Department of Computer Science
Series of Publications A
Report A-2023-18

Pervasive Data Science: From Data Collection to End-User Applications

Agustin Zuniga Corrales

Doctoral dissertation, to be presented for public examination with the permission of the Faculty of Science of the University of Helsinki in Main Building, Karolina Eskelin U3032 on November 2^{nd}, 2023 at 12 o'clock.

University of Helsinki
Finland

Supervisors
Petteri Nurmi, University of Helsinki, Finland
Pan Hui, University of Helsinki, Finland and HKUST, Hong Kong
Huber Flores, University of Tartu, Estonia

Pre-examiners
Evangelos Pournaras, University of Leeds, UK
Robert LiKamWa, Arizona State University, US

Opponent
George Roussos, Birkbeck College, University of London, UK

Custos
Petteri Nurmi, University of Helsinki, Finland

Contact information

Department of Computer Science
P.O. Box 68 (Pietari Kalmin katu 5)
FI-00014 University of Helsinki
Finland

Email address: info@cs.helsinki.fi
URL: http://cs.helsinki.fi/
Telephone: +358 2941 911

ISSN 1238-8645 (print)
ISSN 2814-4031 (online)
ISBN 978-951-51-9972-0 (paperback)
ISBN 978-951-51-9973-7 (PDF)
Helsinki 2023
Unigrafia

Pervasive Data Science: From Data Collection to End-User Applications

Agustin Zuniga Corrales

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
agustin.zuniga@helsinki.fi
https://agustin-zuniga.com

PhD Thesis, Series of Publications A, Report A-2023-18
Helsinki, November 2023, 72+132 pages
ISSN 1238-8645 (print)
ISSN 2814-4031 (online)
ISBN 978-951-51-9972-0 (paperback)
ISBN 978-951-51-9973-7 (PDF)

Abstract

Pervasive Data Science (PDS) is an emerging paradigm that combines the Internet of Things, Pervasive Computing, and Data Science to address everyday challenges. PDS differs from traditional data science in that it harnesses data from pervasive computing deployments, which affects the way data is produced and how it can be analyzed. To date, PDS has received limited attention as an independent research domain as the research field is fragmented and scattered among many different subfields. This is due to a limited understanding of the characteristics and challenges in PDS, and a lack of end-user applications that demonstrate the benefits of PDS. This thesis paves the way for improving the adoption of PDS by offering (i) insights into the processes that produce data, (ii) demonstrating how pervasive computing deployments can enable wide-range of applications by re-purposing existing sensors and capabilities of pervasive computing devices, and (iii) highlighting the potential benefits of Pervasive Data Science by developing end-user applications for tackling sustainable development.

Computing Reviews (2012) Categories and Subject Descriptors:
Human-centered computing → Ubiquitous and mobile computing
Computing methodologies → Machine learning
General and reference → Cross-computing tools and techniques

General Terms:
pervasive data science, internet of things, pervasive computing, data science

Additional Key Words and Phrases:
doctoral thesis, iot architecture, sensing pipeline, data fusion, data collection, crowdsensing, multi-device collaborative sensing and computing, autonomous vehicles, low-cost sensing, indoors monitoring, optical sensing, thermal imaging, sustainable development, pollution monitoring, food waste, deep learning

To Isabel, Samantha, Valeria, Danna, Mihail and Nicole

Acknowledgements

First of all, I would like to express my sincere gratitude to my thesis supervisor, Professor Petteri Nurmi. Your invaluable mentorship since 2017, when I joined the Pervasive Data Science research group, has contributed to the successful completion of my Master's and Ph.D. studies, as well as building a strong and holistic academic profile. Thank you for your valuable time and support, especially for guiding me to understand that problem solving is based on exploring the foundations and making the things happen.

I would like to thank my pre-examiners Associate Professor Evangelos Pournaras and Associate Professor Robert LiKamWa for the time they spent providing valuable feedback on this dissertation. I would also like to thank Professor George Roussos for agreeing to be my opponent.

I would like to thank Professor Pan Hui and Associate Professor Huber Flores for agreeing to be co-supervisors of my doctoral studies, and to Professor Jussi Kangasharju and Professor Hannu Toivonen for being part of my thesis committee. Your invaluable feedback and support contributed to the completion of my Ph.D. journey.

I would like to thank the people who are part of the Doctoral Programme in Computer Science, especially Dr. Pirjo Moen for her patience in guiding me to successfully complete all the steps of the doctoral process.

I would like to acknowledge the financial support for funding my studies provided by the Academy of Finland projects lead by Professor Petteri Nurmi and Professor Sasu Tarkoma, the Helsinki Institute for Information Technology, the Department of Computer Science of the University of Helsinki and the Nokia Foundation. Your valuable support allowed me to concentrate 100% on my research.

I would like to thank all the co-authors, especially Dr. Naser Hossein Motlagh and Dr. Marko Radeta, for their fruitful contribution. My appreciation to all the researchers that have offered their support, especially Dr. Roberto Morabito, Dr. Gopika Premsankar, Dr. Sara Ramezanian and Dr. Ashwin Rao. My special thanks go to Dr. Ngoc Thi Nguyen, thank you for making the last part of the journey even more interesting and valuable.

I would like to thank all the people who helped me along the way, particularly those who encouraged me to go in this direction, especially Alejandra, Milton and Paulina. My special thanks go to Suvi, thank you for your invaluable support throughout my studies and journey at the University of Helsinki.

I would like to thank my parents, Marcia and Jaime, and my sisters, Marcia, Paulita and Nadya, for their love, support and encouragement, not only during my doctoral studies, but throughout my life. My beloved family, you are fundamental to the achievement of my goals.

Last but not least, I would like to thank myself for giving me the chance to start this journey and the courage to finish it.

Helsinki, October 2023

Agustin Zuniga Corrales

Agradecimientos

En primer lugar, me gustaría expresar mi más sincero agradecimiento a mi director de tesis, Profesor Petteri Nurmi. Su inestimable mentoría desde 2017, año en el cual comencé a formar parte del grupo de investigación Pervasive Data Science, ha contribuido a la finalización con éxito de mis estudios de máster y doctorado, así como a la construcción de un perfil académico sólido y holístico. Agradezco su valioso tiempo y apoyo, especialmente por guiarme a entender que la resolución de problemas se basa en explorar los fundamentos y hacer que las cosas sucedan.

Mis agradecimientos a mis preexaminadores, el Profesor Asociado Evangelos Pournaras y el Profesor Asociado Robert LiKamWa, por el tiempo dedicado a proporcionarme sus valiosos comentarios. También me gustaría dar las gracias a Professor George Roussos por aceptar ser mi oponente.

Mi agradecimientos al Profesor Pan Hui y al Profesor Asociado Huber Flores por aceptar ser co-supervisores de mi doctorado, y al Profesor Jussi Kangasharju y Profesor Hannu Toivonen por conformar mi comité de tesis. Sus inestimables comentarios contribuyeron a completar mi doctorado.

Mi agradecimientos a quienes conforman el Programa de Doctorado en Ciencias de la Computatción, especialmente a la Dra. Pirjo Moen por su paciencia y guía para completar con éxito todo el proceso de doctorado.

Mi agradecimientos por el apoyo económico para la financiación de mis estudios proporcionado por los proyectos de la Academia de Finlandia dirigidos por el Profesor Petteri Nurmi y el Profesor Sasu Tarkoma, el Helsinki Institute for Information Technology, el Departamento de Ciencias de la Computación la Universidad de Helsinki y la Nokia Foundation. Su valioso apoyo me ha permitido concentrarme al 100% en mi investigación.

Mi agradecimientos a todos los coautores, especialmente al Dr. Naser Hossein Motlagh y al Dr. Marko Radeta, por su fructífera contribución. Mi aprecio a los investigadores que ofrecieron su soporte, especialmente al Dr. Roberto Morabito, Dra. Gopika Premsankar, Dra. Sara Ramezanian and Dr. Ashwin Rao. Mi agradecimiento especial a la Dra. Ngoc Thi Nguyen, gracias por hacer de la última parte de la jornada más interesante y valiosa.

Mi agradecimientos a todas las personas que me han ayudado a lo largo del camino, particularmente a las que me animaron a seguir en esta dirección, especialmente a Alejandra, Milton y Paulina. Mi agradecimiento especial a Suvi, gracias por el invaluable apoyo ofrecido a lo largo de mis estudios y jornada en la Universidad de Helsinki.

Quiero agradecer mis padres, Marcia y Jaime, y a mis hermanas, Marcia, Paulita y Nadya, su cariño, apoyo y ánimo, no sólo durante mis estudios de doctorado, sino a lo largo de toda mi vida. Mi amada familia, ustedes son fundamentales para la consecución de mis objetivos.

Por último, pero no por ello menos importante, me gustaría darme las gracias a mí mismo por haberme dado la oportunidad de iniciar esta jornada y el valor para terminarla.

Helsinki, Octubre 2023

Agustin Zuniga Corrales

Original publications

This thesis is based on the following original publications, referred to as **Publications I-X** in the text and printed at the end of the thesis.

I **Tortoise or Hare? Quantifying the Effects of Performance on Mobile App Retention**
Agustin Zuniga, Huber Flores, Eemil Lagerspetz, Sasu Tarkoma, Pan Hui, Jukka Manner, Petteri Nurmi. In *Proceedings of The World Wide Web Conference (WWW),* pages 2517–2528. 2019. DOI: https://doi.org/10.1145/3308558.3313428.

II **COSINE: Collaborator selector for cooperative multi-device sensing and computing**
Huber Flores, Agustin Zuniga, Farbod Faghihi, Xin Li, Samuli Hemminki, Sasu Tarkoma, Pan Hui, Petteri Nurmi. In *Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1-10. IEEE, 2020. DOI: https://doi.org/10.1109/PerCom45495.2020.9127364.

III **Collaboration Stability: Quantifying the Success and Failure of Opportunistic Collaboration**
Huber Flores, Agustin Zuniga, Sasu Tarkoma, Leonardo Tonetto, Tristan Braud, Pan Hui, Yong Li, Mostafa Ammar, Petteri Nurmi. *Computer*, 55(8):70-81, 2022. DOI: https://doi.org/10.1109/MC.2021.3112850.

IV **Smart Plants: Low-Cost Solution for Monitoring Indoor Environments**
Agustin Zuniga, Naser Hossein Motlagh, Huber Flores, Petteri Nurmi. *IEEE Internet of Things Journal*, 9(22):23252–23259, 2022. DOI: https://doi.org/10.1109/JIOT.2022.3188475.

V **How Low Can You Go?: Performance Trade-offs in Low-Resolution Thermal Sensors for Occupancy Detection: A Systematic Evaluation**
Mikko Rinta-Homi, Naser Hossein Motlagh, Agustin Zuniga, Huber Flores, Petteri Nurmi. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, volume 5, pages 1-22. ACM 2021. DOI: https://doi.org/10.1145/3478104.

VI **See No Evil: Discovering Covert Surveillance Devices Using Thermal Imaging**
Agustin Zuniga, Naser Hossein Motlagh, Mohammad A. Hoque, Sasu Tarkoma, Huber Flores, Petteri Nurmi. *IEEE Pervasive Computing*, 21(4):33-42, 2022. DOI: https://doi.ieeecomputersociety.org/10.1109/MPRV.2022.3187464.

VII **Ripe or Rotten? Low-Cost Produce Quality Estimation Using Reflective Green Light Sensing**
Agustin Zuniga, Huber Flores, Petteri Nurmi. In *IEEE Pervasive Computing.* 20(3):60–67, 2021. DOI: https://doi.org/10.1109/MPRV.2021.3074474.

VIII **Toward Blue Skies: City-Scale Air Pollution Monitoring using UAVs**
Naser Hossein Motlagh, Matti Irjala, Agustin Zuniga, Eemil Lagerspetz, Valtteri Rantala, Huber Flores, Petteri Nurmi, Sasu Tarkoma. *IEEE Consumer Electronics Magazine*, 12(1):21-31, 2022. DOI: https://doi.org/10.1109/MCE.2022.3167800.

IX **Toward Large-Scale Autonomous Marine Pollution Monitoring**
Huber Flores, Naser Hossein Motlagh, Agustin Zuniga, Mohan Liyanage, Monica Passananti, Sasu Tarkoma, Moustafa Youssef, Petteri Nurmi. *IEEE Internet of Things Magazine*, 4(1):40-45, 2021. DOI: https://doi.org/10.1109/IOTM.0011.2000057.

X **Deep Learning and the Oceans**
Marko Radeta, Agustin Zuniga, Naser Hossein Motlagh, Mohan Liyanage, Ruben Freitas, Moustafa Youssef, Sasu Tarkoma, Huber Flores, Petteri Nurmi. *Computer*, 55(5):39-50, 2022. DOI: https://doi.org/10.1109/MC.2022.3143087.

Contents

Chapter 1

Introduction

Pervasive Data Science (PDS) has emerged as a powerful paradigm that harnesses the potential of data to uncover new insights and drive innovation [20, 59]. PDS builds on three research fields, Internet of Things (IoT), Pervasive Computing and Data Science, and on an exponential increase in the number of smart devices connected to the Internet. The number of smart devices are projected to almost double from 15.14 billion in 2023 to more than 29.42 billion in 2030 ($5.5 - $12.5 trillion in value)[50, 61]. This has been accompanied by a significant increase in academic studies in these fields, as highlighted in Figure 1.1 and several reports [44, 61]. Leveraging these advances, Pervasive Data Science has gained a significant momentum in recent years, empowering research in areas such as environmental monitoring [1, 72], healthcare [3, 25, 81], smart cities [16, 18, 30, 86] and precise agriculture [91, 93, 105].

Pervasive Data Science draws on the use of data from IoT and pervasive computing deployments for data science. For example, PDS applications for monitoring air quality [16, 18, 24] use IoT and pervasive computing technologies to deploy a distributed network of smart devices that periodically collect samples of multiple air compounds at different locations, and data science techniques to define the sampling process and transform the data into insights about spatio-temporal patterns of air quality. The structure of Pervasive Data Science systems is two-fold: (i) the system architecture and (ii) the sensing pipeline. The former specifies the overall composition of devices and data processing, comprising four functional layers: sensing, networking, data management, and privacy and security, while the latter specifies the algorithms and methods for gaining insight from sensor measurements, comprising four stages: data collection, cleaning and pre-processing, feature extraction, and modelling and evaluation.

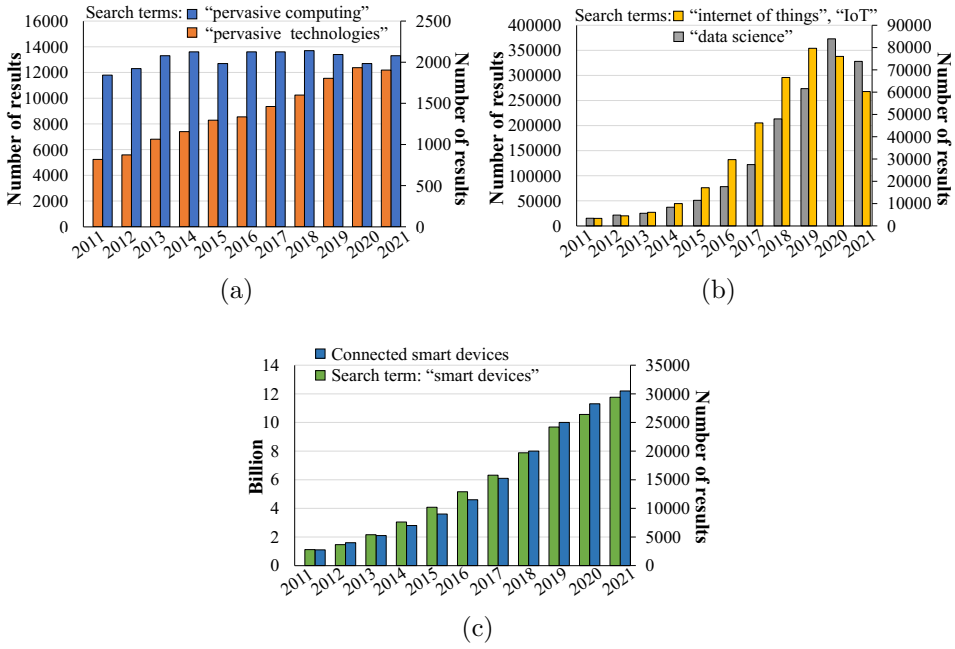


Figure 1.1: Research interest in topics related to pervasive computing, pervasive technologies, Internet of Things, data science and smart devices, and comparison with the number of connected smart devices globally since 2011. (a–c) Number of results corresponds to the search output in Google Scholar. (c) Number of connected smart devices retrieved from [44, 64].

Despite offering an integrated design that promises the acquisition of richer data and a deeper understanding of it, Pervasive Data Science has yet to gain as widespread recognition as the paradigms it builds on [20, 21]. In fact, there has been limited progress in the field of Pervasive Data Science, with only 69 studies explicitly focusing on this specific term, highlighting how the research field is scattered. This is due to a limited understanding of the characteristics and challenges in PDS, and a lack of end-user applications that demonstrate the benefits of PDS. This thesis paves the way for improving the adoption of PDS by offering (i) insights into the processes that produce data, (ii) demonstrating how pervasive computing deployments can enable wide-range of applications by re-purposing existing sensors and capabilities of pervasive computing devices, and (iii) highlighting the potential benefits of PDS by developing end-user applications for tackling sustainable development.

1.1 List of contributions

The core contributions of this thesis are solutions that address key challenges in pervasive data science: how to improve the quality of available data and how to enable new and better ways to collect data. The thesis also demonstrates the benefits of the contributions through novel applications addressing sustainable development in different environments. The core contributions are:

1. **Enhancing data quality: data analysis and multi-device collaboration:** The complexity of Pervasive Data Science applications stems from the many factors that affect data collection, such as different users, devices, environments or contexts. This complexity and the variety of factors impacting data quality result in challenges related to the limited usability of PDS applications and the quality of the data that is available. This thesis provides solutions to address these limitations, and to improve data quality. In **Publication I** we demonstrate the potential benefits of data fusion to generate an understanding of complex interactions, and provide ways to handle limitations with large-scale datasets. We rely on mobile computing data to provide a method for quantifying the relationship between energy consumption, network latency and application retention, highlighting the trade-off between data size and quality. **Publications II and III** address the challenges of efficient data collection in large-scale, multi-device collaborative environments to improve data collection and quality. **Publications II** introduces a novel regularity-based method for selecting long-term collaborators, and **Publication III** demonstrates how spatio-temporal characteristics of mobility in different locations affect collaboration opportunities in everyday contexts.

2. **Enabling novel pervasive data science applications through low-cost sensing**: Pervasive computing has the potential to expand further into new domains that benefit the society. In real-world deployments, key challenges include developing affordable and reliable solutions and guidelines on how to configure sensors. We demonstrate the potential and the benefits of low-cost sensing technologies and the importance of evaluating performance trade-offs to enable pervasive data science applications. We demonstrate how pervasive data science applications can be used for low-cost end-to-end monitoring of indoor environments. **Publication IV** shows how CO_2 accumulation, occupancy and face mask usage can be estimated from air conditioning measurements collected by a prototype

smart plant container that integrates low-cost commercial environmental sensors. **Publication V** demonstrates the importance of sensor configuration and the cost benefit of using low-cost sensors by evaluating the performance trade-offs and costs of using low-resolution thermal array sensors for occupancy detection and counting, and providing guidelines for sensor configuration. **Publication VI** demonstrates how PDS and low-cost sensing can be used to provide new insights and support existing applications. We introduce a novel privacy-preserving method for indoor environments that can accurately detect covert surveillance devices using thermal imaging, and that can complement existing approaches.

