UNIVERSITY OF TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE

DOCTORAL THESIS

LEARNING WHEN TO ASK QUESTIONS ABOUT THE CURRENT CONTEXT

Author

Haonan Zhao

Advisor

Prof. Fausto Giunchiglia

April, 2025

Acknowledgements

The journey of completing my PhD has been a challenging yet rewarding experience. Throughout this process, I have been fortunate to receive support and encouragement from numerous individuals, without whom this work would not have been possible. I would like to express my deepest gratitude to my advisor, Prof. Fausto Giunchiglia, whose guidance, wisdom, and patience have been invaluable. The opportunity to embark on this journey has been transformative, teaching me how to conduct PhD-level research and the essential methodologies of the field. His insightful feedback and unwavering support shaped my research and helped me grow into a real young researcher. The research abroad period at LIPADE has been possible thanks to Prof. Themis Palpanas and the group. Special thanks go to my colleagues and friends at the KnowDive group, whose camaraderie and intellectual discussions have been a source of inspiration and motivation. We have spent countless hours discussing the right approaches to research, and their invaluable suggestions have greatly benefited my work. In addition, I must thanks also to the reviewers, Prof. Styliani Kleanthous and Prof. Wanyi Zhang, whose valuable comments helped me improve this thesis a lot. On a personal note, I am deeply thankful to my family. To my parents, your unconditional love and belief in me have been my foundation. To my wife, Xiaoyue Li, your patience, encouragement, and endless support have been my anchor throughout this journey. I will never forget my time in PhD, because of PhD, Trento has become my second home!

Haonan Zhao

April 2025

Abstract

Mobile phones, being ubiquitous and always accessible, provide an opportunity for symbiotic interaction between humans and machines, enabling the understanding of human behavior at any time and from any location through sensor data. While sensor data offers an objective view of reality, it fails to capture the subjective motivations behind an individual's behavior; to learn the subjective motivations, the general way is to ask context questions to the user. However, a key problem was raised because the answers from people are always not of good quality. This thesis aims to solve the problem of inadequate human input quality, typically marked by a high number of wrong answers, for three personas: researchers, participants, and the platform. For researchers, the thesis demonstrates how users' reaction times before starting to respond significantly affect answer quality, and it designs a comprehensive methodology for enabling participants to provide high-quality input data. For participants, it identifies optimal moments or situational contexts when they are most willing to provide high-quality input data. For the platform, it designs a novel system to automatically collect high-quality human input data, comprising a smartphone application for sensor data and user feedback collection, a calendar for human-machine collaboration configuration, a dashboard for qualitative and quantitative experiment progress feedback, user-centric notifications, and a machine-learning module (pbgc-Forest algorithm) to facilitate human-machine interactions. This thesis paves the way for new opportunities in mobile user interfaces and interaction technologies to enhance user experiences and improve the quality and reliability of data collected from participants, ensuring more accurate and meaningful research outcomes.

Keywords: Human-Machine interaction, Quality of user answers, Context, Data quality, Mobile user interfaces.

Contents

List of Tables

List of Figures

Glossary

Annotation: The label pertains to data derived from sensors, which serve to enrich the content of this sensor and enhance its design, validation, and user interaction. Such labels are characteristic of Human Activity Recognition (HAR) methodologies and are typically generated by experts who annotate sensor databases.

Answer Quality: In this context, answer quality refers to the inaccuracies in answers provided by humans to questions automatically dispatched by machines. However, in contemporary AI research, quality is often quantified by the number of missing answers. This thesis, conversely, focuses on ensuring participants provide correct answers that accurately reflect the real-world context surrounding them. By doing so, machines can better learn and understand the world, ultimately aiding participants in adopting improved lifestyles.

