission

To streamline dataset acquisition and analysis, UniTN will provide technical support via email by setting up a help desk. In addition, the data catalog website collects frequently asked questions, instructions on how to request the data and the data documentation. We expect the submission to be a short paper of a maximum of four pages (excluding references) reflecting on, analyzing or testing the dataset. The paper should describe the motivation, methodology, results, future analysis, and potential negative societal impacts. Optionally, authors can attach additional material. Given that the authors have access to the datasets for the first time and have limited time, we expect the papers to present only preliminary analysis and dataset exploration. Each paper is evaluated by two reviewers The main evaluation criteria are *creativity* (how novel the analysis is), *multidisciplinary* (how well it combines ideas and approaches from multiple disciplines), *presentation* (how clear and well-written the paper is), and *impact* (how likely it is that the analysis can lead to impactful results if the work is further extended and studied). We will award prizes to the best paper and runner-up best paper, and we will evaluate the awarding with vouchers.

7 Workshop structure

Table 1 presents the tentative full-day workshop schedule, which will be adapted according to the number of accepted papers. The workshop attendance is open, and thus, participants can attend without submitting a paper. We expect around 20-25 participants, and we consider the workshop successful with 15-20 submissions.

- **Introduction**. We will welcome the participants and outline the structure of the workshop.

- **Interactive Paper Presentation**. Participants will present their work in a five-minute presentation and a five-minute discussion[2]. Presenters will showcase their insights on the dataset and gain feedback from the audience. We value the discussion phase, which aims at encouraging an exchange of ideas and approaches.
- **Keynote Talk**. Based on the number of submissions, we will invite one or two expert researchers in related areas.
- **Next steps and closing remarks**. We will summarize the key messages from the discussions and identify the next steps to stimulate multidisciplinary work around large longitudinal studies on human behavior.

Table 1: Tentative Workshop Timeline

Time	Session
10 min	Opening
40 min	Keynote + Q/A
50 min	Participants' presentation and discussion
15 min	Break
50 min	Participants' presentation and discussion
	Lunch break
40 min	Keynote + Q/A
50 min	Participants' presentation and discussion
15 min	Break
50 min	Participants' presentation and discussion
15 min	Awards and closing remarks

8 Post-Workshop plan

The workshop papers will be included in the ACM Digital Library (DL) as adjunct proceedings of the UbiComp conference. In addition, the presentations, papers and additional material will be published on the workshop webpage to encourage further analysis and exploration in multiple research directions. The most promising papers will be invited to submit an extended version to IEEE Pervasive Computing. After the workshop, we plan to summarize the results of the workshop in a position paper. If the workshop is successful, as we are optimistic, we plan to organize further editions of the workshop, which may include additional datasets resulting from the European WeNet project about to be published. For this reason, we will administer a brief questionnaire to the participants to understand how we can improve the workshop and the dissemination of the dataset.

Acknowledgments

The authors acknowledge the technical support of Munkhdelger Bayanjargal and Ali Hamza in distributing the dataset to the requesters.

References

[1] Karim Assi, Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Miriam Bidoglia, Sally Stares, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, et al. 2023. Complex daily activities, country-level diversity, and smartphone sensing: A study in Denmark, Italy, Mongolia, Paraguay, and UK. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–23.

[2] Wageesha Bangamuarachchi, Anju Chamantha, Lakmal Meegahapola, Haeeun Kim, Salvador Ruiz-Correa, Indika Perera, and Daniel Gatica-Perez. 2025. Inferring Mood-While-Eating with Smartphone Sensing and Community-Based Model Personalization. *ACM Transactions on Computing for Healthcare (HEALTH)* (2025).

[3] Andrea Bontempelli, Stefano Teso, Fausto Giunchiglia, and Andrea Passerini. 2020. Learning in the Wild with Incremental Skeptical Gaussian Processes. In *IJCAI*.

[4] Emma Bouton-Bessac, Lakmal Meegahapola, and Daniel Gatica-Perez. 2022. Your Day in Your Pocket: Complex Activity Recognition from Smartphone Accelerometers. In *International Conference on Pervasive Computing Technologies for Healthcare*. Springer, 247–258.