3. **Supporting end-user applications for sustainable development:** Adopting a new paradigm requires demonstrating its potential benefits to society. We demonstrate the societal benefits that Pervasive Data Science can bring by addressing environmental sustainability challenges and supporting sustainable development. We also present the challenges and roadmap to be followed to implement these applications in real life. **Publication VII** contributes with a novel low-cost sensing method to reduce food loss and waste that can characterise the quality of produce using contact-based reflective green-light sensing and can be used at all stages of the supply chain. Publications VIII and IX present a vision for multi-ecosystem pollution monitoring using PDS integrated onto smart devices. **Publication VIII** focuses on city-scale air quality monitoring using unmanned aerial vehicles (UAVs), demonstrating that UAV-based pollution monitoring complements ground measurements by providing information on the vertical distribution and potential dispersion mechanisms of pollutants. **Publication IX** presents how large-scale pollution monitoring can be achieved by using coordinated groups of autonomous underwater vehicles (AUVs) equipped with low-cost sensors. **Publication X** contributes by showing how the reorientation and adoption of the technology requires overcoming the challenges inherent in its implementation. We analyse the performance of deep learning in underwater environments, and present the roadmap for using these models in aquatic environments.

1.2 Individual Contributions to Original Publications

In the following, we detail the individual role and contribution of the current author to the original publications that form part of this thesis.

1.2.1 Insights and Methods for Data Collection

Publication I: Tortoise or Hare? Quantifying the Effects of Performance on Mobile App Retention
The original idea of quantifying the effect of performance on mobile app retention and using data fusion to evaluate the combined effect was originated with the current author under the supervision of Eemil Lagerspetz, Huber Flores and Petteri Nurmi. Data fusion and analysis was performed by the current author with feedback from Eemil Lagerspetz, Huber Flores and Petteri Nurmi. Modelling was done by the current author in collaboration with Huber Flores. The first draft of the manuscript was prepared by the current author, while the subsequent writing was a collaborative effort involving all co-authors.

Publication II: COSINE: Collaborator selector for cooperative multi-device sensing and computing
The concept of evaluating different approaches to selecting collaborators in multi-device environments originated with the current author, Huber Flores and Petteri Nurmi. The dataset preparation and experimental evaluation was a joint effort between the current author and Huber Flores. The first draft of the manuscript was prepared by the current author and Huber Flores. The subsequent writing was a collaborative effort involving all co-authors.

Publication III: Collaboration Stability: Quantifying the Success and Failure of Opportunistic Collaboration
The innovative approach of evaluating different methods for selecting collaborators in multi-device environments was jointly conceived by the current author, Huber Flores and Petteri Nurmi. Dataset preparation, stability analysis, quantification and modelling were performed by the current author and Huber Flores, with feedback from Petteri Nurmi. The first draft of the manuscript was prepared the current author and Huber Flores, while the subsequent writing was a collaborative effort involving all co-authors.

1.2.2 Low-Cost PDS Systems for Indoor Monitoring

Publication IV: Smart Plants: Low-Cost Solution for Monitoring Indoor Environments
The vision of instrumenting plant containers to monitor indoor environments originated from the current author under the supervision of Petteri

Nurmi. The experiments, analysis and modelling were conducted by the present author, with feedback from Huber Flores and Petteri Nurmi. The initial draft of the manuscript was prepared by the current author, while the preparation of the final version was a collaborative effort of all co-authors.

Publication V: How low can you go? performance trade-offs in low-resolution thermal sensors for occupancy detection: A systematic evaluation
The origin of the idea to evaluate the trade-offs in low resolution thermal sensors was Mikko Rinta-Homi and Petteri Nurmi. The methodology design, analysis, and evaluation were done by the current author and Mikko Rinta-Homi and under the supervision of Naser Hossein Motlagh and Petteri Nurmi. The first draft of the manuscript was produced by the current author and Mikko Rinta-Homi. The subsequent writing was a joint effort of all co-authors.

Publication VI: See No Evil: Discovering Covert Surveillance Devices Using Thermal Imaging
The novel concept of using thermal imaging to detect covert surveillance devices was conceived by the current author under the supervision of Petteri Nurmi. The experiments, analysis and modelling were carried out by the current author with feedback from Huber Flores and Petteri Nurmi. The first draft of the manuscript was prepared by the current author, while the subsequent writing was a collaborative effort involving all co-authors.

1.2.3 Pervasive Data Science for Sustainable Development

Publication VII: Ripe or Rotten? Low-Cost Produce Quality Estimation Using Reflective Green Light Sensing
The original idea of using green-light sensing to estimate produce quality came from the current author under the supervision of Huber Flores and Petteri Nurmi. The experiments were performed, evaluated and modelled by the current author with feedback from Huber Flores and Petteri Nurmi. The first draft of the manuscript was prepared by the current author. The preparation of the final version was a collaborative effort of all co-authors.

Publication VIII: Toward Blue Skies: City-Scale Air Pollution Monitoring using UAVs
The idea of using low-cost sensors and UAVs for air pollution monitoring originated from the present author, Aeromon Oy and Naser Hossein Motlagh

under the supervision of Petteri Nurmi and Sasu Tarkoma. The analysis and modelling were carried out by the present author with feedback from Petteri Nurmi. The first draft of the manuscript was prepared by the current author and Hossein Motlagh. The final version is a joint effort of all co-authors.

Publication IX: Toward large-scale autonomous marine pollution monitoring
The vision of using AUVs for marine pollution monitoring originated with the current author, Naser Hossein Motlagh, Huber Flores and Petteri Nurmi. Data collection, analysis and modelling were performed by the current author and Huber Flores with feedback from Petteri Nurmi. The first draft of the manuscript was prepared by the current author, Naser Hossein Motlagh and Huber Flores, while the final version was a collaboration of all co-authors.

Publication X: Deep Learning and the Oceans
The idea of evaluating the performance of deep learning underwater started with the current author and Marko Radeta under the supervision of Huber Flores and Petteri Nurmi. The field experiments were carried out in coordination with the current author and Marko Radeta. The analysis and evaluation were performed by the current author and Marko Radeta, with feedback from Huber Flores and Petteri Nurmi. The first draft of the manuscript was prepared by the current author and Marko Radeta, while the subsequent writing was a collaborative effort involving all co-authors.

1.3 Structure of the Thesis

This thesis is divided into four parts. Chapter 2 describes the contributions of Publications I, II and III on how to improve the quality of data and collaboration in large-scale contexts. Chapter 3 presents the results of Publications IV, V and VI on the use of low-cost sensors in novel applications and the evaluation of performance trade-offs. Chapter 4 introduces the findings of Publications VII, VIII, IX and X on the potential of PDS to support sustainable development. Chapter 5 summarises the contributions of this thesis and the future work, followed by reprints of the original publications.

Chapter 2

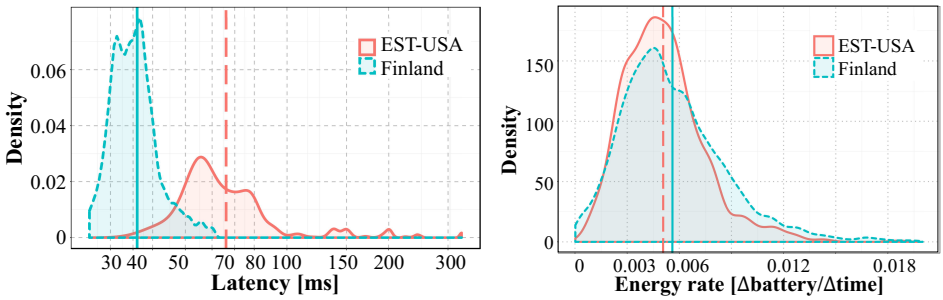
Insights and Methods for Data Collection

Pervasive Data Science harnesses data from diverse smart devices that are carried around by users, integrated into the infrastructure, or operate autonomously. The data that is available for analysis depends on the interplay of these devices, which makes it essential to understand how the available data is being generated and how the interactions between devices affect it. In this chapter we focus on *mobile crowdsensing*, the opportunistic collection of data from mobile devices, and provide insights on the processes that generate data and provide methods for optimizing the availability of measurements that can be used for data science investigations.

2.1 Retention and Data Quality

Today, there is an app for almost everything, with the major marketplaces offering millions of apps to users [26]. However, over a quarter of installed apps are only used once [47] and the rest are unlikely to remain relevant for more than few weeks [97]. Despite the widespread awareness of low app retention [9, 13], there is a lack of understanding regarding the threshold at which negative user perceptions lead to app abandonment. As apps are the main way to obtain data from smart devices, understanding the processes that govern application use and abandonment is critical for understanding how PDS deployments generate data.

Publication I first quantifies the relationship between mobile application performance and retention, i.e. whether users are willing to continue using an app. We perform our analysis by fusing two large-scale datasets: one comprising of crowdsensed measurements of network latency (NetRadar [95])



(a) Latency (b) Energy rate

Figure 2.1: Comparison of latency and energy rate distributions between Finland and Eastern USA. Previously published in Publication I [117].

and the other comprising of energy consumption and mobile app (Carat [77]) data. These are key components influencing user perceptions of apps [39, 49]. Our research shows that (i) latency and energy consumption have a strong impact on app retention and abandonment across app categories, and that (ii) the impact of performance varies depending on the level of performance people are used to and the functionality and category of the app. Our findings are supported by a performance-retention model, demonstrating the predictive capability of our approach. In addition, we demonstrate the power of data fusion to provide new insights that cannot be obtained by analysing datasets individually. Studying app retention improves the understanding of the relationship between user behaviour and factors affecting performance have significant academic and commercial interest. For instance, by exploring this relationship, we can gain deeper insights into mobile interactions and their contextual influences [5, 8, 22, 79]. Additionally, marketers can benefit from understanding the key factors that drive app success or failure [14, 31], while developers can use this knowledge to improve their apps effectively [4, 85].

In the experiments, we focus on the locations with the highest amounts of data: Finland and USA (EST - Eastern Standard Time). For the NetRadar dataset, we focus on the samples collected during cellular network connectivity as this allows to assess performance variations [23] and capture a broad range of usage contexts and higher spectrum of mobility patterns. As unit of analysis we consider *network latency*. The difference of the latency distribution between the two locations (see Figure 2.1a) suggests differences in network infrastructure, or mobile subscriptions, within the two locations. For the Carat dataset, we restrict our analysis to samples collected from

Table 2.1: Importance of performance on retention for top 5 categories (Cat) and mobile applications (App). Darker colours reflect statistical significance ($p \leq 0.05$). C: Communications, P: Productivity, T: Tools, S: Social, M: Music, w: Whatsapp, fm: Facebook Msg., fa: Facebook, d: Dropbox, t: Twitter. Table adapted from the published Publication I [117].

		Significance: Energy			Significance: Latency				Significance: Energy			Significance: Latency		
Location	**Cat.**	**Day 1**	**Day 7**	**Day 15**	**Day 1**	**Day 7**	**Day 15**	**App.**	**Day 1**	**Day 7**	**Day 15**	**Day 1**	**Day 7**	**Day 15**
Finland	C	0.028	0.005	0.006	0.003	0.685	0.921	w (C)	0.028	0.040	0.011	0.007	0.066	0.109
EST-USA		0.019	0.026	0.244	0.313	0.001	0.053		0.315	0.194	0.724	0.460	0.293	0.125
Finland	P	0.654	0.600	0.378	0.033	0.001	0.002	fm (C)	0.027	0.107	0.007	0.203	0.638	0.381
EST-USA		0.263	0.636	0.756	3.0E-04	1.0E-04	0.584		0.050	0.017	0.011	0.186	0.096	0.158
Finland	T	3.0E-04	0.001	1.0E-04	2.0E-04	1.0E-05	0.007	fa (S)	0.239	0.431	0.022	0.009	0.001	0.002
EST-USA		0.059	0.005	5.0E-04	0.499	0.029	0.016		0.013	0.004	0.009	0.035	0.050	0.010
Finland	S	0.223	0.284	0.027	0.010	0.246	0.0669	d (P)	0.665	0.478	0.792	0.039	0.004	1.0E-04
EST-USA		2.0E-04	4.0E-05	0.003	0.0612	0.022	0.099		0.377	0.216	0.134	0.105	0.074	0.313
Finland	M	0.004	0.050	0.304	0.803	0.288	5.0E-04	t (N)	0.089	0.040	0.147	0.231	0.065	0.232
EST-USA		0.027	0.389	0.908	0.244	0.007	0.013		0.471	0.033	0.077	0.030	0.041	0.198

Android devices due to their accessibility and high sampling granularity. As unit of analysis we consider *energy rate* which correspond to the relative change in battery in a given time interval [77]. The two locations show similar energy consumption distributions (see Figure 2.1b), but differ in terms of application usage patterns.

Data fusion is performed using a combination of timestamp and coarse-grained location information, including Mobile Country Code (MCC), Mobile Network Code (MNC), and GPS reverse geocoding. Since the datasets have different sampling periods, we align the records by creating hourly bins and mapping each sample in NetRadar and Carat to the closest bin. The combined dataset comprises 243 applications, 1241 users, over 1 million latency measurements and over 2.8 million energy measurements. Data fusion validity is evaluated by comparing statistical characteristics between combined and individual datasets, as well as sample distributions.

The main results of the study, as summarized in Publication I, include (i) evaluating the effect of performance on retention, (ii) determining the level of *critical point* in performance, (iii) assessing the difference in the effect of performance, (iv) understanding the effects on highly-rated apps and other factors, and (v) analysing the combined effect of Latency and Energy. In the first case, we observe that performance certainly impacts retention, but this relationship is influenced by application category and app popularity. Different interaction patterns are observed across application categories. Users exhibit varying levels of tolerance for poor performance based on the application category and duration of use (see Table 2.1).

We identify *critical points* of performance where a decrease in performance leads to lower retention. Improving performance beyond these points has no impact on retention. The analysis of the critical point reveals that ap-

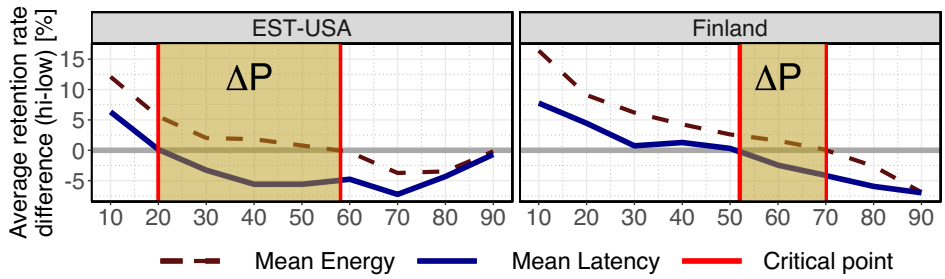


Figure 2.2: Overall average retention difference for *high* and *low* groups combining latency and energy "critical point" thresholds. Previously published in Publication I [117].

plication category, application characteristics, and location play a significant role in determining these point. Latency and energy have varying effects on retention depending on location, with differences potentially attributed to network infrastructure and user preferences. In Figure 2.2, when comparing the difference in critical points (CP) between energy and latency, we observe that the effect of latency is perceived earlier, with energy becoming significant at later percentages. Latency, being a short-term phenomenon, is noticed faster by users and can also affect energy consumption, leading to a faster decrease in retention. In Publication I we show that this trend is more pronounced for individual apps compared to categories.

We extend our analysis to the top-rated apps, demonstrating that latency has a greater impact on apps that heavily rely on online content display, while energy variations are more influential for personalization apps. Critical points, which indicate user tolerance, tend to align for apps with similar functionalities. Our findings also demonstrate the robustness of performance effects on user perceptions and retention across different factors. Analysing energy data from a newer sample, we observed a consistent and increased importance of energy on retention. We also examined low-energy apps, confirming the direct impact of performance on user perceptions and the persistent effect of energy. Finally, the cost-benefit analysis to evaluate the combined effect of energy and latency on retention shows a non-linear relationship between latency and energy, with neither variable clearly dominating the other. Publication I also contributes by modelling the degree to which performance affects retention, which is important to estimate how users will respond to apps during their evolving life-span. The results for individual factors show successful retention value prediction values for Finland, but have slightly higher error rates in cross-country

scenarios. Mixing data from Finland and EST-USA significantly improves the model's performance, with accurate predictions for all categories and factors. Categories with similar usage patterns shows consistent results across countries. Combined factor prediction leads to a significant reduction in error compared to individual analysis.

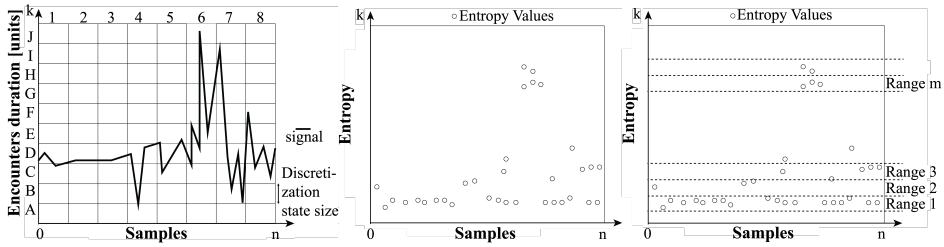
Overall, Publication I motivates the need to explore more about mobile interactions and how they are influenced by the context, specifically studying collaborator selection and collaboration opportunities in large scale environments. These topics are explored in Sections 2.2 and 2.3.

2.2 Collaborator Selection

Pervasive Data Science harnesses the pervasive availability of programmable smart devices. This availability allows for collaborative computing scenarios between multiple users or devices. Finding optimal and appropriate collaborators is critical for the success of these scenarios as the costs of finding and managing collaboration may otherwise offset the benefits they bring [2, 45, 59]. The difficulty on finding collaborators arises from the impact of human movement and the spatio-temporal context, which affects the accessibility of devices for collaboration [6, 67].

Publication II presents COSINE, a novel approach for selecting collaborators in multi-device computing scenarios. The challenges in collaboration selection is to adapt selection to different types of collaborations or task characteristics, and to account for individual variations in mobility. COSINE overcomes these issues by using novel information theoretic measure based on Markov trajectory entropy [26] to efficiently rank and recommend collaborators. COSINE significantly improves the benefits of collaboration by (i) increasing the expected duration of collaboration and (ii) reducing the variability of collaborations. We focus on collaborative multi-device sensing and computing as this is a critical part of PDS to create global platforms that provide high quality services based on device interaction. Compared to the existing research [54, 55, 63], our approach explores how to find the best possible collaborators for the current context, rather than being limited to analysing how to support collaboration. This allows maximising the benefits and characteristics of collaboration based on the context where it takes place. Indeed, recommending collaboration opportunities that have the most predictable duration facilitates scheduling and allocation of tasks across the available devices.

COSINE uses Markov trajectory entropy as measure of regularity. This allows to overcome the limitations related to (i) incapability to adapt to



(a) (b) (c)

Figure 2.3: Phases of our COSINE for selection of collaborators finding suitable collaborator devices. (a) Selection of collaborators, (b) Extraction of regularity, (c) Quantization of measurements. Available devices and collaboration duration. Figure previously published in Publication II [34].