Big Data: Generally, the term refers to datasets that are too vast or complex to be managed using traditional approaches and software. In our context, Big Data encompasses all currently produced data surrounding an individual, starting from social media streams (posts, photos, messages), extending to sensor data used to observe and measure human behavior (e.g., smartphone GPS), and including data from phone applications, streaming services, e-commerce platforms, and more. At the same time, the primary argument of this thesis is that such data fails to fully capture the individual within their context; big data can not capture the motivations behind each individual's behavior.

Context: Context should encompass all that surrounds an individual in their daily life, as perceived from the person's perspective and other information sources. This thesis begins with the premise that the individual is the foremost expert on their own context, given their immersion and capacity to describe it. At any given moment and location, a person can typically recount where

they were, what activities they were engaged in, who accompanied them, and their emotional state. Additionally, contextual information can be augmented through other sources, such as sensors and databases pertinent to the observed situation. The continuous interaction between these two information sources over time delineates the individual's situational context. This concept facilitates the quantitative generation of Big Data and Thick Data.

ESM: The Experience Sampling Method (ESM), also known as the daily diary method or Ecological Momentary Assessment (EMA), represents an intensive longitudinal research methodology. This approach relies on queries, including questions and psychometric scales, to assess participants' thoughts, feelings, behaviors, and environments at multiple points over time. Unlike cross-sectional or traditional longitudinal studies, the ESM approach provides real-time observations of individuals' behaviors (e.g., identifying specific times of day when a person experiences particular emotions or tracking the progress of therapy) with greater granularity, reflecting daily life contexts. However, the ESM method does not incorporate sensor data (e.g., GPS) or potential interactions with these sensors. Consequently, ESM fails to account for various aspects of daily life and their interactions with individuals. Moreover, its prolonged application may increases participant burden and leads to annoyance and lower-quality responses.

GDPR: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons concerning the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). In this thesis, all the experiments we ran follow the GDPR.

Human-in-the-loop: The notion of human-in-the-loop lacks a precise definition in existing literature. We define it as a (symbiotic) type of human-machine collaboration where:

- The machine assists the user in her daily life in various ways, such as

suggesting tasks, reminding her of duties, and autonomously reacting to external inputs.

- The human aids the machine in understanding her life, including personal and reference contexts, based on the data collected by the machine. This assistance involves the human providing semantic interpretations of the collected data, an extensive form of what is known in machine learning as human supervision, which facilitates supervised learning.

- The machine helps the human by proactively seeking assistance when needed, as the human typically remains unaware of the machine's difficulties unless informed. This represents a novel domain in the design of human-machine interactions.

I-Log: i-Log is an innovative application developed by the KnowDive group within the Department of Information Engineering and Computer Science at the University of Trento. Its primary purpose is to collect data for research with the user's informed consent. i-Log leverages the smartphone's internal sensors to gather data and dispatches context-sensitive questions to users. The final objective is to analyze user habits, enabling the provision of personalized services and the creation of valuable research datasets for future studies.

Sensors: A sensor is defined as any device that autonomously detects a physical phenomenon and generates a corresponding signal as output. In our experiment, we specifically refer to sensors that monitor an individual's daily life and their interactions with the environment. Consequently, the primary examples of sensors are those integrated into smartphones, as collected by i-Log, given their ubiquitous presence. These sensors facilitate the observation of physical phenomena such as movement (e.g., GPS, accelerometer), atmospheric and environmental conditions (e.g., temperature), and the individual's interaction with their device (e.g., number and type of applications used, screen touch events).

Additionally, other devices, such as smartwatches that track biometric information (e.g., heartbeat), and various IoT home automation sensors, are considered. Therefore, the data from these sensors can contribute to the generation of quantitative Big Data.

Thick Data: Different from big data, the term of thick data refers to qualitative data that offers insights into the context, moods, and social dynamics that quantitative data alone cannot provide. In our context, Thick data encompasses in-depth, narrative-rich information that reveals the underlying motivations, perceptions, and cultural nuances of individuals. It goes beyond surface-level metrics, capturing the stories and experiences behind human behavior through methods like ethnography, interviews, and participant observation. Unlike big data, which provides broad, numerical insights, thick data dives deep into understanding the 'why' behind actions and decisions, providing a richer, more contextualized picture of individuals in their environments. The primary argument of this thesis is that while big data offers extensive breadth, thick data provides essential depth, enabling a more comprehensive understanding of individuals within their unique social and cultural settings.