[5] Matteo Busso, Andrea Bontempelli, Leonardo Javier Malcotti, Lakmal Meegahapola, Peter Kun, Shyam Diwakar, Chaitanya Nutakki, Marcelo Rodas Britez, et al. 2025. DiversityOne: A Multi-Country Smartphone Sensor Dataset for Everyday Life Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (2025). https://arxiv.org/pdf/2502.03347

[6] Nicolò Alessandro Girardini, Simone Centellegher, Andrea Passerini, Ivano Bison, Fausto Giunchiglia, and Bruno Lepri. 2023. Adaptation of student behavioural routines during Covid-19: a multimodal approach. *EPJ Data Science* 12, 1 (2023), 55.

[7] Fausto Giunchiglia and Xiaoyue Li. 2024. Big-Thick Data generation via reference and personal context unification. In *ECAI 2024*. IOS Press, 1975–1984.

[8] Fausto Giunchiglia, Mattia Zeni, Elisa Gobbi, Enrico Bignotti, and Ivano Bison. 2018. Mobile social media usage and academic performance. *Computers in Human Behavior* 82 (2018), 177–185.

[9] Nathan Kammoun, Lakmal Meegahapola, and Daniel Gatica-Perez. 2023. Understanding the Social Context of Eating with Multimodal Smartphone Sensing: The Role of Country Diversity. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 604–612.

[10] Peter Kun, Amalia de Götzen, Miriam Bidoglia, Niels Jørgen Gommesen, and George Gaskell. 2022. Exploring diversity perceptions in a community through a Q&A chatbot. In *DRS2022: Bilbao Design Research Society*. 1–19.

[11] Aurel Ruben Mäder, Lakmal Meegahapola, and Daniel Gatica-Perez. 2024. Learning About Social Context from Smartphone Data: Generalization Across Countries and Daily Life Moments. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[12] Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Chaitanya Nutakki, Shyam Diwakar, et al. 2023. Generalization and personalization of mobile sensing-based mood inference models: an analysis of college students in eight countries. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 6, 4 (2023), 1–32.

[13] Lakmal Meegahapola, Hamza Hassoune, and Daniel Gatica-Perez. 2024. M3BAT: Unsupervised Domain Adaptation for Multimodal Mobile Sensing with Multi-Branch Adversarial Training. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–30.

[14] L. Michael, I. Bison, M. Busso, L. Cernuzzi, A. de Götzen, S. Diwakar, K. Gal, S. Ganbold, G. Gaskell, D. Gatica-Perez, J. Heesen, D. Miorandi, N. Osman, S. Ruiz-Correa, L. Schelenz, A. Segal, C. Sierra, H. Xu, and F. Giunchiglia. 2025. Towards Open Diversity-Aware Social Interactions. *ArXiv, under submission* (2025).

[15] Will Orr and Edward B Kang. 2024. AI as a Sport: On the Competitive Epistemologies of Benchmarking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1875–1884.

[16] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366* (2021).

[17] Qiang Shen, Haotian Feng, Rui Song, Stefano Teso, Fausto Giunchiglia, Hao Xu, et al. 2022. Federated multi-task attention for cross-individual human activity recognition. In *IJCAI*. IJCAI, 3423–3429.

[18] Mattia Zeni, Ilya Zaihrayeu, and Fausto Giunchiglia. 2014. Multi-device activity logging. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. 299–302.

[19] Mattia Zeni, Wanyi Zhang, Enrico Bignotti, Andrea Passerini, and Fausto Giunchiglia. 2019. Fixing Mislabeling by Human Annotators Leveraging Conflict Resolution and Prior Knowledge. *IMWUT* 3, 1 (2019).

[20] Wanyi Zhang, Qiang Shen, Stefano Teso, Bruno Lepri, Andrea Passerini, Ivano Bison, and Fausto Giunchiglia. 2021. Putting human behavior predictability in context. *EPJ Data Science* 10, 1 (2021), 42.

[2]The final timing will be defined based on the number of submissions. If the number of submissions is higher than expected, we will add a poster session.