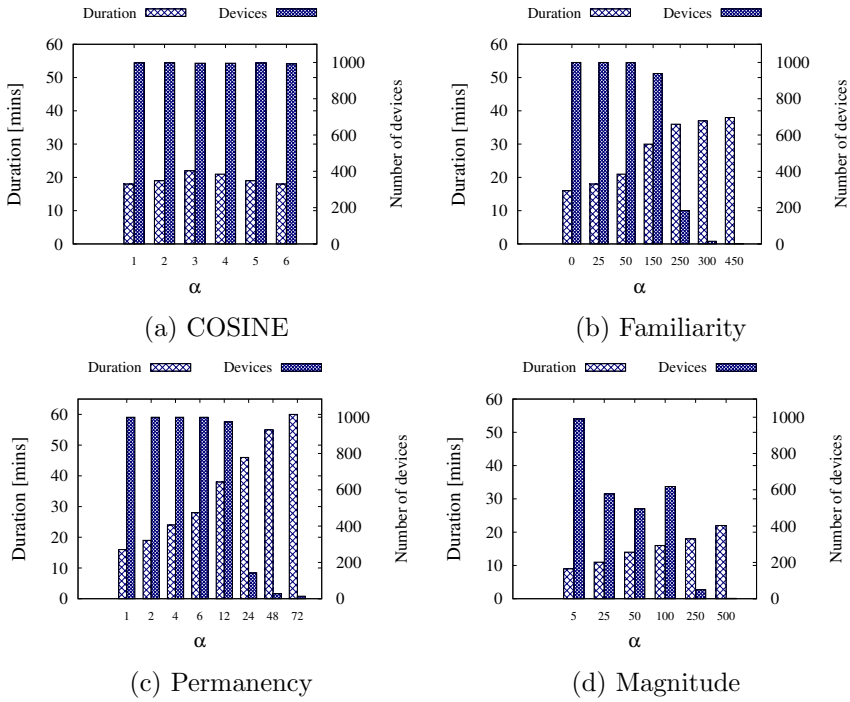
varying mobility patterns; and (ii) incapability to account for different task and collaboration characteristics. The overall approach for collaboration selection consists of three phases, which are illustrated in Figure 2.3. The first phase is signal quantisation where duration values for device encounters are quantized into a finite set of states. This allows us to represent the signal as a discrete trajectory and measure regularity through state transitions (see Figure 2.3a). Quantization is performed using k-means clustering with the desired number of quantization levels, given by the the number of clusters k. In the second phase, regularity extraction (see Figure 2.3b), we quantify encounter regularity using a Markov chain approach, creating a transition probability matrix to measure the consistency of the encounters and identify device pairs with a predictable duration. By constructing a Markov trajectory entropy matrix, it is possible to assess the regularity of each encounter pair. The matrix is obtained by estimating the probabilities of transitions between different states (quantized duration values) and then calculating the Markov entropies. The last phase is the selection of collaborators (see Figure 2.3c). We establish entropy ranges and rank them based on frequency and duration, selecting collaborators with consistent availability and prioritize those with longer duration.

We utilize a large-scale crowdsensed dataset collected by a cellular operator in Shanghai over a one-week period as our primary data source. This dataset contains real-world mobility traces and app usage patterns from 137 495 devices and 10 363 base stations, ensuring the data is representative of human mobility and enables the study of collaboration opportunities in a large context. The dataset provides coarse-grained information about base stations in the metropolitan area and session details of user connec-

tions, including device identifiers, session duration, data transferred, base station IDs, and GPS coordinates. In our experiments, we focus on a 20 km^2 area with the highest density of users and the 1000 users with the highest amount of samples ($\approx$ 1 000 000 D2D encounters). We separately assessed the representativeness of the data, showing that distribution of encounters has the same shape as other datasets that investigate device collaboration [62]. We focus on collaborations lasting at least five minutes, which is the minimum time for a sensing task to be beneficial in a collaboration between two devices [62], and exclude base stations and users with measurement counts below the 90^{th} percentile. Sessions are aligned and matched between users at hourly intervals. Our approach is compared against three baselines that specify a threshold condition α for selecting collaborators, and evaluate the changes in collaborations for different values of α. The first baseline considered is *Familiarity*, which is determined by the frequency of device encounters, with α representing the number of encounters. Higher familiarity corresponds when devices interact frequently. Next baseline is *Permanency*, which measures the duration that devices stay within proximity. The collaboration time required for a candidate device is denoted as α. The last baseline, *Magnitude*, represents the α number of devices needed to distribute a task.

The findings of our evaluation, summarized in Figure 2.4 and presented in Publication II, illustrate that device encounters contain adequate regularity to serve as a valuable source for identifying collaborations. Our method also effectively detects collaborations characterised by prolonged duration and predictability. The results include: (i) validation of the process of regularity extraction from candidates, (ii) assessing the selection of collaborators, (iii) comparing the results of COSINE with other selection approaches and (iv) evaluating selection performance in different contexts. Entropy values are used to measure regularity, with low entropy indicating consistent encounter duration and high entropy indicating fluctuating encounter duration. Three representative traces with different encounter frequencies are selected for analysis, i.e. high, medium and low encounter frequencies. The results show that regularity, represented by entropy values, can be used to characterise collaboration candidates and their encounter duration.

We demonstrate that COSINE can identify collaborators with long and stable periods of proximity. By selecting devices optimally, collaborations of up to 22 minutes can be achieved. Other α values result in shorter durations of collaboration, despite a similar number of devices being available. When comparing COSINE with other approaches, we observe that a similar distribution of collaboration opportunities between all the approaches, indi-



(a) COSINE (b) Familiarity

(c) Permanency (d) Magnitude

Figure 2.4: Selection of collaborators for the considered baselines for various α-values. Figure previously published in Publication II [34].

cating consistency in collaboration principles over time. The variations in collaboration opportunities based on different collaboration requirements are illustrated in Figure 2.4, showing, for example that duration of collaboration opportunities increases with higher familiarity between devices. However, as the familiarity requirement increases, the likelihood of finding collaborators decreases. The Mean Absolute Deviation (MAD) for regularity remained consistent at 5.66 across all duration values, while Familiarity and Permanency showed higher variance (MAD = 13.74), making them unreliable for identifying long-term collaborations. Finally, we demonstrate the adaptability of COSINE to select collaborators in different context. The results of COSINE are consistent even when considering specific locations with oscillating mobility patterns, like a park.

We demonstrate how collaborator selection influences overall system design, but also highlight how different contexts affect the characteristics of available collaborators. This motivates us to further explore how human mobility and contextual spatio-temporal differences impact collaboration opportunities. We discuss this in Publication III, summarized in Section 2.3.

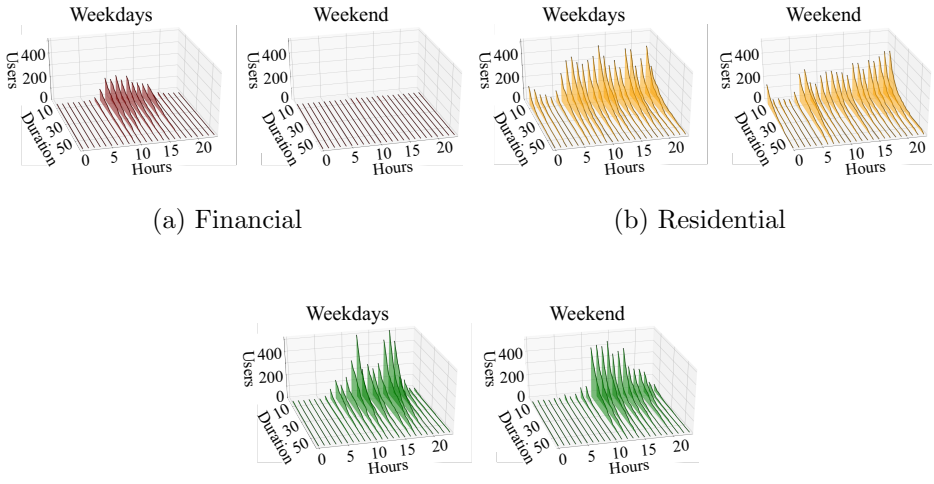
2.3 Collaboration Stability

Collaboration is crucial in Pervasive Data Science for accomplishing tasks specially in large-scale multi-device environments [59]. Stability plays a vital role in collaboration as it ensures consistent execution of shared tasks over time. Without stability, a collaborative application or service can experience a degradation in performance or even a complete failure [101].

Currently, there is a limited understanding of the attributes that govern collaborative opportunities, making it challenging to establish the circumstances in which participatory applications are likely to succeed or fail [41, 90]. Publication III solves this lack of understanding by quantifying the collaborative stability of human mobility and characterising its impact on the ability to find a reliable set of collaborators for multi-device computing and networking scenarios. Our contribution includes (i) presenting representative applications that benefit from collaboration stability, (ii) introducing a general model for collaboration stability based on the variation of collaboration opportunities, and (iii) empirically investigating the impact of stability on networking applications. The evaluation is carried out using a representative large-scale dataset of device-to-device encounters in everyday contexts. Our evaluation reveals a strong correlation between collaboration opportunities and the spatio-temporal context in which collaboration occurs. Furthermore, our study highlights the significance of stability in collaborative sensing applications within PDS, offering a way to significantly enhance the selection of collaborators. For example, understanding the stability of devices in a context allows efficient schedules to be set for task execution based on their computational complexity and processing requirements.

Collaboration stability can be defined as the persistence of collaborative or cooperative interaction patterns. The requirements for stability are likely to depend on the spatio-temporal characteristics of a location [100] or application. We define *collaboration stability* formally for an application A as the function $f_A(d, t, s)$ where d is the number of collaborators (devices), t is the temporal context for collaboration, and s is the spatial context. Understanding the stability can be highly beneficial for a wide range of multi-computing applications, such as contact tracing, fog computing, autonomous vehicles, and crowd computing [70, 80].

We use the same large-scale dataset of one-week real-world mobility traces that we used to study collaborator selection (refer to Section 2.2). By using data collected from a mobile operator, it is possible to ensure that the patterns of human mobility and interaction captured are representative of real-world scenarios of multi-device applications. We focus our analysis in the same 20 km^2 area as in Publication II. This guarantees having a



(a) Financial

(b) Residential

(c) Shopping

Figure 2.5: Availability of users for D2D collaborations over time in different contexts. Figure adapted from the previously published Publication III [37].

large-scale representative amount encounters for analysis, i.e., $\approx 1\,000\,000$ encounters, and cover six areas that differ in terms of user concentration and land use, and are representative of locations where encounters are part of everyday life routine [110]: Residential, Pubs, Park, Shopping, Train Station and Financial.

Our experimental design is two-fold: (i) quantifying collaboration stability and analysing the factors influencing it, and (ii) evaluating collaborator selection using collaborative sensing as a case study. We first analyse collaboration stability in human mobility by monitoring continuous proximity between devices, quantifying stability in different areas and parts of the week. Daily patterns of users are used to compare stability in different areas and evaluate its potential to be used for predicting the classification of an area. Next, we compare two methods for selecting collaborators: familiarity [62] and regularity (see Section 2.2). Both use a threshold parameter to determine collaborators, applied either on encounter frequency or encounter entropy, respectively.

The results of our experiments are summarised in Figure 2.5 and presented in Publication III and show that collaboration stability is highly dependent on context with both the type of location and the time-of-day affecting it. Understanding collaboration stability can help improve the selection of collaborators. We first examine how the number of devices

Table 2.2: Stability for different contexts, duration time (Dt) and number of (Nd) available devices based on different daily routines (R). Table previously published in Publication III [37].

R	Category (time window)	Time range	Financial		Residential		Shopping	
			Dt	Nd	Dt	Nd	Dt	Nd
1	Daytime	08:00-20:59	47	117	56	339	57	299
	Nighttime	21:00-07:59	14	5	53	192	36	28
2	Rest (Early Morning)	01:00-08:59	18	16	51	170	32	35
	Work (Morning-Afternoon)	09:00-16:59	53	161	56	342	57	267
	Leisure (Afternoon-Evening)	17:00-00:59	27	21	57	322	55	250
3	Rest (Early Morning)	01:00-05:59	5	2	46	82	17	7
	Rush hours (Morning)	06:00-07:59	34	18	59	297	59	56
	Work (Morning)	08:00-11:59	55	166	54	336	57	190
	Lunch break	12:00-14:59	55	161	57	393	56	344
	Work (Afternoon)	15:00-17:59	49	112	57	279	57	323
	Rush hours (Evening)	18:00-20:59	27	14	55	350	57	377
	Leisure (Evening)	21:00-00:59	15	2	58	279	48	41
		Total	**33**	**66**	**55**	**282**	**49**	**185**

available for collaboration varies across contexts, confirming the influence of location and time on collaboration. This can be observed in Figure 2.5, which shows that the number of devices is higher during weekdays than weekends for all the areas, indicating a higher device usage and higher degree of user activity. Collaboration groups during the weekend are more focused on supporting short-term tasks, as the weekend has lower stability levels. The availability of devices also depends on the time of day, indicating that the stability of collaboration in a location is dynamic.

We assess the differences in the stability of collaboration in different domains. We model stability based on mobility patterns that are indicative of users' routines [111]. Collaboration stability changes significantly depending on the granularity of daily routines (see Table 2.2). More granular routines provide information that is more consistent with the type of area and human mobility. For example, as the second routine (Work) shows, the Financial location has most activity during working hours (9:00-16:59). However, a more detailed quantification (i.e. the third routine) shows that the likelihood of having more and longer collaborations is higher during the morning working hours (8:00-10:59). The most consistent stability of collaboration occurs in the Residential area.

When assessing whether stability is a factor that can be used to predict the context, and specifically the type of location, we observe that stability is a characteristic of the location. The classification performance for predicting the context using stability and location as input features in simple ML algo-

rithms reaches 70% classification accuracy. This suggest that applications used in specific contexts must align their collaboration patterns with the context's characteristic patterns to ensure successful collaboration. Stability improves selection of collaborators in multi-device sensing and computing tasks. As expected from the results of Publication II, the duration of collaboration opportunities increases with the level of familiarity. Applications that require high familiarity can only benefit from short collaborations. When regularity is used, the duration that the collaborators are available is more stable, meaning that selection using devices that have higher collaboration stability can improve system stability.

Overall, these results highlight how the availability of devices varies over time and location. The suitability of tasks for different locations also varies, e.g., train stations are better suited for collaborative sensing rather than collaborative computing due to the lack of support for stable groups that have long duration. Our analysis highlights that understanding stability is critical for making optimal decisions about scheduling collaborative tasks and predicting their success or failure based on context.

Chapter 3

Low-Cost PDS Systems for Indoor Monitoring

In Chapter 2, we demonstrated how PDS systems can take advantage of the deployment of a massive amount of devices to collect and analyse large amounts of diverse data and establish stable multi-device sensing collaborations. Although these devices can provide valuable data, their use is usually limited to the specific application for which they were originally installed. Exploring the possibility of using these devices for additional applications can bring benefits, such as gaining deeper understanding of the environment, complementing existing systems and efficient use of energy.

This chapter demonstrates the potential and the benefits of re-purposing low-cost sensing devices to provide an easy and cheap-to-deploy alternative for real-world tasks and for augmenting existing technologies. We also discuss the challenges associated with the costs of deploying low-cost sensors. The presented use cases, which are part on the contributions of the Thesis, focus on indoor monitoring, as humans tend to spend a significant amount of their time in indoor environments [56]. The quality of indoor environments is linked with human health, productivity, comfort, and quality of life in general. It is therefore necessary to provide efficient sensing solutions that can effectively support the monitoring of indoor environments.

When re-purposing low-cost sensors is their ease of implementation and ability to offer valuable data for addressing emerging challenges. In Publication IV, we show how existing low-cost environmental sensors can be integrated into plants containers to provide information about the environmental conditions of indoor areas. Plants are a common sight in indoor spaces to decorate and improve the diversity of the environment that bring several well-being benefits [10, 66] and are increasingly integrating sensors for monitoring plant condition. Through comprehensive experiments, we

demonstrate the potential of smart plants to estimate overall CO_2 accumulation, detect occupancy, and identify the usage of face masks. These are representative examples of the applications that can be implemented by re-purposing smart plant sensors (see Section 3.1).

Accuracy and precision are crucial factors to validate the reliability of the measurements. Low-cost sensing devices are normally equipped with sensors that compromise accuracy and precision for smaller size and lower energy drain [83, 109]. These factors can be also significantly affected by the sensor configuration and by how the sensor is physically installed. In Publication V, we assesses the performance trade-offs and associated limitations of low-resolution thermal array sensors with varying sensor resolutions. These sensors are a cost-effective solution for detecting and counting occupants in indoor environments. Compared to other monitoring solutions, the key benefits are energy-efficiency and non-intrusiveness [7, 17, 92]. However, the performance of these sensors is significantly affected by the camera's resolution, frame rate, and field-of-view [103]. To address this, we determine the minimum resolution and frame rate required for reliable detection, and provide guidelines for determining optimal sensor resolution and configuration in real-world deployments.

We conclude this chapter by exploring how existing sensing methods can be used to implement new applications. We extend our analysis of thermal sensors from occupancy monitoring to privacy-preservation to privacy preserving applications. Publication VI demonstrates how thermal imaging integrated into off-the-shelf consumer devices can be combined with simple processing pipelines to offer a non-invasive approach to accurately detect covert surveillance devices in a wide range of environments and settings. We investigate factors such as, distance to other electrical objects, environment, luminosity, camera type, and partial occlusion (see Section 3.3). We show that thermal imaging can detect covert devices more easily and efficiently than the conventional techniques in everyday contexts.

3.1 Smart Plants for Indoor Monitoring

Publication IV presents the idea of utilizing smart plants as a cost-effective and easily deployable solution for monitoring indoor environments. Plants are typically used in close proximity to people and are increasingly being placed in containers that incorporate low-cost sensors that are used to monitor plant growth and environmental conditions. In the publication, we demonstrate how these sensors can serve as an alternative technology for monitoring and enhancing indoor environments, eliminating the need

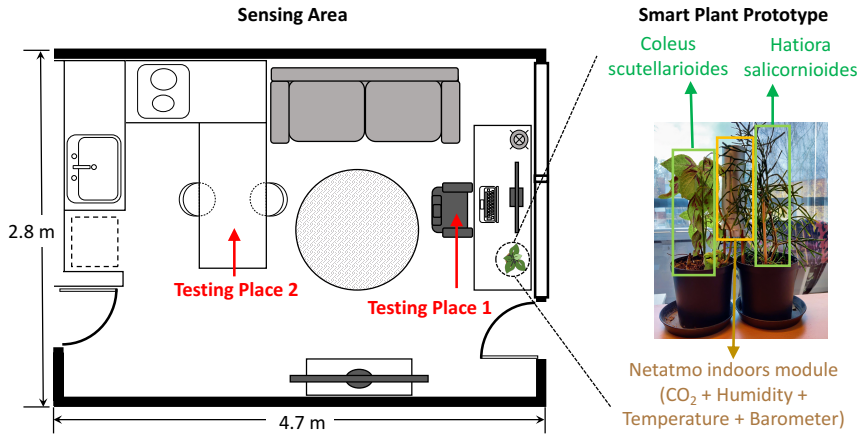


Figure 3.1: Experimental Setup of Smart Plant Prototype. Previously published in Publication IV [119].

for separate, often proprietary, sensors. The contribution comprises of two parts: (i) showing how smart plants can be used to estimate overall CO_2 accumulation, occupancy information, and whether people use protective face masks or not, and (ii) establishing a research roadmap that identifies the key challenges in employing smart plants to monitor indoor environments.

In general, the characteristics and variety of sensors change depending on the context of use. Solutions for monitoring individual plants [78, 104] integrate fewer sensors than solutions for monitoring greenhouses, gardens, and grow rooms [58]. Miniaturisation and cost reduction in the production of sensors can reduce the gap between these solutions and enable more sensors to be integrated into plant containers for individual plants. We build on this direction and consider a design that combines aspects of both types of sensor devices integrating a portable weather station (Netatmo) with a plastic plant container. The Netatmo station is equipped with sensors that measure carbon dioxide levels, temperature, barometric pressure, and relative humidity. The sensors were calibrated in accordance with the manufacturer's guidelines. The collected data is transmitted from the weather station unit to cloud services through WiFi communication. This connection allows access to the device configuration and download of measurements from the cloud.

To assess the benefits of our design, we collected measurements with different time granularity and in different experimental conditions to estimate how well the sensors in the container capture variations in air quality. Figure 3.1 illustrates the test environment, which corresponds to an area

that is representative of a typical living or office space. The plant container (pot) was placed in a fixed location. For the experiments we consider ornamental plants that are easy to grow, care and are commonly placed in indoor spaces: Coleus scutellarioides (common name: painted nettle) and Hatiora salicornioides (common name: bottle cactus). The greenery was planted in separate pots covering the sensors used for monitoring to ensure a realistic monitoring context. The ground truth corresponds to the mean value of CO_2 when the indoor area is empty. The measurements are then pre-processed using simple outlier removal. Feature extraction involves statistical summaries of the measurements for different time window sizes. These features are used for growth curve analysis and for modelling face mask use and occupancy estimation. Between experiments,CO_2 levels are returned to ground truth levels. Samples were collected in different experimental conditions, including varying the number of participants (0–2), the time of data collection (3–h, 6–h, 1-day), the location of the participants relative to the smart plant (80–cm, 380–cm), and the use of FFP2 face masks (none wearing face mask, both wearing face mask, one wearing face mask while the other does not).