Time Diary: Similar to the Experience Sampling Method (ESM), the Time Diary approach is a subset of daily diary methods. It is specifically designed to capture an individual's daily activities, considering various aspects such as location, activities, and companions throughout the day. This approach is particularly relevant for understanding citizens' habits and the challenges they encounter daily. Consequently, it is employed by social research institutes such as EUROSTAT and ISTAT in Europe and Italy. In our case, the Harmonized European Time Use Surveys (HETUS) standard (or the American Time Use Survey, ATUS, in the United States) is used to pose daily questions on the key aspects of an individual's context.

Chapter 1

Introduction

1.1 Motivation

Smartphones have become an integral part of our daily lives, enabling us to receive notifications and respond to them anytime and anywhere [159]. Numerous studies have illustrated how predictable various aspects of human behavior are through smart devices, including research on mobility [34, 85, 174, 51, 5], social interactions [60, 59], and preferences for favorite places [6] and friends [134]. Contextual information has been shown to be valuable in applications such as health and physical activity monitoring [148, 102, 208, 207], mental health monitoring [195, 197, 210], and elderly care [122, 20, 198, 209], as well as in predicting individual behaviors and traits [56, 90, 196, 145]. The challenge in this context is to develop high-quality characterizations of this information [117]. Early and recent work on the general topic of context recognition include studies such as [21, 156, 92, 182, 96] and [149, 189, 190, 33]. However, many such studies have primarily focused on the use of sensor data alone. This focus has led to various types of errors, most notably in terms of data validity, the accuracy of the indicators measuring the phenomenon, and data completeness, where missing data should ideally occur randomly rather than systematically.

Smartphone sensors are widely used to collect data, which, in turn, provide features that allow us to comprehend the world [211]. This type of data is often

termed as *big data* [52], see the top-left of Figure 2.1. The main limitation is that, while sensors provide an excellent objective representation of reality, they do not account for the subjective motivations behind an individual's activities. On the other hand, *thick data*, defined in [27], is a class of data sources that align with ethnographically collected and meticulously analyzed observational data, middle-right part of Figure 2.1. Essentially, while big data provides large-scale quantitative insights, thick data delivers qualitative insights into human behavior, experiences, and motivations. When both are integrated, termed as *big-thick data*, they render a more holistic view of human needs and preferences to machines, see the top-right corner of Figure 2.1.

Recognizing that human behavior is based on individuals' subjective perceptions of their current context, which is learning from the big-thick data. A few studies have centered the role of context in their analyses [214]. Which use sensor data as additional, albeit crucial, information. They demonstrate that when a user provides a subjective description of their current context, any target modality (e.g., location or activity) becomes significantly more predictable by also incorporating information from other modalities (e.g., time, user characteristics, social ties). This opens up the possibility for machines to gather information and learn about every aspect of a person's daily life, with high-impact applications across research areas focusing on the flow of individual behaviors and thoughts [98, 202] and the ecology of human development [35]. Examples of such applications include medical behavior [180], clinical psychology, social sciences, human-computer interaction, and more recently, human-in-the-loop AI [26].

1.2 Problem

Accurately capturing human behavior based on big-thick data necessitates active collaboration between users and machines. However, a significant limi-

tation is the inaccuracy of human input, as discussed in [100, 193]. Frequent notifications from smartphones can be particularly intrusive, especially when they interrupt users during active periods, leading to annoyance and decreased efficiency in work and study [155]. For instance, when an individual receives a notification while driving or in a meeting, they cannot provide immediate feedback, which poses a problem since people may not accurately recall their behavior at the time of receiving the notification. Another major challenge is that machines require users to answer a high volume of questions, often resulting in low-quality responses [133]. Existing methods rely on fixed or random schedules, which may not align well with participants' availability or willingness to respond [214, 26], thus diminishing the effectiveness of human-AI collaboration. To collect accurate big-thick data, which we can call collecting big-thick data with high quality, for understanding human behavior, the key problem is *the absence of a platform capable of collecting high-quality big-thick data by understanding participants*. This primary issue leads to four subsequent problems based on different personas, namely, researcher, participant, and platform.