The results of the experiments are detailed in Publication IV and include (i) validating the characterisation of CO_2 levels from the measurements, (ii) assessing the robustness of our approach to different factors affecting the sensing environment, (iii) analysing the changes in the measurements over time and (iv) assessing the possibility of modelling face mask use and occupation estimating using the characterised measurements.

We demonstrate a consistent and significant correlation between the measurements obtained from the smart plant sensor and the dedicated sensor. The collected measurements of accumulated variations of CO_2 in unoccupied spaces are not statistically different between the two devices. This suggest that sensors placed in a plant container can effectively measure indoor ambient conditions, comparable to using a separate sensor device. We observe that distance has a significant effect on CO_2 levels. Concentration is higher when individuals are closer to the smart plant (see Figure 3.2). This can be useful to make coarse-grained estimates of the distance where people are located. The number of people in the room also has a statistically significant effect on the CO_2 measurements (see Figure 3.2), showing the possibility of obtaining information about the number of occupants. In addition, smart plants can provide insight into compliance with the requirements for the use of face masks by analysing the increase in CO_2 levels (slope) over time. The rate of CO_2 accumulation can be used to identify mask use, regardless of distance. We also analysed robustness to watering

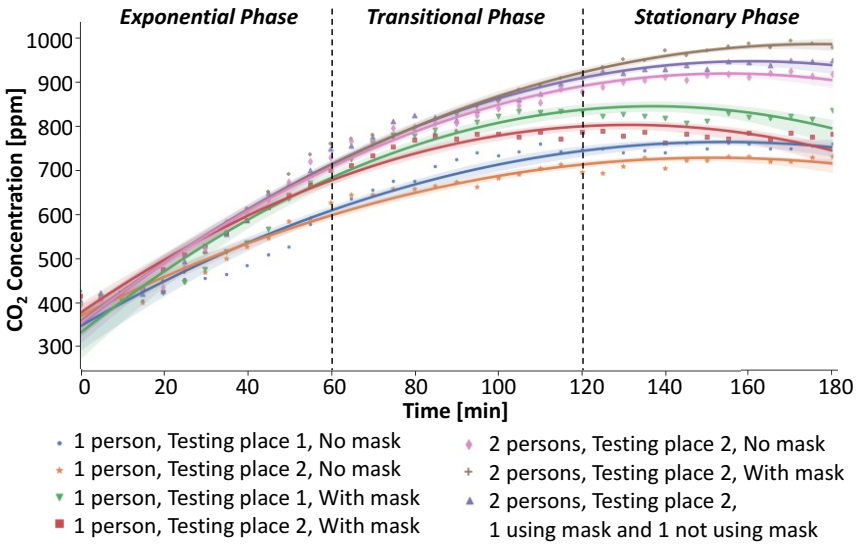


Figure 3.2: CO_2 concentration curve for different experimental conditions. Previously published in Publication IV [119].

as this can potentially affect humidity through water evaporation and plant transpiration, and CO_2 concentration. No significant differences between CO_2 concentrations were found, confirming that watering does not affect the sensor measurements.

In the publication IV we also identify three phases in the growth curve of CO_2 concentration: 1) an exponential initial growth phase; 2) a near linear transition period; and 3) a stationary phase where the concentration is saturated (see Figure 3.2). When analysing the average concentration and slope in each phase, it can be seen that face masks increase the growth rate in the exponential phase, but do not necessarily result in significantly higher concentrations in the saturation phase. The differences in the slope of the CO_2 concentrations in all phases and experimental conditions are significant. Approximately 150 min are required to reduce the amount of CO_2 back to the unsaturated level, showing that the sensors of the smart plant can also provide information to support proper ventilation in an area.

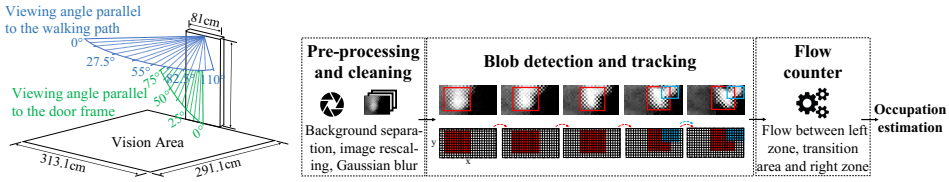
Finally, we demonstrate that smart plants can support coarse-grained classification of face mask use and indoor occupancy estimation. For the face mask classification, a model based only on CO_2 concentration achieves an average accuracy of 65%. A more informative feature model increases the classification accuracy ($\approx$ 70% when temperature measurements are added). The main source of error is the case where mask use is mixed

between occupants. When estimating indoor occupancy, the models achieve higher results. The accuracy ranges from $\approx 85\%$ with a model based only on CO_2 levels to $\approx 89\%$ when incorporating both CO_2 levels and temperature. The highest misclassification rates occur when distinguishing between one or two occupants. The inclusion of time window information does not significantly affect the model performance across all time windows in both cases. Overall, the classification results show that smart plants are able to model different indoor scenarios and provide valuable insights into the indoor context. However, the performance depends on the complexity of the task and the characteristics of the space.

3.2 Thermal sensing for Occupancy Detection

Human occupancy data is crucial for building management systems as it enables optimisation of energy usage and ensures a comfortable and healthy indoor environment. Low-resolution thermal array sensors are a promising solution for occupancy detection and counting. However, a major challenge with low resolution sensors is their sensitivity to measurement parameters, such as overall resolution, frame rate and camera field-of-view, which significantly affect their performance [103]. Publication V contributes with a systematic evaluation of low-resolution thermal array sensors for occupancy detection, considering performance trade-offs and costs such as privacy loss and deployment cost. We envision an increase in the pervasiveness of low-resolution thermal sensor arrays as a potential solution for supporting building management systems. Through our investigation, we are able to determine the necessary minimum resolution and frame rate for reliable detection. Additionally, we analyse the impact of different viewing angles on performance. Furthermore, we provide guidelines for selecting the most suitable sensor parameters based on deployment specifications, taking into consideration various performance and cost criteria.

Our evaluation considers two datasets, an open source dataset of tripwire-triggered thermal images (TIDOS [17]) and a proprietary dataset used to validate the generality of the findings and to investigate the effect of differences in camera field of view. Both datasets were collected using Melexis MLX906401 thermal array sensors, which provides a resolution: 32×24 pixels and two types of fields-of-view (FOV): standard ($55° \times 35°$) and wide angle ($110° \times 75°$). The design of the sensing pipeline consists of using the sensing units in different configurations to collect thermal radiation measurements of the environment. The measurements were collected by installing the thermal camera above the doorway of different occupation



(a) (b)

Figure 3.3: (a) Testing environment and (b) processing pipeline for occupancy estimation. Previously published in Publication V [87].

scenarios. For TIDOS, the thermal camera was located at a height of 240 cm, the side of the larger viewing angle was parallel to the door frame and the vision area was 249.9 cm × 151.34 cm. For the controlled testbed measurements, the dimensions of the door frame are 205 cm × 81 cm and the vision area on both sides of the door was 313.1 cm×291.1 cm (see Figure 3.3a). Ground truth was based on manual annotation of the measurements for both datasets. We follow similar cleaning and pre-processing steps to state-of-the-art occupancy counting systems using thermal array sensor data [15, 17]. The steps include: scene separation, deletion of long sequences with no activity, background removal, image rescaling and Gaussian blurring. These methods reduce the effects of noise and improve the robustness of occupancy detection and counting. Occupancy detection and counting is estimated using a blob detection algorithm, which involves detecting, tracking and counting the flow of blobs by evaluating blob transitions in successive frames. Feature extraction considers three parameters that impact on the processing cost and the cost of the cameras themselves: image resolution, frame rate and field-of-view. We use this information as input in to a Tree-Structured Parzen Estimator (TPE) algorithm to find the combination that maximises the accuracy of the occupancy count. This information is used as input to identify lower bounds for resolution and frame rate to establish a minimum point at which reliable detection is possible. The sensing pipeline is illustrated in Figure 3.3b.

We next summarise the results of Publication V.When analysing the impact of camera parameters, frame rates between 4 and 16 generally perform very similarly and do not have a significant impact on the overall performance of the occupancy counting algorithm. The counting accuracy is consistently around 90% for resolutions higher than 2 × 2, which reduces accuracy by around 10% (see Figure 3.4). Lower resolutions allow detecting only one person at a time, but provide better privacy protection. Analysis of viewing angle shows that for horizontal angle (see Figure 3.5b) a viewing

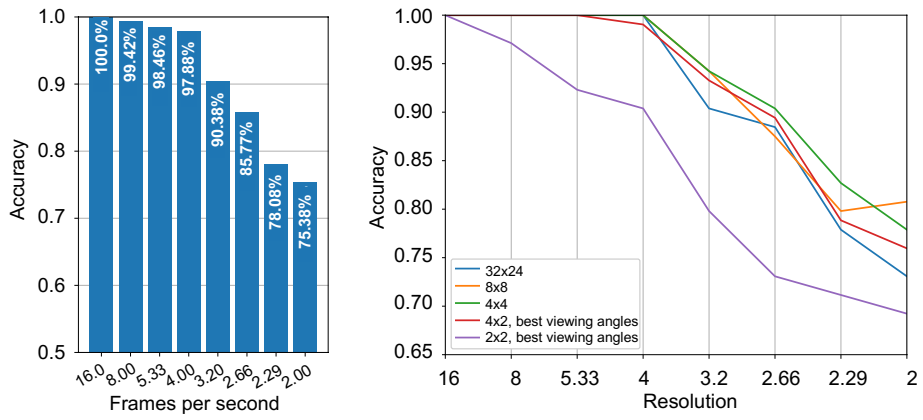


Figure 3.4: (a) Average counting accuracy of different frame rates and (b) counting accuracy at different resolutions and for different frame rates in the self-recorded dataset. Previously published in Publication V [87].

angle of 50° works optimally with resolutions of 2×2 and 4×4 whereas viewing angles between 50° and 75° provide the best results when the resolution is 8×8. For the vertical angle (see Figure 3.5a) all resolutions perform well when the angle is between 42° and 55° as long as the frame rate is at least 4 FPS. The viewing angle of a camera depends on the door area configuration and the height of the door frame. Optimal results are achieved when the camera captures the immediate area beyond the door frame, monitoring the physical entry and exit point. Higher resolutions provide wider angles, while lower resolutions work if aligned with entry/exit. Next, we evaluate the count error for different scenarios. A resolution of 8×8 and frame rates equal to or higher than 4 frames per second provide a lower counting error. Lower frame rates contribute to a higher number of missing scenes. Our analysis indicates that scenes with a larger crowd size are more prone to counting errors. A frame rate below 4 fps leads to an increased number of missed scenes, likely due to the absence of rapid blob transitions or the merging of multiple blobs during detection.

A trade-off analysis of thermal array sensors was conducted using various real-life smart deployments. The optimal density of sensor deployment depends on the application scenario [7, 75]. We consider two smart offices [71, 73] that include rooms for rooms, rest and study. Although the comparison focused on smart offices, the results can be applied to residential buildings as well. The recommended deployment for the first space ($2431.52\,\mathrm{m}^2$) to continuously monitor occupancy and ensure continuous coverage would be

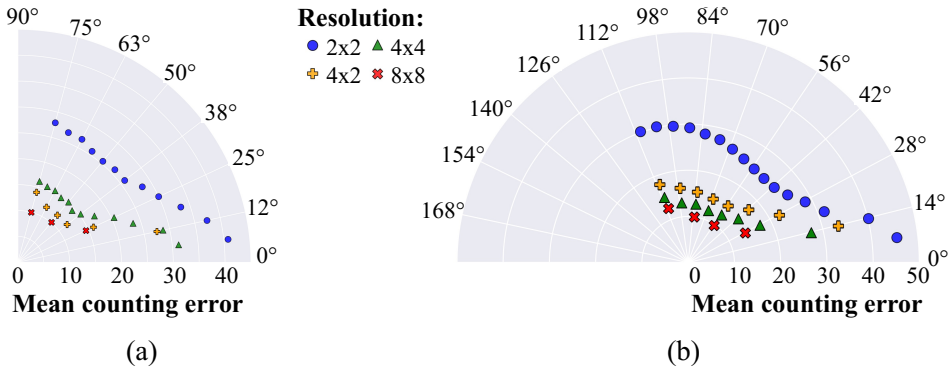


Figure 3.5: (a) Mean counting errors for different degrees (a) parallel to the door frame and (b) parallel to the walking path in the self-recorded dataset using different resolutions. Previously published in Publication V [87].

486 sensors (currently 352 sensors), while for the second space (400 m^2) the recommended deployment would be 80 sensors (currently 10 sensors). Supplementing current deployments with thermal sensor arrays would cost much less than using other sensor technologies. The difference would be up to \$116 580 in the first case and \$465 600 in the second, depending on the sensor technology. The results show that low-resolution sensors can provide significant savings for occupancy detection applications, while helping to protect occupant privacy.

Our vision is to enable environments with a mix of low and high resolution sensor technologies so that they can complement each other to achieve efficient monitoring in multiple contexts. Finally, we discuss factors that can affect our findings. In our experiments, we focused on performance and cost trade-offs without analysing the processing pipeline itself in detail. In general, the performance is comparable to solutions using a more complex pipeline or sensing method [15, 116]. In failure analysis, we manually reviewed the causes of failures to gain insights into performance and limitations. For example, close proximity between individuals, obstructing the floor view, resulted in system malfunctions and occupant tracking loss. In terms of computational performance, occupancy counting improves rapidly up to 1000 iterations but slows down afterwards. Using 4096 iterations ensures accurate data collection without excessive resource consumption or noise. Camera parameter experiments were compared between two datasets, with the controlled testbed offering better control and a wider range of parameters. Similar results were observed for resolution and viewing angle range in both datasets.

3.3 Thermal Sensing for Covert Device Detection

Growing concerns about covert surveillance devices highlight the need for effective detection methods as users are increasingly interested in knowing if they are being monitored [68, 94, 98]. Current detection methods include manual inspection, reflectance detection, magnetic field analysis and network traffic analysis [60, 88, 108]. These methods are vulnerable to the operational status of the devices and require significant user involvement [108, 114]. The use of thermal cameras to measure thermal emissions in the environment has been studied [32, 33, 102] but the detection of covert surveillance devices using this approach remains unexplored.

Publication VI contributes a re-purposing of thermal imaging technology for discovering covert surveillance devices. This technology that can be easily integrated into readily available consumer devices such as smartphones. The contribution (i) demonstrates the effectiveness of our approach through extensive and systematic evaluations that consider different types of covert cameras and deployment contexts, including factors that may affect detection performance and (ii) discusses how the simplicity and efficiency of using thermal imaging to detect hidden surveillance devices can support existing methods and improve the detection domain. Our approach uses thermal imaging to create a map of the thermal emissions in the environment and to uses simple processing techniques to identify the areas with thermal emissions corresponding to potential covert surveillance devices. We use thermal imaging due to (i) it is a portable solution that is increasingly available on consumer devices and (ii) it offers the possibility of detecting devices that have recently been used for surveillance [19, 27, 33].

The experimental design consist of (i) surveillance devices and (ii) a sensing device. The first component corresponds to four commercial off-the-shelf IP cameras and a smartphone. These devices are examples of consumer devices that can be used or repurposed to be used as surveillance tool, and cover representative features that are relevant for this purpose, including: video resolution (1080p, 720p), audio transmission (enable, disable), motion detection (enable, disable), power source (power cord, battery-powered) and access mode (online: wireless, offline: local storage). Communication with the devices is via WiFi making it possible monitoring the traffic over the network. Camera vendor apps are used to access the configuration and local storage of the devices. The second component corresponds to a FLIR One Pro camera equipped with a USB-C port to connect to a mobile device.

We implement a custom app that interacts with the FLIR Mobile API to extract the temperature matrix. Initially, the thermal camera is positioned in front of the area of interest to be scanned. The camera

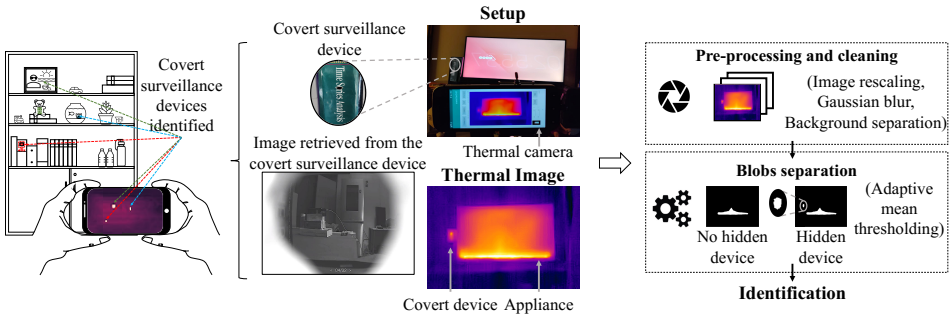


Figure 3.6: Thermal imaging targeting a covert surveillance device in a living room and ambient light conditions, and the processing pipeline for recognizing covert surveillance devices from thermal camera input. Previously published in Publication VI [120].

collects thermal emissions as a temperature matrix, which is transformed into a thermal image. To eliminate thermal noise, a Gaussian blur filter is applied, and the image is converted to grayscale. We employ the same blob separation technique described in Section 3.2 to identify regions with thermal radiation higher than certain threshold [27, 87]. These regions possess characteristic features such as temperature, shape, and size, which are indicative of the components and operating conditions of the devices being detected. The average temperature of these regions is compared to the ambient environment, and regions that significantly deviate from the it are marked as potential surveillance devices. These annotations can be utilised by other systems to enhance their detection of surveillance devices or incorporated into intelligent models to monitor the environment or identify suspicious heat signature. Figure 3.6 illustrates the sensing pipeline.

The purpose of the experiments is to assess the reliability of thermal imaging in detecting surveillance devices in various conditions and against common countermeasures. In the experiments, we cover the cameras with manufactured goods representative of common household materials that can be used to hide a camera, such as a cardboard box, a fabric shopping bag, and a thermos bottle. A 0.5 cm diameter hole was drilled in the cardboard and the shopping bag to expose the lens of the hidden camera. The experiments involve analysing of the fingerprints in two settings: a controlled testbed and in-the-wild. In the controlled testbed, we minimize heat reflection and thermal dissipation noise from other sources. In the real-world scenario, an office room is used, and furniture and other objects are employed to hide the cameras. We take measurements at intervals of 10 seconds for a total duration of 2 minutes. This helps to reduce measurement

variance and aligns with the interval between successive camera calibration cycles [65]. During the intervals between measurements, the thermal camera is kept at a controlled temperature between 2°C to 4°C for 2 minutes to prevent potential errors due to camera overheating. Overall, we examine three experimental designs to assess the effectiveness of thermal imaging in covert camera detection and analyze variations in thermal fingerprints across diverse contexts, including: (i) assessing performance under various light conditions and distances, (ii) exploring different camera operating configurations such as video resolution, audio transmission, and motion detection, and (iii) testing different camera masking materials.

Publication VI demonstrates the validity of our approach to robust detection of covert surveillance devices. We perform experiments to characterise the thermal fingerprint of the surveillance devices operating in different modes and under different environmental conditions. The results show a clear difference in the thermal fingerprint of the different cameras. The devices with extra-functionalities (e.g., climate sensing, telephony, mobile connectivity) emit higher thermal radiation due to the presence of more internal components that generate heat. A similar behaviour occurs when more features are enabled during operation. Video resolution has the largest impact on temperature, with an increase of 2°C to 6°C when using Full HD video resolution. This difference is statistically significant across all devices. These results demonstrate that thermal sensing can not only capture the thermal fingerprint of devices but also provide information about the camera's configuration mode, such as resolution, voice recording, or image recognition.