The problems from researcher's perspective are as follows:

- **Problem 1:** Despite extensive research aimed at collecting human behavior data via sensors and questionnaires, there is currently no standard for determining the quality of human input data, especially during the data collection process.
- **Problem 2:** Currently, there is no comprehensive methodology that helps researchers design experiments to enable participants to provide high-quality input data.

The problem from participant's perspective is as follows:

- **Problem 3:** Participants often receive context questions about their daily behavior at any time, which can lead to annoyance as their daily routines

are disrupted by numerous notifications. Bailey et al. [14] found that when peripheral tasks interrupt the execution of primary tasks, users require 3% to 27% more time to complete tasks and experience 31% to 106% more annoyance. Therefore, it is crucial to identify the optimal moments, or situational contexts, when participants are most willing to provide high-quality input data.

The problem from platform's perspective is as follows:

- **Problem 4:** There are numerous research platforms that collect and process sensor and human input data from smart devices, aiming to gather big-thick data. However, these platforms often overlook the quality of the data they collect, focusing instead on quantity. As a result, there is a pressing need for a novel system or platform that not only collects sufficient data but also ensures high-quality data by understanding participants' daily behaviors and answer patterns.

Given the above challenges, it is crucial for humans to collaborate with AI machines. Initially, humans assist machines in learning their context information and answer behavior, and subsequently, the machine can help minimize annoyance by optimizing the timing of notifications. This necessitates designing a methodology to identify the key factors that impact human input data quality, and then indicate in which situational contexts humans consistently provide high-quality input data. In essence, the goal is to ensure that humans are not disrupted by intrusive notifications, thereby fostering a more seamless and productive interaction between humans and AI.

1.3 Solution

To deliver a platform that can automatically collect high-quality big-thick data and ensure alignment between machine perception of human behavior and hu-

man description over the lifespan of the AI, the key challenges to consider are:

- Data Accuracy: Ensuring that the data collected is accurate and reliable, minimizing errors and inconsistencies.
- Context Understanding: Capturing both objective sensor data and subjective human input to fully understand the individual's context.
- Scalability: Designing the platform to handle large volumes of data and a wide range of contexts efficiently.
- Interdisciplinary Integration: Combining insights from various fields to develop a comprehensive understanding of human behavior.
- Adaptive Learning: Enabling the AI to continuously learn and adapt to changes in human behavior and context over time.
- User Interface: Creating an intuitive and user-friendly interface to facilitate seamless interactions between users and the AI.

It is fundamental that the machine comprehends the situational contexts in which participants provide high-quality input data. Therefore, the research objective of this thesis is to *develop a configurable platform that enables the automatic collection of high-quality big-thick data from participants while ensuring minimal disturbance from the AI system.* To achieve this, our platform and methodology must address the aforementioned problems with the following solutions:

- **Solution 1:** Identify key factors that impact human input data, ensuring these factors are easily calculable during the data collection process, thereby providing an intuitive impression of data quality. This solution will be handled in Chapter 4.

- **Solution 2:** Design a novel methodology with a machine learning algorithm that can continuously and automatically learn the answer behavior of participants while achieving high performance in predicting the moments when participants are most likely to provide high-quality input data. Also, develop a novel language that offers flexibility for researchers to design questions and sensor collections for gathering big-thick data while providing participants with enhanced flexibility in responding to questions. This solution will be handled in Chapter 5.