When examining the effect of light and distance on the measurements, we find that a short placement distance of the sensor is sufficient to obtain accurate measurements (two metres for an object of ≈ 1 cm size). A shorter distance increases the number of pixels that cover the object, making identification easier. Thermal signatures are more intense in darkness, thus detecting covert devices is more likely in the absence of ambient light. The results of the office environment experiments show a statistically significant reduction in the thermal fingerprint when different materials were used to cover the camera (see Figure 3.7). The difference in visibility is also significant across different materials. Materials such as cardboard and fabric, which require the lens to be exposed, allow a small area of thermal fingerprint to be visible. The thickness of the cover material simply reduces the intensity of the thermal signature, bringing it closer to the ambient environment, but the signature remains distinguishable for all cameras that were tested. Thicker glass can absorb most of the thermal radiation if the

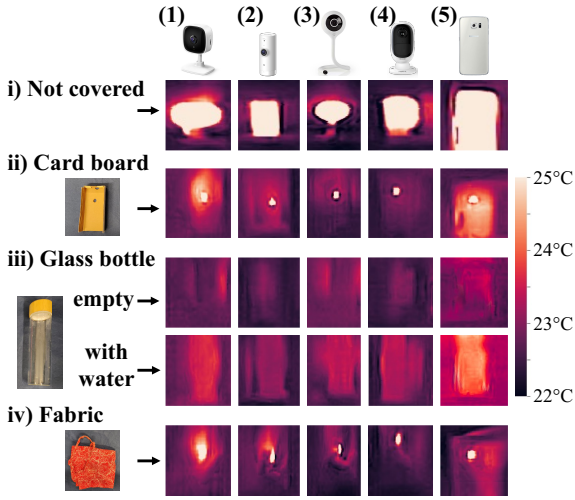


Figure 3.7: Thermal fingerprint of hidden cameras with different materials. Low-cost surveillance cameras: (1) TP-Link Tapo C100, (2) D-Link DCS-8000LH, (3) Nedis WIFICI11CWT, (4) Reolink Argus 2, (5) Samsung Galaxy S6. Previously published in Publication VI [120].

camera is not directly in contact with the glass. In addition, filling a glass container with water causes the thermal fingerprint to reappear due to changes in the container's reflective properties.

When assessing the effect of proximity to other sources of thermal emissions (i.e., TV, microwave, wall mounted radiator and fluorescent tube light), we observe that detection may require analysing differences in thermal dissipation between the objects [32] or by detecting the thermal signature as an extension of the heat source. In both cases, discovery is possible but it may require the use of more advanced processing techniques.

Publication VI also compares our approach against magnetic field-based detection and network traffic analysis. We consider the devices with the highest and the lowest thermal fingerprints: Reolink and Samsung Galaxy S6 (see Figure 3.7). Magnetic field-based detection struggles to discover distant devices or those placed behind glass. Network analysis can identify abnormal traffic conditions and device types but requires the device to be on the local network, longer monitoring, and does not provide the source's location. Overall, our approach serves as a good example of a low-cost, robust, and effective Pervasive Data Science system for detecting privacy threats in indoor environments. It can also complement existing techniques.

Chapter 4

PDS for Sustainable Development

In Chapters 2 and 3 we showed how Pervasive Data Science can be used to collect measurements at large-scale across different contexts and how low-cost devices and sensing methods can be re-purposed for innovative uses. In this Chapter, we leverage these attributes of PDS to enable applications that tackle environmental sustainability challenges. Sustainable development broadly speaking aims at balancing the needs of the current and future generations with the resources that are available [11].

Food loss and waste is one of the leading sustainability goal of the the United Nations [82]. Efforts are being made to minimise the amount of food that goes to waste at different stages of the food supply chain, such as releasing communication campaigns about food waste [42] or implementing regulations for responsible food management practices. However, still at least 30% of the food is either lost or wasted globally [82, 96] Pervasive computing and IoT technologies are being increasingly utilised in food production and agriculture to enhance productivity and reduce waste across the entire supply chain. In Publication VII, we use these technologies to improve quality estimation, a key factor in reducing food waste and identifying problems in the supply chain. We present an innovative method for characterising various organic produce using contact-based reflective green light sensing and demonstrate that our method is able to provide insights into the internal and external quality factors of produce at different stages of ripeness (see Section 4.1).

Pollution monitoring is crucial for achieving sustainable development. Pollution is a multi-ecosystem problem that produce significant risks to human health and the environment demanding urgent action for mitigation and preservation [52, 99]. The challenges of pollution monitoring are related to the limitations of current methods, which only provide information on specific pollutants, have limited spatial and temporal resolution, and ignore

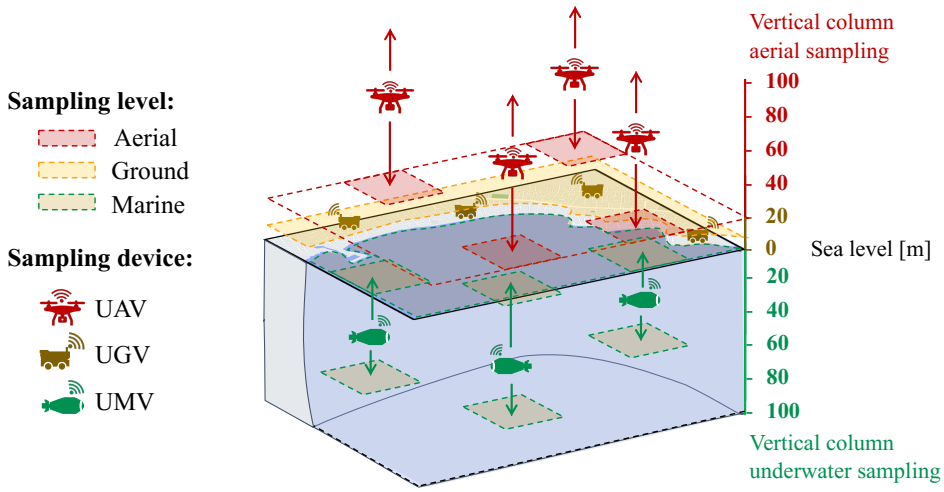


Figure 4.1: Pervasive Data Science vision for improving environmental pollution monitoring using coordinated unmanned vehicles equipped with low-cost sensors.

interactions with other ecosystems. PDS can help to improve pollution monitoring. For example, autonomous and unmanned vehicles equipped with low-cost sensors can facilitate the collection of measurements of different pollutants in different contexts and ecosystems, i.e. air, land, water (see Figure 4.1). This allows better understanding the pollution processes and the complex linkages within ecosystems. In this thesis, we focus on aerial and marine pollution monitoring, as these are less-explored fields compared to ground-based pollution monitoring [113]. Firstly, Publication VIII introduces our approach for city-scale air quality monitoring using unmanned aerial vehicles (UAVs). We show that UAVs-based pollution monitoring provides information on the vertical distribution and potential dispersion mechanisms of pollutants that can complement ground measurements and cover larger areas (see Section 4.2). Secondly, in Publication IX, we present our vision for large-scale pollution monitoring using coordinated groups of underwater vehicles (UVs). Our experiments demonstrate the potential for collecting accurate information on the properties of pollutants at high spatio-temporal granularity (see Section 4.3). Both publications also highlight the requirements, challenges and enablers for implementing these solutions.

Finally, Publication X contributes with a research vision for using deep learning (DL) in marine environments. DL models can be integrated directly into underwater vehicles or equipment carried by scuba divers, and can

provide significant benefits in supporting real-time deep-sea operations, particularly those involving computer vision. Our contribution encompasses a comparative analysis of deep learning performance for classification tasks in both surface and underwater environments. We also highlight the existing research challenges and provide a roadmap for the use of deep learning in aquatic environments (see Section 4.4).

4.1 Low-Cost Produce Quality Estimation

Food loss and waste is one leading sustainability goal of the the United Nations [82]. Despite the efforts to reduce the amount of food that goes to waste still at least 30% of the food is either lost or wasted globally [82, 96]. Quality estimation is essential in food production and agriculture to minimise waste and identify supply chain issues. Existing solutions for produce quality assessment, include visual inspection [115], automated image processing [57] or different forms of spectral imaging [51, 106]. However, these methods are difficult to adopt as they either support only specific stages of the supply chain or are expensive and difficult to operate.

Publication VII contributes with an innovative low-cost approach to fresh produce characterisation by repurposing inexpensive green-light sensors for quality estimation. The practicality of our approach is (i) discussed by showing its potential contribution at all stages of the supply chain and (ii) validated by extensive empirical benchmarks to correctly distinguish organic produce from non-organic items, establish unique fingerprints for different produce, and estimate the quality or ripeness of produce. Our approach works similarly to spectral imaging [106]. We focus only on the green light wavelength because (i) it can penetrate the surface of organic products, allowing variations in light reflection and absorption to be assessed, and (ii) it involves simple components that are affordable, energy efficient, easy to implement on low-cost microcontrollers, and widely available on commercial off-the-shelf devices.

The experimental design includes a COTS smartwatch (Samsung Gear S3 Frontier) that integrates two green LED lights and a photo-receptor, which are used for sampling heart rate information. Green light is short wavelength and effectively penetrates produce epicarp. The sensor collects photoplethysmography (PPG) values corresponding to the light reflected by an item exposed to the green light. The sensing sensing unit is placed in contact with the object being measured and collect measurements for 90 seconds. These measurements are subsequently filtered to extract statistical summaries (i.e., mean and standard deviation), which correspond to the

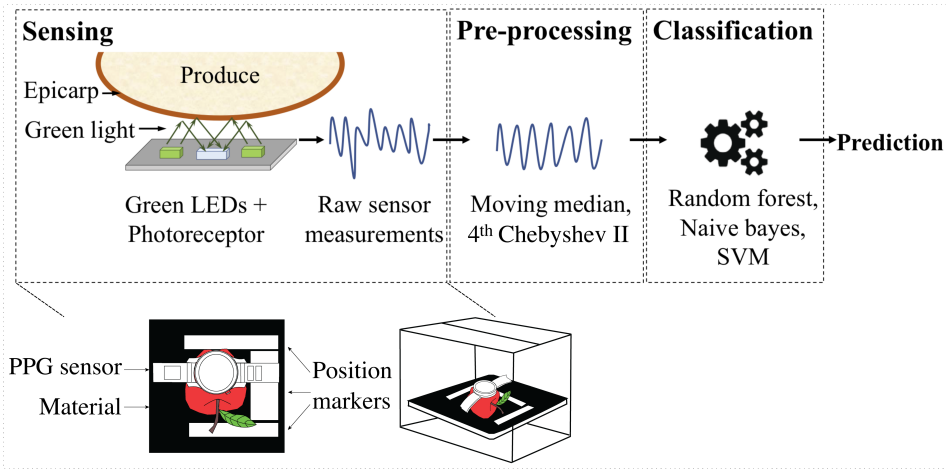


Figure 4.2: Sensing pipeline for produce quality estimation. Previously published in Publication VII [118].

object's fingerprint. These fingerprints can be used independently to distinguish between various items or item classes, or they can serve as input for ML/AI models to assess the quality of the produce [35]. The sensing pipeline is summarised in Figure 4.2.

The produce considered for the experiments are samples that cover the main categories of a structural classification of produce: berry (tomato, banana, avocado, kiwi), hesperidium (lemon, mandarine), drupe (mango, peach, plum), pome (pear, apple) and pepo (passion fruit, watermelon, melon). The inorganic objects are representative of common waste: plastic bottle (polyethylene terephthalate PET), a metal spoon (stainless steel), a small ceramic plate (feldspar), a wooden toy box (solid walnut oak) and a empty bottle of carbonated water (container glass). The controlled testbed used for the measurements (see Figure 4.2) ensures that we can carefully control the sensor position relative to the material, the luminosity level, the noise caused by possible background reflections and the room temperature.

The experimental design comprises of three parts: (i) collection of five sets of fresh produce and waste measurements to assess the performance of the characterisation, (ii) a 15-day decomposition experiment to assess the correlation between light reflectivity and changes in organic material characteristics, and (iii) an 8-day follow-up decomposition experiment to assess the generalisation of our approach to material recognition and produce decomposition capture. We use the green light sensor to collect light intensity values reported by the photo-receptor from produce and inorganic

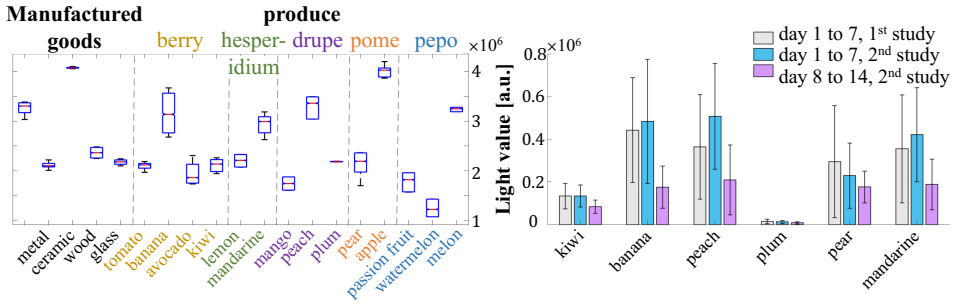


Figure 4.3: Characterization of different objects using green light sensing (left). Mean absolute deviation of intensity measurements over a 15 days period for different fruits (right). Previously published in Publication VII [118].

objects. For each material sample, the sensor takes measurements with 20 Hz frequency over a 90 second period. The information about the material and experimental condition is manually annotated after sampling individual material. Cleaning and pre-processing includes removing the first and last 15 seconds of the signal to avoid possible sources of external light and applying moving median and 4^{th}– order Chebyshev Type II filters to reduce signal noise. The extracted features correspond to the descriptive statistics of the pre-processed signal from each material and experimental condition. For the modelling stage, we use three simple classifiers to model three quality stages: ripe, overripe and decayed. We rely on simple models to minimise overfitting and ensure the models can be easily integrated and operated on energy constrained devices.

The results of the experiments are detailed in Publication VII. Firstly, we demonstrate that the variations in light intensity among different types of items can establish a unique fingerprint for each item, enabling the distinction between them. Characterisation results are presented in Figure 4.3 (left part) and show that there are clear differences in the light values between the objects. Manufactured objects tend to have a lower average light dispersion than produce. In addition, epicarp characteristics influence light reflection and absorption, with warmer epicarp colours having higher values compared to cooler epicarp colours. These differences have been statistically validated for their significance. Secondly, we evaluate the robustness of our approach to variations of distance and luminosity. The results indicate that there is no statistically significant effect on the light values for distances lower that one centimetre regardless the luminosity conditions. This suggests that the sensor should be in close proximity to the measured fruit while being able to tolerate small gaps (up to one centimetre). This aspect is

crucial in preventing manipulation of produce that may indirectly impact its quality. Furthermore, we analyse the potential of green light sensing to capture the quality of organic produce by evaluating the change in the fingerprint of the objects over a 15-day period and validating the findings with a 7-day follow-up experiment. Figure 4.3 (right part) shows that the magnitude of deviations correlates with the changes in the produce over time. Fruits experiencing significant alterations in epicarp color and firmness exhibit higher deviations. The differences in measurements between the two decomposition studies were not statistically significant, likely due to differences in the initial state of the produce.

In addition, we examine the correlation between the degree of variation and the thermal properties of produce used for modelling decomposition risks within different stages of the supply chain. These properties include: skin mass transfer coefficient (k_s), the transpiration coefficient (k_t) and the respiratory heat generation rate (W) [12, 38, 43]. The results indicate a strong relationship between variations in green light and produce decomposition. The final contribution of Publication VII demonstrates that green light measurements can support coarse-grained classification of produce into different stages of ripeness: ripe, overripe and decayed. The generic model (i.e., light reflectivity only) achieves a classification accuracy of $\approx 60\%$. Firmness of the produce is the most informative feature ($\approx 74\%$ accuracy combining firmness and light intensity). By incorporating multiple features the performance improves even further ($\approx 83\%$ accuracy combining firmness, light intensity, epicarp colour and produce name). If only light intensity values are available, then our approach is best suited for classifying produce or distinguishing between organic and non-organic items rather than on estimating quality. This demonstrates the general applicability and effectiveness of our approach in complementing various stages of the produce supply chain, which is crucial to reduce food loss and waste.

4.2 Air Pollution Monitoring

Reducing air pollution is one of the most important environmental sustainability challenges as poor air quality has significant health and economic impacts affecting millions of people worldwide [52]. Current solutions to mitigate air pollution rely on using professional-grade measurement stations [72] and low-cost sensors [18]. However, these sensors capture only the pollution distribution close to ground level without being able to provide information of the vertical distribution of pollutants or to explain their dispersion in the environment.

Publication VIII contributes a vision for city-scale air monitoring using commercial off-the-shelf unmanned autonomous vehicles (UAVs) equipped with portable air quality sensors. We demonstrate the benefits of the proposed approach through benchmark measurements from industrial and residential locations. We show that UAVs can provide the vertical profile of pollutant concentration and dispersion at different locations. The results highlight the importance of vertical modelling, showing the differences in pollutant distributions both vertically and between locations, compared to a background profile from a professional-grade measurement station. Capturing these differences is essential for accurate modelling and estimation of dispersion effects, providing new opportunities for atmospheric studies, understanding of pollution dispersion patterns and supporting sustainable development. To move from measurements collected by a remote operator to fully autonomous measurements, advances in the functional stages of the PDS architecture and sensing pipeline are needed to overcome the limitations and challenges of current technologies. Key requirements and challenges for autonomous monitoring are summarised in Tables 4.1 and 4.2.

The measurements are collected using a low-cost COTS UAV model X4S equipped with an Aeromon BH-12 sensor, which measures airborne gaseous compounds and particulate matter (PM) in the air. The PM sensor can detect concentrations ranging from $0.01\,\mu g/m^3$ to $1500\,mg/m^3$. Environmental sensors measured relative humidity (RH%), temperature (T), and pressure (P). The drone was manually controlled using the UAV's command to have a better control of the vertical sampling while maintaining line-of-sight with the UAV. Pollution measurements are collected at 11 different heights, with a 10-meter interval, starting from ground level to a height of 100 meters. Three particle sizes were measured: $PM_{1.0}$, $PM_{2.5}$, PM_{10}. These sizes refer to particles with diameters of at most 1.0 µm, 2.5 µm and 10.0 µm respectively. Sampling was carried out in an industrial and a residential area in southern Finland to evaluate changes in pollutant concentrations in locations with different characteristics. The gold standard measurements correspond to professional-grade measurement stations located close to the sampling areas. Cleaning and pre-processing included removing atypical values and synchronising the samples collected at the different locations with the gold standard measurements. The extracted features correspond to the representative statistics (i.e., mean, standard deviation and range). This values were used as input in a linear regression model that describes the changes in pollutant concentration at different altitudes and locations.

The results of the experiments of Publication VIII show the benefits of UAVs-based Air Pollution Monitoring. We demonstrate that our ap-

Table 4.1: Summary of requirements for environmental pollution monitoring using UAVs and AUVs. Adapted from Publications VIII and IX.

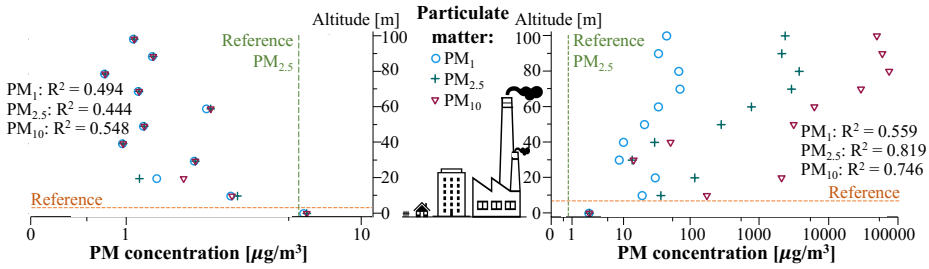
Requirement	Environment	Emerging Challenges
Pollution Detection, Identification and Localization	Air	Comprehensive air quality monitoring requires sensors to sample particulate, gaseous and other pollutants. 3D-maps help to understand the pollution situation, gain insights and coordinate UAVs to seal leaks and locate emissions. 5G and 6G technologies enable accurate location with minimal power consumption.
	Underwater	Pollutants come in different forms.and interactions between contaminants affect their formation. Detailed information is necessary for effective cleaning. Large-scale monitoring requires affordable AUVs with limited capabilities to coordinate surface-based cleaning.
Coordination	Air	Coordination enables horizontal and vertical sampling strategies. Sampling locations for UAVs should complement ground-based solutions, citizen sensors, and monitoring stations. Real-time coordination is needed for efficient sampling plans and maximise data quality.
	Underwater	Coordination allows covering large areas. Collaborative AUV deployments enable to capture the full range of pollutants at different depths. Coordination is essential for communication range and collaborative sensing.
Lightweight Sensor and Hardware Designs	Air	UAVs need lightweight sensors for efficient monitoring in large areas. Sensor placement must be optimised to prevent air flows from affecting pollution measurements. New computing units should enable real-time monitoring, i.e., processing, data analysis and coordination.
	Underwater	Interfaces for remote operators are required for specifying bounds and constrains to AUVs and coordinate sampling and monitoring areas. AUVs need to be fault-tolerant to operate for long periods without human intervention and require reliable casing solutions.

proach provides comprehensive information on the vertical distributions of particulate matter for the two sampling locations (refer to Figure 4.4). As expected, higher concentrations are observed at ground level in the residential area. The vertical column at $1.0\,\mu g/m^3$ in the figure represents negligible concentrations. In contrast, the industrial location displays considerably higher concentrations compared to the residential area. While the concentrations at ground level are comparable, the plume resulting from industrial processes substantially elevates particle concentrations at higher

Table 4.2: Research challenges and emerging topics for pollution monitoring in different environments (E) using aerial (A) and underwater (U) autonomous vehicles. Adapted from Publications VIII and IX.

Type	E	Key Research Challenges	Emerging Challenges
Sensing	A	Improve the accuracy of sensors using calibration	Opportunistic calibration integrating precise sampling and compact sensors
	U	Methods for categorizing pollution sources and identifying multiple pollutants. Robustness under diverse conditions.	Model the inner change of materials exposed to the environment.
Power and Operational Time	A	Innovative approaches for energy management and optimising sampling	Collaborative processing and computing augmentation
	U	Novel energy management and power solutions, i.e., wireless charging stations, underwater energy harvesting	Resource augmentation and energy efficiency using collaborative processing and offloading
Networking	A	Ultra low-latency data transfer, and precise localization	Integrate UAVs with edge and 5G infrastructure
	U	High band short-range technologies and robust long-range technologies capable of interacting with existing infrastructure	Robustness of technologies against water characteristics, e.g., water flows, temperature, and salinity.
Localization	A	Accurate and robust 3D localization	Localization using emerging technologies, e.g., mmWave
	U	Positioning schemes for 3D absolute localization, robust relative positioning.	Hybrid localization that offers relative and absolute positioning
Situational awareness	A	Models to analyze air pollutants in diverse contexts	Advanced models for collecting air quality measurements
	U	High-resolution and energy-efficient situational awareness techniques	Coordinated data collection that supports scientific sampling

altitudes. The distribution of particles varies with size and altitude. Larger particles have the highest concentration at higher altitudes due to hygroscopic growth. As altitude increases, pollutants either fall to the ground or disperse in the environment. The differences in vertical distribution are significant and depend on the sampling location. This demonstrates the importance of employing a UAV-based solution for capturing pollutants at different altitudes to monitor the vertical column, and understand the influencing factors and the mechanisms that control dispersion in different environments, as well as to support health risks assessment.



(a) Residential (b) Industrial

Figure 4.4: Concentration of particulate matters (PMs) at different air column altitude within different environmental profiles. Previously published in Publication VIII [74].

The power of UAVs in this application is on capturing a three-dimensional view of the distribution of pollutants. We showed the significant effect of location and altitude on PM concentrations, and the no interaction effect of variations in pollution distribution between the two locations and different altitudes. Statistically significant differences occur from 70 meters onward, highlighting the importance of monitoring both the horizontal and vertical dimensions in order to accurately measure the full extent of pollutants as emissions can vary significantly at different altitude.

When comparing $PM_{2.5}$ measurements collected using the UAV with the gold standard, differences between the reference stations primarily result from variations in the deployment location. In Figure 4.4, the reference lines for the residential location demonstrate a close agreement with the experimental values. Conversely, measurements taken at the industrial site show higher concentrations compared to those at the reference station. Concentration levels correlate with the degree of pollutants. The divergence from the reference values suggests that the elevated pollutant concentrations are localised, with the pollutants dispersing across a wider area.

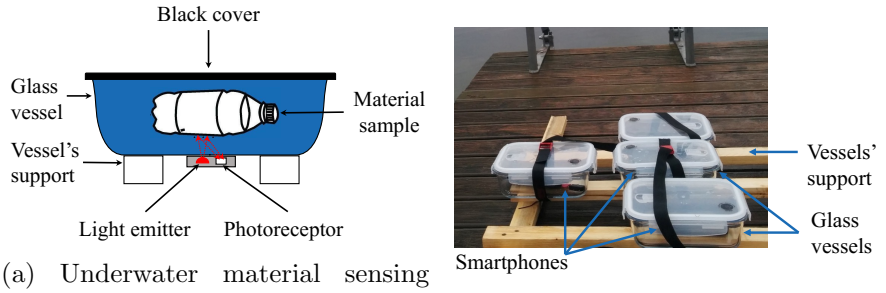
4.3 Marine Pollution Monitoring

Marine pollution is a growing global challenge that has adverse effects on the well-being of marine ecosystems, weather patterns, and even human health [28, 99]. To effectively mitigate underwater pollution, it is crucial to implement appropriate measures and clean-up operations, which in

turn require reliable information on the extend of marine pollutants [28]. Obtaining information about the extent of marine pollutants is currently challenging due to existing measurement solutions are laborious costly, restricted to specific pollutants, and offer limited spatio-temporal resolution. Similarly, sensing solutions are poorly suited for underwater sensing due to being sensitive to environmental characteristics [48, 107].

Publication IX presents a research vision for large-scale autonomous marine pollution monitoring using coordinated groups of autonomous underwater vehicles (AUVs) to assess the extent and characteristics of pollutants in aquatic environments and green light sensing to classify marine debris. The feasibility of our vision is addressed by evaluating two key research challenges through small-scale controlled experiments: (i) pollutant detection and classification and (ii) distributed and cooperative processing. Controlled experiments and evaluation focus on preparing the PDS systems to operate in marine ecosystem to gather detailed information on pollutants at high spatial and temporal resolution. Similar as in Section 4.2, the Publication presents the key requirements challenges, and enablers for enabling our approach (see Tables 4.1 and 4.2).

Our experimental setup for debris classification relies on the same COTS device used in Section 4.1 (green light LED combined with a photoreceptor) We use the device to collect light intensity measurements reflected by different materials (paperboard, high-density polyethylene, polyethylene terephthalate, aluminium, feldspar and solid walnut oak) under different sensing medium (air cf. water) and luminosity conditions (ambient cf. darkness). The objects are placed individually into a glass container covered with a non-reflective (black) lid. The smartwatch is taped outside the container, directly below the measured object (see Figure 4.5a). Sensing medium emulates a water-proof casing on the surface (empty container) and in underwater environment (container filled with water). Luminosity conditions emulate the case where the container is unobstructed and ambient light penetrates into the water (ambient) and when the sensor is in direct contact with the object (darkness). The light intensity measurements are sampled at a 100 Hz frequency over a 90-second period, and annotated for each material and experimental condition. Cleaning and pre-processing are performed using the same methods as in Section 4.1, i.e., extracting the more stable part of the signal and applying noise filters that preserve the temporal characteristics of the measurements. The extracted features correspond to the average reflected light of the pre-processed signal, which serve as input for classification models [35, 118]. We focus on simple classifiers to ensure that they are as energy efficient as possible to operate in AUVs.



(a) Underwater material sensing setup.

(b) Micro-cloud deployment.

Figure 4.5: Controlled testbed for underwater material sensing and collaborative processing. Adapted from Publication IX [36].

The experimental setup to evaluate collaborative processing underwater consists of four LG Nexus 5 smartphones placed inside sealed glass containers mounted on a wooden structure equidistant from each other (see Figure 4.5b). The arrangement forms a micro-cloud in which devices communicate via a WiFi interface. We implement an Android prototype for collaborative object recognition that runs on the devices using a master-slave approach. The master device randomly assigns computing tasks to other devices (workers). The master receives a sequential video feed and sends frames to the workers in a round-robin fashion. Workers send results back to the master upon task completion. The sensing pipeline involves an experimental object recognition task using a video feed. It simulates the computational requirements for recognising debris underwater. The video feed consists of 50 images (224×224 resolution) from ImageNet. The task is performed on the surface and underwater with the devices in the glass containers. The gold standard corresponds to the task performed on the surface without encase. Each device in the micro-cloud has installed a pre-trained and quantized convolutional neural network model (MobileNet) for object recognition. As the evaluation metrics we use task completion and success rate.

The results obtained from Publication IX for underwater material sensing confirm that light reflectivity measurements of different materials have significant variation to support debris identification in all tested conditions. Publication IX also demonstrates that optical sensing can provide a coarse-grained classification of debris by using light intensity measurements to build simple classifiers that can be deployed on AUVs at low energy cost. The performance analysis shows that changes in the sensing medium have minimal effect compared to changes in luminosity. In fact, the best classification performance, just over 80%, is achieved when measurements come

from similar luminosity environments. This suggest that classification of marine pollutants using optical sensing is feasible as long as the models are trained with sufficiently similar samples to the AUV operating environment.

For the underwater distributed processing, Publication IX confirms that processing time of the micro-cloud is affected by the number of participant devices. Increasing the number of workers reduces processing time, which is important for achieving higher video rates and image resolutions. Underwater micro-clouds have the potential to support tasks that require real-time analysis of camera images. Experimental results demonstrate that even with just one worker device, a processing frequency of 30 fps can be achieved. When assessing the effect of encasing or submersion on computing performance, we observe a marginal increase in the response time, i.e., $\approx$ 5ms for the encasing and $\approx$ 20 for underwater. The task success rate, percentage of frames that the workers return to the master, is 100% when the devices are above the surface. As expected, success rate decreases when the micro-cloud is underwater. When assessing the effect of distance between devices, we observed that success rate drastically drops to 0% beyond 10 cm. The result suggests the need to improve the networking layer with efficient short-range communication solutions for underwater operations.

4.4 Deep Learning in Marine Environments

Pervasive Data Science systems provide valuable possibilities for real-time exploration, monitoring, and modelling of environments. In the domain of marine science and oceanography, machine learning (ML) techniques are increasingly employed to analyse underwater video footage [46, 53], playing a central role in various emerging underwater applications aimed at sustainable development [69]. The utilization of vision-based data holds significant potential for leveraging deep learning (DL) in supporting deep sea applications. Nevertheless, this also imposes limitations on the platforms that handle such data and poses challenges in ensuring accurate operation of the techniques [112] and operating ML algorithms in-situ [53].

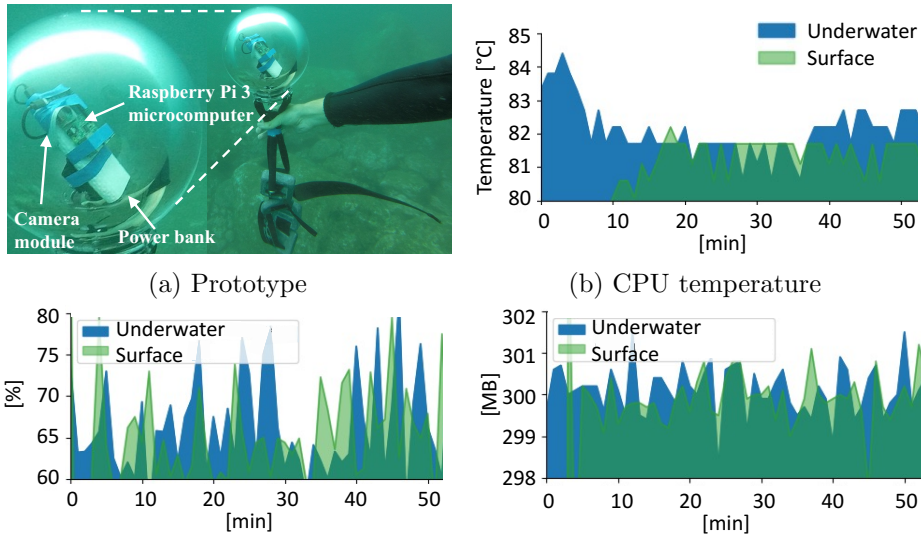
Publication X demonstrates the potential of DL to enrich existing tasks in aquatic environments. We envision integrating DL directly as part of underwater operations to offer timely access to data and insights about the underwater environment. We use underwater marine litter detection as case study, which allows to show the impact of using embedded DL to increase the scale of marine pollution monitoring by supporting automated in-situ analysis. PDS DL-based systems can bring significant benefits to underwater computing, supporting automated analysis of data in underwater monitoring

and providing a mechanism to support the operation of underwater vehicles or other infrastructure integrating computing capabilities, such as seabed sensor networks and buoys.

The potential of underwater DL extends to several domains that require real-time analysis of underwater data, such as audio or video signals, collected in different ways, e.g. by underwater vehicles (AUVs or ROVs), divers or marine fauna. Embedding new technologies into underwater vehicles is becoming increasing possible, but current attempts mainly rely on using hardware build to work in the surface. Deploying computing resources in marine environments is highly challenging due to the variation in the operation conditions, e.g., increment in depth derives on higher pressure and water density, requiring robust casings and balancing the inside-outside pressure. Summarising the current state of submersible hardware allows to understand current limitations of existing deployments and the opportunities of using underwater vehicles prepared to efficiently run DL models.

Current deep-sea monitoring rely on static cable connected underwater stations or observatories, which are fixed and communicate with central ground control stations through cables that provide real-time communication and power supply. Relying on a mobile platform for communication is highly costly and offer limited computing capability. AUVs and ROVs are currently popular for underwater explorations, however their processing and battery autonomy is highly limited. More advanced platforms that integrate computing features for running DL are extremely costly to be consider for large-scale low-cost COTS solutions. Hybrid Technologies is a solution to powerful processing power through the cloud, but this works best close to shore and calm areas close to a surface-based hub or gateway. Combining underwater operations with surface-based cloud is also prone to transmission errors and limits real-time operation to the maximum depth at which devices can operate.

Integrating deep learning directly into underwater platforms has the potential to significantly enhance underwater investigations by providing real-time access to data and insights of the aquatic environment. Publication X contributes to this by demonstrating the potential of DL to enrich pollution detection in underwater environments. The field experiments for underwater marine litter detection enclose a Raspberry Pi 3 microcomputer with camera module and power bank placed in a sealed container (see Figure 4.6a) and transported during a 50 minute dive. The DL model was implemented using TensorFlow Lite, which is the standard framework for multi-platform deployment of DL models and big data applications. A GPU infrastructure was utilised for training and obtaining the quantized model for object



(a) Prototype

(b) CPU temperature

(c) CPU usage

(d) RAM usage

Figure 4.6: In-the-wild test for litter classification with (a) Deployed microsphere for real-time marine litter detection, (b) CPU temperature (Celsius), (c) CPU usage (percentage) and (d) and RAM usage (MB). Previously published in Publication X [84].

recognition. For the experiments we use Trash-ICRA19 [40] annotated trash dataset. We extended annotations to include four different litter categories: plastic, wood, metal, and other. We collected the ground truth by repeating the same experimental setup on the surface. The device is then set up to perform continuous object recognition at 15 FPS captured by the camera. The modelling is based on the MobileNetv2 [89] DL architecture with Single-Shot Detection (SSD), which has lightweight structure and good object recognition performance. We used intersection of unions (IoU) bounding boxes, CPU temperature and usage and RAM usage as evaluation metrics, as they allow us to assess the performance of the model and the effect of the environment on hardware performance.

Publication X provides the insights from the analysis of litter classification and hardware performance during in-the-wild tests. Classification results show a median classification accuracy of 87% for all debris categories. The materials with the highest accuracy are wood (97%) and metal (95%). Performance and recall drop to 56% and 60% respectively for images containing multiple litter objects. This demonstrates the potential of using DL to detect (and classify) marine litter in real-time and to automate

underwater tasks. Analysis of the detection successes and failures highlights the challenges of extending these systems to operate robustly in the underwater environment. For example, individual categories of marine debris are generally easy to detect at close range. At longer distances, light and water conditions reduce overall accuracy. Complex backgrounds, or where debris are covered by sediment, add to the detection complexity. As mentioned above, the model has problems detecting multiple objects or object categories. In the evaluation of hardware performance, CPU temperature shows an early spike during the underwater operation (see Figure 4.6b), which responds to the lack of air outlet inside the container. The colder water starts to offer cooling for the container when dive submerges deeper. CPU and RAM usage results show a similar performance underwater and in the surface (see Figures 4.6c and 4.6d). These results suggest that even low-cost micro-controllers could be integrated with readily available AUVs and ROVs to develop cost-effective underwater computing platforms that support the execution of deep learning models. However, deploying these systems at large requires further work, e.g., novel casing designs as well as cooling and energy management features to avoid interfering the underwater vehicles operation. The results of the experiments demonstrated the potential of deploying underwater vehicles integrated with DL-models to automate different tasks. However, it is necessary to overcome important challenges for helping to scale up DL operation for underwater monitoring, such as operating in extreme and changing conditions and differing water temperatures. The key research challenges are summarised in Table 4.3.

Table 4.3: Research challenges and emerging topics for enabling deep learning deployment. Adapted from Publication X.

Type	Key Research Challenges	Emerging Challenges
Sensing	Migration of existing sensors for underwater operations	New portable, lightweight and energy-efficient sensing solutions
Processing resources	Augmenting resources with additional infrastructure	Advanced augmentation with distributed and collaborative processing
Fault tolerance and operational time	Recovery and replacement of components without extraction from underwater	Multi-modal techniques to provide robust recovery and continuous operations
Communication and cooperation	Adoption of different communication technologies, e.g., electromagnetic, acoustic and optical	Emerging mature interfaces and integration with new paradigms, e.g., 5G and 6G
Resource intensive processing	Design of better encasing to improve thermal absorption of heavy processing	Emerging approaches to reduce thermal overhead based on distributed processing
Advance autonomy	New autonomous functionalities that reduce human intervention, e.g., back to home routines	Total autonomy for underwater solutions, e.g., self-healing, optimisation and configuration
Open SDKs and extendable APIs	Open firmware to build a wider ecosystems of solutions	Adoption of a common and reusable platform
Data diversity and massive datasets	Adoption of static monitoring solutions in different aquatic environments	Emerging integration of dynamic monitoring and data collection with underwater vehicles

Chapter 5

Discussion and Conclusion

In this thesis, we have study Pervasive Data Science as an emerging paradigm that can generate new knowledge across domains from data collected by multi-device platforms. We have addressed important challenges and limitations in PDS to improve its adoption as an independent research domain. Specifically, we have provided insights into the processes that govern the way data is generated, offered methods to improve and understand data collection from multiple devices, demonstrated how new applications can be enabled by re-purposing existing sensors, and provided examples of end-user applications in the field of sustainable development. In the following, we present some insights into the limitations and future potential of the contributions, as well as conclude the thesis.