- **Solution 3:** Based on the descriptive statistics and machine learning methods to determine the situational contexts and the number of questions in which users are most likely to provide high-quality answers. This solution will be handled in Chapter 6.

- **Solution 4:** Create a configurable platform that enables the collection of high-quality big-thick data from participants while ensuring minimal disturbance from the AI system. This solution will be handled in Chapter 7.

1.4 Contribution

Our goal is to design a platform that can learn from human input data to find a way for individuals to provide high-quality data. We aim to predict answer quality based on the user's context at any given moment, using data streams from their smart devices. Figure 1.1 provides a simplified overview of the scenario. The input data from participants is obtained from sensors and their responses to questions. Our AI system, named Planner, can learn each participant's answer behavior and daily routine to devise a strategy for sending questions that elicit high-quality answers while minimizing disruptions to their daily lives. Specifically, the contributions of this work are as follows:

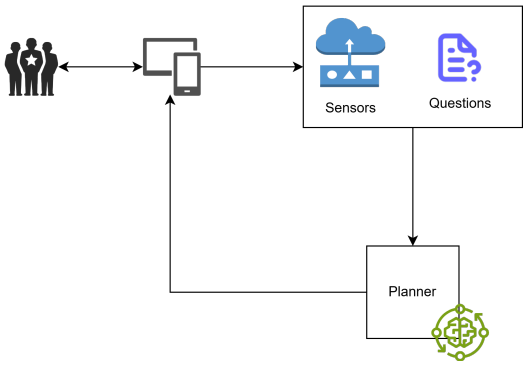


Figure 1.1: Simplified overview of the proposed solution.

1. Demonstrating that long reaction and completion times can lead to low answer quality, and the reaction time is the key factor that can impact the human input data quality. Thereby introducing a novel parameter that provides a visual reference for both researchers and participants during the data collection process.

2. Developing a situational context model to analyze human behavior from five different dimensions, namely temporal context, spatial context, internal context, object context, and social context. We found that participants are more likely to provide high-quality answers when in a negative mood, in settings such as a university, library, or home, while using social media applications or during public transportation, and when spending time with family or alone.

3. Defining a planning language based on the iCalendar standard that offers flexibility in designing recurrence rules for sending questionnaires to users. This flexibility, via the calendar, also allows users to adjust the answer times for each questionnaire.

4. Designing a novel machine learning algorithm named pbgc-Forest, which combines the advantages of deep forest and PCA, to enhance the planner component's ability to accurately predict the moments when participants

are most likely to provide high-quality answers.

5. Designing a strategy to optimize the number of questions to obtain high-quality answers while minimizing interruptions. Our results show that presenting ten questions per day maintains the highest quality of answers, minimizing potential disruptions to participants' daily routines and maximizing high-quality responses.

6. Creating a user-friendly platform that integrates sensor data collection and user feedback. This platform includes a machine learning model named Planner, which learns individual participants' answer behaviors and daily routines. We demonstrate the effectiveness of our approach through case studies.

By achieving these objectives, we hope to enhance the quality of data collected from participants, ultimately contributing to more accurate and meaningful research findings.

1.5 List of publications

This thesis is mainly based on the following publications:

- Ivano Bison, **Haonan Zhao.** *Factors Impacting the Quality of User Answers on Smartphones [C].* The Workshop on Diversity-aware Hybrid Human-Artificial Intelligence: HHAI conference workshop. 2023. https://ceur-ws.org/Vol-3456/short5-2.pdf

- **Haonan Zhao,** Fausto Giunchiglia. *Scheduling Real-Time Acquisition of Context Information [C].* European Conference on Artificial Intelligence: ECAI conference workshop. 2023. https://mrc.kriwi.de/2023/download/paper-1-2.pdf

- Ivan Kayongo, **Haonan Zhao,** Leonardo Malcotti, Fausto Giunchiglia. *A Methodology and System For Big-Thick Data Collection [C].* Informatik Festival. 2024.
https://dl.gi.de/server/api/core/bitstreams/346e0026-1454-4bf5-81cd-3ccca1cbed61
content

- Andrea Bontempelli, Marcelo Rodas Britez, Xiaoyue Li, **Haonan Zhao,** Luca Erculiani, Stefano Teso, Andrea Passerini, and Fausto Giunchiglia. *Lifelong Personal Context Recognition [C].* The First International Workshop on Human-Centered Design of Symbiotic Hybrid Intelligence co-located with The First International Conference on Hybrid Human-Artificial Intelligence (HHAI). 2022
https://arxiv.org/abs/2205.10123

- Ivan Kayongo, Leonardo Malcotti, **Haonan Zhao,** Fausto Giunchiglia. *A methodology and a platform for high-quality rich personal data collection [J].* AI Communications Journal 2025. (Accepted)
https://arxiv.org/pdf/2501.16864

- **Haonan Zhao,** Ivano Bison, Fausto Giunchiglia. *What Impacts the Quality of the User Answers when Asked about the Current Context? [J].* the Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT), 2025. (Ready to submit)

- **Haonan Zhao,** Xiaoyue Li, Matteo Busso, Ivano Bison, Fausto Giunchiglia. *Leveraging Personality Traits and Context in Mood Prediction. [C].* European Conference on Artificial Intelligence: ECAI conference, 2025. (Ready to submit)

1.6 Structure

Despite the abundance of data about people and some research on human behavior, uncertainties about the quality of human input data and the absence of fundamental variables hinder accurate big-thick data representation. By considering individuals' perspectives on their contexts, it is possible to generate meaningful information that fosters advanced research and new applications. However, while these statements are attractive in theory, they require scientific and interdisciplinary operationalization and validation. Currently, there is no end-to-end methodology capable of handling high-quality big-thick data. Therefore, a method is needed to put theory into practice.

The main contribution of this thesis is a methodology that identifies factors indicating the quality of human input data during experiments. Additionally, it introduces a platform for automatically collecting high-quality personal big-thick data for future use. We show the structure of this thesis as follows:

Chapter 2 presents the current state of the art in the field of personal context, as our goal is to thoroughly understand human behavior. To achieve this, we begin by examining how researchers have designed models to represent human behavior. Following this, we analyze current applications used to collect context information, referred to as big-thick data, and explore the methods employed in existing research to evaluate the quality of this data.

Chapter 3 describes the two databases we used in this thesis, namely the databases collected from the Smart Unitn 2 and WeNet projects. First, we provide an overview of the i-Log application, the smartphone app used to collect big-thick data. Then, we detail the two datasets, including the length of the experiments, the sensors used for data collection, and the questions asked to the participants.

Chapter 4 describes how we can identify the key factors that monitor human answer quality, specifically reaction time and completion time during the

question-answering process. Initially, we draw inspiration from the total survey error paradigm and then employ statistical analysis methods to demonstrate the relationship between answer quality and reaction time. We use GPS coordinates as a verification method. The results show that reaction time and completion time are factors that can be easily collected during the experiment and have a direct relationship with human answer quality. By analyzing these factors, we can better understand and improve the quality of data collected from participants, ensuring more accurate and reliable research outcomes.

Chapter 5 outlines our methodology for assessing human answer quality, which includes the design and presentation of personal context. Through personal context, we collect big-thick data that represents human behavior. We then demonstrate our planning language, which is used to represent temporal context information and provide flexibility for recurring question collection. Finally, we delve into the specifics of the pbgc-Forest algorithm, the central element of this thesis, which enables the future platform to gather high-quality human responses automatically.

Chapter 6 demonstrates the practical application of our methodology in the real world. We present three case studies to illustrate how our methodology enhances the quality of answers collected from participants when asked context questions. Firstly, we explain how understanding situational context increases the likelihood of obtaining high-quality responses and identify the specific situational contexts in which participants can provide high-quality answers. Secondly, we use machine learning algorithms to validate our results, showing that human differences also significantly impact answer quality. Thirdly, we demonstrate that reducing the number of questions and timing them appropriately can lead to higher-quality responses.