5.1 Discussion and Future Work

Data Fusion and analysis of performance: Publication I provided insights into the processes that govern how crowdsensing produces measurements. The amount and quality of measurements is essential for providing the best insights, and it is also essential for enabling accurate AI models. Normally one app only provides one type of measurements and optimally measurements from multiple apps need to be combined. Combining datasets, however, can be challenging as the intersection of the datasets is often very small. If the data is limited to the part that is common, this reduces the benefits gained from crowdsensing and can lead to unbalanced and biased samples. Methods for data fusion specifically designed for large-scale datasets, which evaluate the balance between data size and data quality, are necessary. Beyond data fusion, it is also necessary to better understand factors that affect data production processes. Publication I examined the

critical factors of energy and latency in relation to user perception and demonstrated how they affect data availability through retention. Publication III, in turn, studied how spatio-temporal context and device availability affect data availability and Publication II demonstrated how mobile system design can be harnessed to improve data quality. It is essential to investigate also other factors that are linked to performance, such as diverse usage patterns, mobile interactions [76] and human mobility [100], and to develop further methods that help to improve the quality and availability of data.

Applications and Stakeholders: We have demonstrated several new applications for pervasive data science, yet naturally there are many other ones that our work enables. Long and stable collaborations, as enabled by Publication II, enables collaboration scenarios involving resource-intensive sensors and hardware, asynchronous machine learning, edge intelligence, as well as small-scale data centers powered by smartphones. In Publication VI, we contributed with an innovative method for detecting covert surveillance devices using thermal sensing. This could be adopted in markets such as Airbnb and Uber where there is often a lack of trust and a high risk of being surreptitiously monitored. The results of Publications VII, VIII, and IX similarly can be harnessed for many purposes and adopted by stakeholders ranging from industry to local authorities and end-users.

Other Contexts: As is common, our experiments largely considered settings that offer sufficient degree of control. Increasing the sample sizes and conducting further data collections can further improve the insights that can be gathered. For example, educational campaigns on food loss could be used to collect further data to support the methods in Publication VII. Publication II examines stable collaborations in an individual location, but it would be possible to extend the work also in scenarios where people are mobile as mobility patterns have been shown to have a high degree of regularity [29]. Further research should also explore stability in different types of locations and time spans, considering different contexts and factors such as cultural differences, mobility variations, urban design and size of geographic area. Similarly, Publication VI studied surveillance cameras that can be easily purchased in any market without restrictions and the evaluation could be expanded to devices that are specifically designed for covert surveillance, such as cameras embedded in pens or USB sticks, and to additional test environments, such as home-type environments. Nevertheless, our work has offered new ways to harness pervasive data science to deliver innovative end-user applications.

Accuracy and privacy of low-cost sensors: Low-cost sensors are designed to be easy to deploy and consume low power, making them accessible

to a wide range of applications. In Section 3.2 we evaluate the performance trade-offs associated with low-cost sensors and examine how accuracy is affected by sensor configuration. However, it is important to note that the accuracy of low-cost sensors can be susceptible to drift bias, which limits their effectiveness in different contexts and environments. It is crucial to conduct studies aimed at reducing environmental drift and minimising context dependency. By addressing these issues, it may be possible to ensure the reliable and robust operation of sensors across different contexts and environments, which is important to enable our vision of multi-ecosystem pollution monitoring presented in Sections 4.2 and 4.3. Low-cost sensors have the potential to collect valuable data that can complement existing applications and address new challenges. However, this feature also raises concerns about the unauthorised collection of data from users or the possibility of combining this data with existing data and revealing sensitive information. Further research on context-aware privacy of sensor data and data aggregation, such as differential privacy or federated learning, should be conducted to ensure privacy in multi-sensor environments.

Large-Scale Adoption: The applications presented in this thesis were intended as proof-of-concept demonstrators that showcase the potential for PDS. Before these can be brought to large-scale use, there are further challenges that need to be addressed. For example, Publication IX demonstrated the potential of optical sensing to detect and identify underwater pollution, but further work is needed to improve the fabrication of materials that are suitable for underwater use and for system designs that can avoid disturbing and damaging underwater ecosystems. Similarly, the use of UAVs at city scale, as envisioned in Publication VIII, requires the incorporation of mechanisms to ensure reliable operation, such as coordination and collision avoidance, and regulations on the operation of UAVs that prevent carrying excessive equipment or to invade personal privacy.

Complementary Solutions: In many cases, the best results are obtained by combining multiple different solutions, each of which has their own benefits and disadvantages. For example, Publication VII studied the use of low-cost optical sensing for estimating produce quality, but it could be integrated with other methods, such as computer vision or even multispectral imaging, to offer more accurate information about produce quality. Similarly, the smart plant solution described in Publication IV, has been designed to be compatible with infrastructure-based sensors, such as thermal array sensors used in Publication V.

Advancing sustainable development: In Chapter 4 we presented food waste reduction and pollution monitoring as representative examples of

how sustainable development applications can benefit from PDS. We envision having low-cost sensors and sensing methods that can be seamlessly integrated into all stages of a process, ensuring traceability and providing information on the operational context. For example, integrating our quality assessment solution (see Section 4.1) throughout the food supply chain to assess the efficiency of the storage and transport system. Beyond the limitations on data collection discussed above, we have identified certain limitations in multi-device platforms that need to be overcome to enable effective data collection within and between environments. For example, pollution monitoring requires the use of robust communication technologies and coordination algorithms to enable covering larger areas and ensure efficient sampling. Communication and coordination should be extended to allow interaction between UAVs and AUVs, enabling multi-ecosystem pollution monitoring. These issues warrant further research and development.

Enabling AI underwater: Section 4.4 presents an evaluation of deep learning (DL) models in aquatic environments, focusing on image classification tasks due to their relevance in this context. However, it is important to note that other models, such as Large Language Models (LLMs) or Long Short Term Memory (LSTM), can also be explored to gain complementary insights into the energy and performance costs associated with different types of underwater tasks. In our analysis, we conducted experiments using a diver to transport the computing unit underwater. However, future research can extend this work by using Autonomous Underwater Vehicles (AUVs) to assess the readiness and robustness of hardware and software to host DL models, including assessing interference from other operations, computational overload and variations in energy consumption. By considering these advances, we can gain a deeper understanding of the potential applications of DL in underwater environments and improve the overall efficiency and effectiveness of these models in real-world scenarios.

5.2 Summary and Conclusion

Pervasive Data Science (PDS) is an emerging area that integrates the Internet of Things (IoT), Pervasive Computing, and Data Science to address everyday needs in multi-device environments. While the individual subfields of PDS have been highly active in recent years, progress and widespread adoption of PDS itself has been hampered by challenges that affect the collection, analysis and use of sensor data produced by smart devices. In this thesis, we have contributed insights, methods, new applications, and system designs for fuelling the uptake of PDS and to produce new applications.

Firstly, the quality of sensor data has been improved by illustrating the power of data fusion in understanding the relationships between different factors collected by individual datasets, and the limitations of current data fusion methods in handling large-scale datasets. The impact of network latency and energy consumption on mobile application retention was quantified using two large-scale crowdsensed datasets. Opportunistic collaborative sensing and computing in multi-device environments was analysed, showing the correlation between user mobility and collaboration opportunities, and introducing a novel collaborator selection method based on Markov trajectory entropy. Secondly, the potential of re-purposing low-cost sensors has been studied by considering them as affordable and easy-to-deploy solutions for data collection in various applications. In this thesis we studies the use of smart plants integrated with sensors for indoor environmental monitoring, the use of thermal array sensors for occupancy sensing, and the use of thermal imaging to detect covert surveillance devices to protect the privacy of indoor users. Finally, this thesis has demonstrated the potential of Pervasive Data Science to support environmental sustainability and sustainable development. We introduced PDS solutions for product quality assessment along the supply chain, air pollution monitoring using aerial drones, and large-scale marine pollution monitoring using underwater vehicles. The performance and challenges of using deep learning models in new environments, i.e. aquatic environments for image classification, have also been illustrated.

Taken together, the contributions of this thesis help pave the way for further adoption of PDS. Naturally, there are many other challenges that also need addressing, such as integration of suitable AI models, privacy, and security. Nevertheless, this thesis can serve as catalyst for further research in PDS and helps to demonstrate the challenges and benefits of PDS, while also providing some solutions to some of its key challenges.

References

[1] Tommaso Addabbo, Ada Fort, Marco Mugnaini, Lorenzo Parri, Alessandro Pozzebon, and Valerio Vignoli. Smart Sensing in Mobility: a LoRaWAN Architecture for Pervasive Environmental Monitoring. In *Proceedings of the 2019 IEEE International Forum on Research and Technology for Society and Industry*, pages 421–426. IEEE, 2019.

[2] Ardalan Amiri Sani, Kevin Boos, Min Hong Yun, and Lin Zhong. Rio: A System Solution for Sharing I/O Between Mobile Systems. In *Proceedings of the International Conference on Mobile Systems, Applications, and Services*, pages 259–272. ACM, 2014.

[3] Byeong Wan An, Jung Hwal Shin, So-Yun Kim, Joohee Kim, Sangyoon Ji, Jihun Park, Youngjin Lee, Jiuk Jang, Young-Geun Park, Eunjin Cho, Subin Jo, and Jang-Ung Park. Smart Sensor Systems for Wearable Electronic Devices. *Polymers*, 9(8):303, 2017.

[4] Kumaribaba Athukorala, Eemil Lagerspetz, Maria von Kügelgen, Antti Jylhä, Adam J. Oliner, Giulio Jacucci, and Sasu Tarkoma. How Carat Affects User Behavior: Implications for Mobile Battery Awareness Applications. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 1029–1038. ACM, 2014.

[5] Ricardo A. Baeza-Yates, Di Jiang, Fabrizio Silvestri, and Beverly Harrison. Predicting The Next App That You Are Going To Use. In *Proceedings of the International Conference on Web Search and Data Mining)*, pages 285–294. ACM, 2015.

[6] Dimitri Belli, Stefano Chessa, Luca Foschini, and Michele Girolami. A Probabilistic Model for the Deployment of Human-Enabled Edge Computing in Massive Sensing Scenarios. *IEEE Internet of Things Journal*, 7(3):2421–2431, 2020.

[7] Alex Beltran, Varick L Erickson, and Alberto E Cerpa. Thermosense: Occupancy Thermal Based Sensing for HVAC Control. In *Proceedings*

of the ACM Workshop on Embedded Systems For Energy-Efficient Buildings. Association for Computing Machinery, 2013.

[8] Matthias Böhmer, Brent Hecht, Johannes Schöning, Antonio Krüger, and Gernot Bauer. Falling Asleep with Angry Birds, Facebook and Kindle: A Large Scale Study on Mobile Application Usage. In *Proceedings of the ACM International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 47–56. ACM, 2011.

[9] Team Braze. 10 Essential Mobile App KPIs and Engagement Metrics (and How to Use Them). May 2016. https://www.braze.com/resources/articles/essential-mobile-app-metrics-formulas. [Accessed on 2023-06-01].

[10] Federico Brilli, Silvano Fares, Andrea Ghirardo, Pieter de Visser, Vicent Calatayud, Amalia Muñoz, Isabella Annesi-Maesano, Federico Sebastiani, Alessandro Alivernini, Vincenzo Varriale, and Flavio Menghini. Plants for Sustainable Improvement of Indoor Air Quality. *Trends in Plant Science*, 23(6):507–512, 2018.

[11] Robert Cassen. Our Common Future: Report of the World Commission on Environment and Development. *International Affairs*, 64(1): 126, 1987.

[12] Khe Van Chau, Robert Romero, Direlle Baird, and Jerome Gaffney. Transpiration Coefficients of Fruits and Vegetables in Refrigerated Storage. *ASHRAE Report 370-RP*, 1987.

[13] Andrew Chen. Number of Apps Available in Leading App Stores as of 3rd Quarter 2018. https://andrewchen.co/new-data-shows-why-losing-80-of-your-mobile-users-is-normal-and-that-the-best-apps-do-much-better. [Accessed on 2019-02-14].

[14] Ning Chen, Jialiu Lin, Steven CH Hoi, Xiaokui Xiao, and Boshen Zhang. AR-miner: Mining Informative Reviews for Developers from Mobile App Marketplace. In *Proceedings of the ACM International Conference on Software Engineering*, pages 767–778. ACM, 2014.

[15] Veena Chidurala and Xinrong Li. Occupancy Estimation Using Thermal Imaging Sensors and Machine Learning Algorithms. *IEEE Sensors Journal*, 21(6):8627–8638, 2021.

[16] Niel Andre Cloete, Reza Malekian, and Lakshmi Nair. Design of Smart Sensors for Real-Time Water Quality Monitoring. *IEEE Access*, 4:3975–3990, 2016.

[17] Mertcan Cokbas, Prakash Ishwar, and Janusz Konrad. Low-Resolution Overhead Thermal Tripwire For Occupancy Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 88–89. CVF/IEEE, 2020.

[18] Francesco Concas, Julien Mineraud, Eemil Lagerspetz, Samu Varjonen, Xiaoli Liu, Kai Puolamäki, Petteri Nurmi, and Sasu Tarkoma. Low-Cost Outdoor Air Quality Monitoring And Sensor Calibration: A Survey And Critical Analysis. *ACM Transactions on Sensor Networks (TOSN)*, 17(2):1–44, 2021.

[19] Farooq Dar, Hilary Emenike, Zhigang Yin, Mohan Liyanage, Rajesh Sharma, Agustin Zuniga, Mohammad A Hoque, Marko Radeta, Petteri Nurmi, and Huber Flores. The Midas Touch: Thermal Dissipation Resulting From Everyday Interactions As A Sensing Modality. *Pervasive and Mobile Computing*, 84:1–18, 2022.

[20] Nigel Davies and Sarah Clinch. Pervasive Data Science. *IEEE Pervasive Computing*, 16(3):50–58, 2017.

[21] Nigel Davies, Nicholas D. Lane, and Mirco Musolesi. Pervasive Data Science and AI. *IEEE Pervasive Computing*, 18(3):7–8, 2019.

[22] Mark de Reuver, Harry Bouwman, Nico Heerschap, and Hannu Verkasalo. Smartphone Measurement: Do People Use Mobile Applications as They Say They Do? In *Proceedings of the 11th AIS International Conference on Mobile Business (ICMB 2012)*, pages 1–10. AISeL, 2012.

[23] Pralhad Deshpande, Xiaoxiao Hou, and Samir R Das. Performance Comparison of 3G and Metro-Scale WiFi for Vehicular Network Access. In *Proceedings of the 10th SIGCOMM Conference on Internet measurement (IMC 2010)*, pages 301–307. ACM, 2010.

[24] Srinivas Devarakonda, Parveen Sevusu, Hongzhang Liu, Ruilin Liu, Liviu Iftode, and Badri Nath. Real-Time Air Quality Monitoring Through Mobile Sensing In Metropolitan Areas. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, pages 1–8, 2013.

[25] Jessilyn Dunn, Ryan Runge, and Michael Snyder. Wearables and the Medical Revolution. *Personalized medicine*, 15(5):429–448, 2018.

[26] Laura Ekroot and Thomas M Cover. The Entropy of Markov Trajectories. *IEEE Transactions on Information Theory*, 39(4):1418–1421, 1993.

[27] Hilary Emenike, Farooq Dar, Mohan Liyanage, Rajesh Sharma, Agustin Zuniga, Mohammad A Hoque, Marko Radeta, Petteri Nurmi, and Huber Flores. Characterizing Everyday Objects Using Human Touch: Thermal Dissipation As A Sensing Modality. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–8. IEEE, 2021.

[28] Marcus Eriksen, Laurent CM Lebreton, Henry S Carson, Martin Thiel, Charles J Moore, Jose C Borerro, Francois Galgani, Peter G Ryan, and Julia Reisser. Plastic Pollution In The World's Oceans: More Than 5 Trillion Plastic Pieces Weighing Over 250,000 Tons Afloat At Sea. *PloS One*, 9(12):e111913, 2014.

[29] Farbod Faghihi and Petteri Nurmi. An Empirical Study On The Regularity Of Route Mobility. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1418–1425, 2016.

[30] Gisela Fernandes. *Pervasive Data Science Applied to the Services Society*. PhD thesis, University of Minho Guimarães, Portugal, 2019.

[31] Huber Flores, Sasu Tarkoma, Petteri Nurmi, and Pan Hui. MobileCloudSim: A Context-Aware Simulation Toolkit for Mobile Computational Offloading. In *Proceedings of the ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp 2018)*, pages 38–41. ACM, 2018.

[32] Huber Flores, Jonatan Hamberg, Xin Li, Titti Malmivirta, Agustin Zuniga, Eemil Lagerspetz, and Petteri Nurmi. Evaluating Energy-Efficiency Using Thermal Imaging. In *Proceedings of ACM HotMobile*, pages 147–152, 2019.

[33] Huber Flores, Jonatan Hamberg, Xin Li, Titti Malmivirta, Agustin Zuniga, Eemil Lagerspetz, and Petteri Nurmi. Estimating Energy Footprint Using Thermal Imaging. *ACM GetMobile*, 23(3):5–8, 2020.

[34] Huber Flores, Agustin Zuniga, Farbod Faghihi, Xin Li, Samuli Hemminki, Sasu Tarkoma, Pan Hui, and Petteri Nurmi. Cosine: Collaborator selector for cooperative multi-device sensing and computing. In *Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–10. IEEE, 2020.

[35] Huber Flores, Agustin Zuniga, Naser Hossein Motlagh, Mohan Liyanage, Monica Passananti, Sasu Tarkoma, Moustafa Youssef, and Petteri Nurmi. Penguin: Aquatic Plastic Pollution Sensing Using AUVs. In *Proceedings of the ACM Workshop on Micro Aerial Vehicle Networks, Systems, and Applications (DroNet@MobiSys)*, pages 1–6. ACM, 2020.

[36] Huber Flores, Naser Hossein Motlagh, Agustin Zuniga, Mohan Liyanage, Monica Passananti, Sasu Tarkoma, Moustafa Youssef, and Petteri Nurmi. Toward Large-Scale Autonomous Marine Pollution Monitoring. *IEEE Internet of Things Magazine*, 4(1):40–45, 2021.

[37] Huber Flores, Agustin Zuniga, Leonardo Tonetto, Tristan Braud, Pan Hui, Yong Li, Sasu Tarkoma, Mostafa Ammar, and Petteri Nurmi. Collaboration Stability: Quantifying the Success and Failure of Opportunistic Collaboration. *Computer*, 55(8):70–81, 2021.

[38] Susana C Fonseca, Fernanda AR Oliveira, and Jeffrey K Brecht. Modelling Respiration Rate of Fresh Fruits and Vegetables for Modified Atmosphere Packages: A Review. *Journal of Food Engineering*, 52(2): 99–119, 2002.

[39] Bin Fu, Jialiu Lin, Lei Li, Christos Faloutsos, Jason Hong, and Norman Sadeh. Why People Hate Your App: Making Sense of User Feedback in a Mobile App Store. In *Proceedings of the 19th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD 2013)*, pages 1276–1284. ACM, 2013.

[40] Michael S Fulton, Jungseok Hong, and Junaed Sattar. Trash-ICRA19: A Bounding Box Labeled Dataset of Underwater Trash. *Interactive Robotics and Vision Lab*, 2020.

[41] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding Individual Human Mobility Patterns. *Nature*, 453 (7196):779–782, 2008.

[42] Afton Halloran, Jesper Clement, Niels Kornum, Camelia Bucatariu, and Jakob Magid. Addressing Food Waste Reduction in Denmark. *Food Policy*, 49:294–301, 2014.

[43] ASHRAE Handbook and Refrigeration Chapter. Thermal Properties of Food. *ASHRAE Refrigeration Handbook, ASHRAE, Atlanta, GA*, 2018.

[44] Mohammad Hasan. State of IoT 2022: Number of Connected IoT Devices Growing 18% to 14.4 Billion Globally. In *IoT Analytics Research 2022*, 2022. https://iot-analytics.com/number-connected-iot-devices. [Accessed on 2022-11-01].

[45] Samuli Hemminki, Kai Zhao, Aaron Yi Ding, Martti Rannanjärvi, Sasu Tarkoma, and Petteri Nurmi. CoSense: A Collaborative Sensing Platform for Mobile Devices. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, pages 34–35. ACM, 2013.

[46] Michael Ho, Sami El-Borgi, Devendra Patil, and Gangbing Song. Inspection and monitoring systems subsea pipelines: A review paper. *Structural Health Monitoring*, 19(2):606–645, 2020.