Chapter 7 details the design of our platform that can autonomously collect high-quality personal big-thick data. This chapter outlines the overall architecture of the platform, highlighting how each component collaborates to ensure

its functionality. Firstly, we provide an overview of the four main components of the platform, namely, ISAC, Dashboard, Calendar, and Planner. The last one is the central element of this thesis. It uses advanced algorithms, including pbgc-Forest, to learn from participants' behavior and situational context, optimizing the timing and delivery of questions to gather high-quality human responses. The chapter delves into how these components collaborate to create a seamless and efficient data collection process. By integrating these elements, our platform aims to minimize disturbances to participants while maximizing the quality of the data collected.

Finally, Chapter 8 discusses the ethical and societal implications of this work and outlines future research directions. Which also concludes by summarizing the main contributions of this thesis and addressing its limitations.

Chapter 2

State of the Art

The objective of this thesis is to find an appropriate context for eliciting high-quality answers to the context questions. So, Section 2.1 first explores the concept of context, with an emphasis on personal context, which can be the key parameter to impact if a person can give an answer with high quality. This section is organized into three parts: the definition of context, the representation of context, and the application of context in various areas. Subsequently, Section 2.2 addresses the collection of context data, which is bifurcated into big data (smartphone sensor data) and thick data (human input data). The methodology and platforms for collecting both big data and thick data will then be examined. Finally, Section 2.3 reverts to the core topic, focusing on the quality of big-thick data. This section encompasses a discussion on the state-of-the-art techniques for analyzing the quality of sensor data and human input data, the application of the total survey error paradigm by sociologists to evaluate human answer quality, and the methods employed by AI researchers to assess human answer quality.

2.1 Personal context

2.1.1 Context definition

Approximately thirty years ago, Schilit et al. [165] introduced the concept of context, defining it as "locations, identities of nearby individuals and objects, and changes to those objects." Similarly, Brown et al. [36] depicted context as "locations, varying user roles, time, seasons." In addition, Dey et al. [55] defines context as closely aligned with our understanding: "the user's physical, social, emotional, or informational states." Dey and Abowd [4] define the context in a more comprehensive manner. They state that "context is any information that can be used to characterize the situation of an entity. An entity is a person or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves." However, to define context, we require a context model that can address the dynamic and open-ended nature of the world, capable of defining environments and concepts. Consequently, we adopt the context definition from Giunchiglia et al. [74], which models context as a local view centered around the individual and comprises the following four dimensions:

- *Activity:* this refers to the primary activity that the person, around whom the context is centered, is performing. This is derived from the question "what are you doing?", e.g., studying and sleeping.

- *Location:* this represents the main location where the person is currently situated, answering the question "where are you?", e.g., home, university and in street.

- *Social:* this includes the individuals who are part of the current environment, answering the question "who are you with?", e.g., colleagues, classmates and parents.

- *Object:* this encompasses the objects that are in the vicinity or being used by the person, answering the question "what are you with?", e.g., computer, pens and books.

Based on the aforementioned definitions, it is apparent that human responses to the four questions exhibit subjectivity; thus, context modeling should be person-centric. Consider the following scenario:

The instructor's interaction with high school students within the classroom setting exemplifies a complex model, which is inherently dependent on the observer's perspective. The spatial context diverges: for the instructor, the classroom is a professional domain, whereas for the students, it serves as a learning environment. This dichotomy demonstrates that a single environment can embody multiple roles contingent upon the viewpoint. Regarding activities, students assume the role of listeners, while the instructor actively disseminates information via the blackboard. Each participant engages with specific tools: the instructor employs chalk for board writing, whereas students use pens for note-taking. Such previous contextual subjectivity is encapsulated through teleologies, as explicated in [77]. Contextual elements are classified into objects, functions, and actions. Objects are tangible and finite entities such as buildings, individuals, and vehicles. Functions denote the intended utility from the user's perspective, while actions encompass the dynamic transitions objects undergo to fulfill these functions.