[47] Dave Hoch. App User Retention Improves in the U.S., but Declines Internationally. 2017. http://info.localytics.com/blog/app-user-retention-improves-in-the-us. [Accessed on 2019-2-14].

[48] Florian Huth, Martin Schnell, Jesper Wittborn, Nenad Ocelic, and Rainer Hillenbrand. Infrared-spectroscopic Nanoimaging With a Thermal Source. *Nature Materials*, 10(5):352–356, 2011.

[49] Selim Ickin, Katarzyna Wac, Markus Fiedler, Lucjan Janowski, Jin-Hyuk Hong, and Anind K Dey. Factors Influencing Quality of Experience of Commonly Used Mobile applications. *IEEE Communications Magazine*, 50(4):48–56, 2012.

[50] Fortune Business Insights. Internet of Things (IoT) Market Size, Share & COVID-19 Impact Analysis, By Component (Platform, Solution & Services), By End-use Industry (BFSI, Retail, Government, Healthcare, Manufacturing, Agriculture, Sustainable Energy, Transportation, IT & Telecom, and Others), and Regional Forecast, 2023-2030. https://www.fortunebusinessinsights.com/industry-reports/internet-of-things-iot-market-100307. [Accessed on 2023-24-04].

[51] Weiwei Jiang, Gabriele Marini, Niels van Berkel, Zhanna Sarsenbayeva, Zheyu Tan, Chu Luo, Xin He, Tilman Dingler, Jorge Goncalves, Yoshihiro Kawahara, and Vassilis Kostakos. Probing Sucrose Contents

in Everyday Drinks Using Miniaturized Near-Infrared Spectroscopy Scanners. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 3(4):1–25, 2019.

[52] Xu-Qin Jiang, Xiao-Dong Mei, and Di Feng. Air Pollution and Chronic Airway Diseases: What Should People Know and Do? *Journal of Thoracic Disease*, 8(1):E31, 2016.

[53] Leilei Jin and Hong Liang. Deep Learning For Underwater Image Recognition In Small Sample Size Situations. In *Proceedings of the International Conference OCEANS 2017-Europe*, pages 1–4. IEEE, 2017.

[54] Matthew Keally, Gang Zhou, Guoliang Xing, and Jianxin Wu. Remora: Sensing Resource Sharing Among Smartphone-Based Body Sensor Networks. In *Proceedings of the 2013 IEEE/ACM 21st International Symposium on Quality of Service (IWQoS)*, pages 1–10. IEEE, 2013.

[55] Philipp H Kindt, Daniel Yunge, Gerhard Reinerth, and Samarjit Chakraborty. Griassdi: Mutually Assisted Slotless Neighbor Discovery. In *Proceedings of the 2017 16th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 93–104. IEEE, 2017.

[56] Neil E Klepeis, William C Nelson, Wayne R Ott, John P Robinson, Andy M Tsang, Paul Switzer, Joseph V Behar, Stephen C Hern, and William H Engelmann. The National Human Activity Pattern Survey (NHAPS): A Resource for Assessing Exposure to Environmental Pollutants. *Journal of Exposure Science & Environmental Epidemiology*, 11(3):231–252, 2001.

[57] Jyoti A Kodagali and S Balaji. Computer Vision and Image Analysis Based Techniques for Automatic Characterization of Fruits - A Review. *International Journal of Computer Applications*, 50(6):1–14, 2012.

[58] Ravi Kishore Kodali, Vishal Jain, and Sumit Karagwal. IoT Based Smart Greenhouse. In *Proceedings of IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pages 1–6, 2016.

[59] Eemil Lagerspetz, Xin Li, Jonatan Hamberg, Huber Flores, Petteri Nurmi, Nigel Davis, and Sumi Helal. Pervasive Data Science on the Edge. *IEEE Pervasive Computing*, 18(3):40–49, 2019.

[60] Brent Lagesse, Kevin Wu, Jaynie Shorb, and Zealous Zhu. Detecting Spies In Iot Systems Using Cyber-Physical Correlation. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom) Workshops*, pages 185–190, 2018.

[61] Daniel Lee. McKinsey IoT 2030 Forecast: MachineFi Economy Explosion is Coming. International Business Times. https://www.ibtimes.com/mckinsey-iot-2030-forecast-machinefi-economy-explosion-coming-3339071. [Accessed on 2023-04-06].

[62] Youngki Lee, Younghyun Ju, Chulhong Min, Seungwoo Kang, Inseok Hwang, and Junehwa Song. CoMon: Cooperative Ambience Monitoring Platform with Continuity and Benefit Awareness. In *Proceedings of the ACM International Conference on Mobile Systems, Applications, and Services*, pages 46–56, 2012.

[63] Chenguang Liu, Jie Hua, Changyong Hu, and Christine Julien. StaCon: Self-Stabilizing Context Neighborhood For Mobile Iot Devices. In *Proceedings IEEE International Conference on Pervasive Computing and Communications (PerCom) Workshops*, pages 361–363. IEEE, 2019.

[64] Knud Lasse Lueth. State of the IoT 2020: 12 Billion IoT Connections, Surpassing non-IoT for the First Time. 2020. https://iot-analytics.com/state-of-the-iot-2020-12-billion-iot-connections-surpassing-non-iot-for-the-first-time. [Accessed on 2022-11-01].

[65] Titti Malmivirta, Jonatan Hamberg, Eemil Lagerspetz, Xin Li, Ella Peltonen, Huber Flores, and Petteri Nurmi. Hot or Not? Robust and Accurate Continuous Thermal Imaging on FLIR Cameras. In *Proceedings of IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–9, 2019.

[66] Giancarlo Mangone, Colin A Capaldi, Zack M van Allen, and Peter G Luscuere. Bringing Nature To Work: Preferences and Perceptions of Constructed Indoor and Natural Outdoor Workspaces. *Urban Forestry & Urban Greening*, 23:1–12, 2017.

[67] Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. P2P-MapReduce: Parallel data processing in dynamic Cloud environments. *Journal of Computer and System Sciences*, 78(5):1382–1402, 2012.

[68] Summer Meza. Airbnb Hosts are Recording Their Guests With Hidden Cameras. https://www.newsweek.com/airbnb-hidden-cameras-recording-guests-739709. [Accessed on 2022-06-20].

[69] Md Moniruzzaman, Syed Mohammed Shamsul Islam, Mohammed Bennamoun, and Paul Lavery. Food Loss And Waste: Measurement, Drivers, And Solutions. In *Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 150–160. Springer, 2017.

[70] Naser Hossei Motlagh, Lauri Loven, Jacky Cao, Xiao Liu, Petteri Nurmi, Schahram Dustdar, Sasu Tarkoma, and Xiang Su. Edge Computing: The Computing Infrastructure For The Smart Megacities Of The Future. *Computer*, 55(12):54–64, 2022.

[71] Naser Hossein Motlagh, Martha Arbayani Zaidan, Eemil Lagerspetz, Samu Varjonen, Juhani Toivonen, Julien Mineraud, Andrew Rebeiro-Hargrave, Matti Siekkinen, Tareq Hussein, Petteri Nurmi, and Sasu Tarkoma. Indoor Air Quality Monitoring Using Infrastructure-Based Motion Detectors. In *Proceedings of the 2019 IEEE International Conference on Industrial Informatics*, volume 1, pages 902–907. IEEE, 2019.

[72] Naser Hossein Motlagh, Eemil Lagerspetz, Petteri Nurmi, Xin Li, Samu Varjonen, Julien Mineraud, Matti Siekkinen, Andrew Rebeiro-Hargrave, Tareq Hussein, Tuukka Petaja, Markku Kulmala, and Tarkoma Sasu. Toward Massive Scale Air Quality Monitoring. *IEEE Communications Magazine*, 58(2):54–59, 2020.

[73] Naser Hossein Motlagh, Pupu Toivonen, Martha Arbayani Zaidan, Eemil Lagerspetz, Ella Peltonen, Ekaterina Gilman, Petteri Nurmi, and Sasu Tarkoma. Monitoring Social Distancing in Smart Spaces using Infrastructure-Based Sensors. In *Proceedings of the IEEE World Forum on Internet of Things (WF-IoT)*, pages 124–129, 2021.

[74] Naser Hossein Motlagh, Matti Irjala, Agustin Zuniga, Eemil Lagerspetz, Valtteri Rantala, Huber Flores, Petteri Nurmi, and Sasu Tarkoma. Toward Blue Skies: City-Scale Air Pollution Monitoring using UAVs. *IEEE Consumer Electronics Magazine*, 12(1):21–31, 2022.

[75] Kazuya Murao, Tsutomu Terada, Ai Yano, and Ryuichi Matsukura. Detecting Room-To-Room Movement By Passive Infrared Sensors In

Home Environments. In *Proceedings of the First Workshop on Recent Advances in Behavior Prediction and Pro-Active Pervasive Computing*, pages 1–12. ACM, 2012.

[76] Ngoc Thi Nguyen, Agustin Zuniga, Hyowon Lee, Pan Hui, Huber Flores, and Petteri Nurmi. (M)Ad To See Me? Intelligent Advertisement Placement: Balancing User Annoyance and Advertising Effectiveness. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, volume 4, pages 1–26. ACM, 2020.

[77] Adam J Oliner, Anand P Iyer, Ion Stoica, Eemil Lagerspetz, and Sasu Tarkoma. Carat: Collaborative Energy Diagnosis for Mobile Devices. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems (SenSys 2013)*, pages 1–14. ACM, 2013.

[78] Kamlesh S Patle, Riya Saini, Ahlad Kumar, and Vinay S Palaparthy. Field Evaluation of Smart Sensor System for Plant Disease Prediction using LSTM Network. *IEEE Sensors Journal*, 22(4):3715–3725, 2021.

[79] Ella Peltonen, Eemil Lagerspetz, Jonatan Hamberg, Abhinav Mehrotra, Mirco Musolesi, Petteri Nurmi, and Sasu Tarkoma. The Hidden Image of Mobile Apps: Geographic, Demographic, and Cultural Factors in Mobile Usage. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–12, 2018.

[80] Alberto Petrillo, Alessandro Salvi, Stefania Santini, and Antonio Saverio Valente. Adaptive Multi-Agents Synchronization For Collaborative Driving Of Autonomous Vehicles With Multiple Communication Delays. *Transportation Research, Part C: Emerging Technologies*, 86: 372–392, 2018.

[81] Lukasz Piwek, David A Ellis, Sally Andrews, and Adam Joinson. The Rise Of Consumer Health Wearables: Promises And Barriers. *PLoS Medicine*, 13(2):e1001953, 2016.

[82] United Nations Environment Programme. Food Waste Index Report 2021. https://wedocs.unep.org/20.500.11822/35280. [Accessed on 2023-02-14].

[83] Xiaohui Qiao, Qiang Zhang, Dongbin Wang, Jiming Hao, and Jingkun Jiang. Improving Data Reliability: A Quality Control Practice for Low-cost PM2.5 Sensor Network. *Science of The Total Environment*, 779:146381, 2021.

[84] Marko Radeta, Agustin Zuniga, Naser Hossein Motlagh, Mohan Liyanage, Ruben Freitas, Moustafa Youssef, Sasu Tarkoma, Huber Flores, and Petteri Nurmi. Deep Learning and the Oceans. *Computer*, 55(5): 39–50, 2022.

[85] Lenin Ravindranath, Jitendra Padhye, Sharad Agarwal, Ratul Mahajan, Ian Obermiller, and Shahin Shayandeh. AppInsight: Mobile App Performance Monitoring in the Wild. In *Proceedings of the 10th Symposium on Operating Systems Design and Implementation*, pages 107–120. USENIX, 2012.

[86] Sharon Richardson. Predicting Presence in Urban Outdoor Spaces. *IEEE Pervasive Computing*, 18(3):21–30, 2019.

[87] Mikko Rinta-Homi, Naser Hossein Motlagh, Agustin Zuniga, Huber Flores, and Petteri Nurmi. How Low Can You Go? Performance Trade-Offs In Low-Resolution Thermal Sensors For Occupancy Detection: A Systematic Evaluation. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, volume 5, pages 1–22. ACM, 2021.

[88] Marc Roessler. How to Find Hidden Cameras. http://www.tentacle.franken.de/papers/hiddencams.pdf. [Accessed on 2022-05-16].

[89] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[90] Andres Sevtsuk and Carlo Ratti. Does Urban Mobility Have a Daily Routine? Learning From the Aggregate Data of Mobile Networks. *Journal of Urban Technology*, 17(1):41–60, 2010.

[91] Dmitrii Shadrin, Andrey Somov, Tatiana Podladchikova, and Rupert Gerzer. Pervasive Agriculture: Measuring and Predicting Plant Growth Using Statistics and 2D/3D Imaging. In *Proceedings of the IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6. IEEE, 2018.

[92] Elahe Soltanaghaei and Kamin Whitehouse. Walksense: Classifying Home Occupancy States Using Walkway Sensing. In *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments*, pages 167–176, 2016.

[93] Andrey Somov, Dmitry Shadrin, Ilia Fastovets, Artyom Nikitin, Sergey Matveev, and Oleksii Hrinchuk. Pervasive Agriculture: IoT-Enabled Greenhouse for Plant Growth Control. *IEEE Pervasive Computing*, 17(4):65–75, 2018.

[94] Yunpeng Song, Yun Huang, Zhongmin Cai, and Jason I Hong. I'm All Eyes and Ears: Exploring Effective Locators for Privacy Awareness in IoT Scenarios. In *Proceedings of ACM Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[95] Sebastian Sonntag, Jukka Manner, and Lennart Schulte. NetRadar-Measuring the Wireless World. In *Proceedings of the 11th IEEE International Symposium on Modeling & Optimization in Mobile, Ad Hoc & Wireless Networks (WiOpt, 2013)*, pages 29–34. IEEE, 2013.

[96] Edward S Spang, Laura C Moreno, Sara A Pace, Yigal Achmon, Irwin Donis-Gonzalez, Wendi A Gosliner, Madison P Jablonski-Sheffield, Md Abdul Momin, Tom E Quested, Kiara S Winans, and Thomas P Tomich. Food Loss and Waste: Measurement, Drivers, and Solutions. *Annual Review of Environment and Resources*, 44:117–156, 2019.

[97] Statista. Number of Apps Available in Leading App Stores as of 3rd Quarter 2018. 2018. https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/. [Accessed on 2019-2-19].

[98] Birru Dereje Teshome. Spy Camera Epidemic in Korea: A Situational Analysis. *Asian Journal of Sociological Research*, 2(1):1–13, 2019.

[99] Florian Thevenon, Chris Carroll, and Joao Sousa. Plastic Debris in the Ocean: The Characterization of Marine Plastics and Their Environmental Impacts. Technical report, International Union for Conservation of Nature (IUCN), 2014.

[100] Leonardo Tonetto, Eemil Lagerspetz, Aaron Yi Ding, Jörg Ott, Sasu Tarkoma, and Petteri Nurmi. The Mobility Laws of Location-based Games. *EPJ Data Science*, 10(1):1–10, 2021.

[101] Tuyen X Tran, Abolfazl Hajisami, Parul Pandey, and Dario Pompili. Collaborative Mobile Edge Computing in 5G networks: New Paradigms, Scenarios, and Challenges. *IEEE Communications Magazine*, 55(4):54–61, 2017.

[102] Moya Tung. How to Find Hidden Cameras. https://www.securitybees.com/blogs/home/48443077-how-to-find-hidden-cameras. [Accessed on 2022-06-20].

[103] Ash Tyndall, Rachel Cardell-Oliver, and Adrian Keating. Occupancy Estimation Using a Low-Pixel Count Thermal Imager. *IEEE Sensors Journal*, 16(10):3784–3791, 2016.

[104] Fabrizio Valpreda and Ilaria Zonda. Grüt: A Gardening Sensor Kit for Children. *Sensors*, 16(2):231–244, 2016.

[105] Mehmet C Vuran, Abdul Salam, Rigoberto Wong, and Suat Irmak. Internet of Underground Things in Precision Agriculture: Architecture and Technology Aspects. *Ad Hoc Networks*, 81:160–173, 2018.

[106] Nan-Nan Wang, Da-Wen Sun, Yi-Chao Yang, Hongbin Pu, and Zhiwei Zhu. Recent Advances in the Application of Hyperspectral Imaging for Evaluating Fruit Quality. *Food Analytical Methods*, 9:178–191, 2016.

[107] Yong Wang, Dianhong Wang, Qian Lu, Dapeng Luo, and Wu Fang. Aquatic Debris Detection Using Embedded Camera Sensors. *Sensors*, 15(2):3116–3137, 2015.

[108] Derek Ward. Hidden Camera Detectors Tested. https://ipvm.com/reports/hidden-cameras-finder. [Accessed on 2022-06-20].

[109] David E Williams. Low Cost Sensor Networks: How Do We Know the Data Are Reliable? *ACS Sensors*, 4(10):2558–2565, 2019.

[110] Feng Xia, Jinzhong Wang, Xiangjie Kong, Zhibo Wang, Jianxin Li, and Chengfei Liu. Exploring Human Mobility Patterns in Urban Scenarios: A Trajectory Data Perspective. *IEEE Communications Magazine*, 56(3):142–149, 2018.

[111] Shili Xiang, Lu Li, Si Min Lo, and Xiaoli Li. People-centric Mobile Crowdsensing Platform for Urban Design. In *Proceedings of the International Conference on Advanced Data Mining and Applications (ADMA)*, pages 569–581. Springer, 2017.

[112] Miao Yang, Jintong Hu, Chongyi Li, Gustavo Rohde, Yixiang Du, and Ke Hu. An in-depth survey of underwater image enhancement and restoration. *IEEE Access*, 7:123638–123657, 2019.

[113] Zhigang Yin, Mayowa Olapade, Mohan Liyanage, Farooq Dar, Agustin Zuniga, Naser Hossein Motlagh, Xiang Su, Sasu Tarkoma, Pan Hui, Petteri Nurmi, and Huber Flores. Toward City-Scale Litter Monitoring Using Autonomous Ground Vehicles. *IEEE Pervasive Computing*, 21 (3):74–83, 2022.

[114] Lingjing Yu, Bo Luo, Jun Ma, Zhaoyu Zhou, and Qingyun Liu. You Are What You Broadcast: Identification of Mobile and IoT Devices from (Public) WiFi. In *Proceedings of USENIX Security Symposium (USENIX Security 20)*, pages 55–72. USENIX Association, 2020.

[115] Liyong Zhou, Vikram Chalana, and Yongmin Kim. PC-based Machine Vision System for Real-time Computer-aided Potato Inspection. *International Journal of Imaging Systems and Technology*, 9(6):423–433, 1998.

[116] Jianhong Zou, Qianchuan Zhao, Wen Yang, and Fulin Wang. Occupancy Detection in the Office by Analyzing Surveillance Videos and Its Application to Building Energy Conservation. *Energy and Buildings*, 152:385–398, 2017.

[117] Agustin Zuniga, Huber Flores, Eemil Lagerspetz, Sasu Tarkoma, Pan Hui, Jukka Manner, and Petteri Nurmi. Tortoise or Hare? Quantifying the Effects of Performance on Mobile App Retention. In *Proceedings of The World Wide Web Conference (WWW)*, pages 2517–2528, 2019.

[118] Agustin Zuniga, Huber Flores, and Petteri Nurmi. Ripe or Rotten? Low-cost Produce Quality Estimation Using Reflective Green Light Sensing. *IEEE Pervasive Computing*, 20(3):60–67, 2021.

[119] Agustin Zuniga, Naser Hossein Motlagh, Huber Flores, and Petteri Nurmi. Smart Plants: Low-Cost Solution for Monitoring Indoor Environments. *IEEE Internet of Things Journal*, 9(22):23252–23259, 2022.

[120] Agustin Zuniga, Naser Hossein Motlagh, Mohammad A Hoque, Sasu Tarkoma, Huber Flores, and Petteri Nurmi. See No Evil: Discovering Covert Surveillance Devices Using Thermal Imaging. *IEEE Pervasive Computing*, 21(4):33–42, 2022.