2.1.2 Context representation

From the above section, we know how to define context, and the following step is to represent context. In this part, we find that the schema and data levels of the knowledge graph proposed in previous works by Giunchiglia et al. [79]. Based on these notions, we model the personal context as the Entity type graph (ETG). The ETG is the schema that defines the relation between the different

concepts, namely objects, functions, and actions. A lot of literature exists on the definition of ETG, e.g., [124, 78].

Also, in recent years, a new concept has been introduced in the context of knowledge representation: the Entity-Topic Graph (ETopicG), which plays a crucial role in organizing and structuring information. The ETopicG focuses on the relationships between entities and their associated topics, providing a high-level view of the knowledge domain [126]. On the other hand, the Entity Graph (EG) delves into the detailed connections between entities, capturing the intricate web of relationships and attributes. By combining these two representations, researchers can achieve a comprehensive and nuanced understanding of the knowledge space, facilitating advanced applications in areas such as semantic search, information retrieval, and knowledge discovery [126].

Recent studies have further explored the potential of ETopicG and EG in various applications. Loureiro et al. [126] proposed a novel approach for topic modeling using entity-based topics extracted from large language models and graph neural networks. Their work demonstrates how ETopicG can be leveraged to uncover latent structures within a corpus of text, leading to more coherent and interpretable topics. Additionally, the concept of EG has been applied in the field of database management, where it is used to optimize query performance by defining templates for related persistence fields. This approach allows for more efficient data retrieval and improved application performance.

2.1.3 Context application

Context in Machine Learning: Recent advancements in machine learning have introduced innovative approaches to context representation. Yang et al. [205] investigated the training dynamics of transformers in their work. They demonstrated how transformers can learn contextual information to generalize to unseen examples and tasks. Also, Munguia-Galeano et al. [137] proposed a framework called Iota explicit context representation (IECR) for discrete envi-

ronments. This framework uses contextual keyframes to extract functions representing the affordances of states, significantly improving learning efficiency. Recently, Qian and Yin [147] introduced a novel model for semi-supervised node classification. Their model captures both local and global dependencies between nodes using graph convolution networks and an improved transformer model, demonstrating superior performance on various open datasets. In reinforcement learning, context-specific representation abstraction has also been explored. Abdulhai et al. [2] discussed how abstracting state representations based on context can improve the learning efficiency of deep options in reinforcement learning.

Context in Human Computer Interaction: Context representation in Human-Computer Interaction (HCI) plays a pivotal role in enhancing the user experience by adapting interactions based on the user's environment and situation. Makkonen et al. [128] discuss various methodologies and state-of-the-art research in context awareness, highlighting its increasing importance in future HCI applications. Chamberlain et al. [45] provides a comprehensive framework that distinguishes between intrinsic and extrinsic contextual factors, which is crucial for analyzing human-computer communication. An extensive body of work exists on the use of context representation in activity recognition through sensor data. For instance, various studies provide insights into users' activities [54], manage evolving data streams [1], and operate effectively even in out-of-the-lab settings [62], all aiming to recognize the context [190]. Mobile sensing research delves into diverse applications such as the investigation of social interactions [157], the recognition of flu-like symptoms [17], and the detection of eating events [16].

2.2 Big-thick data research

2.2.1 Big data

The concept of big data emerged with the explosive increase in digital records from smart devices. It is characterized by its *massive Volume*, *high Velocity*, and *Variety of data forms*, known as the three *Vs*. AI research has extensively analyzed big data to reveal patterns, particularly relating to human behaviors. For instance, human activities can be predicted using motion or inertial sensors (e.g., accelerometers and gyroscopes) [37, 151]. In-vehicle sensors, in conjunction with roadside infrastructure, such as cameras and sensors, have been employed in intelligent transportation systems [89].

Big data facilitates computational and predictive analytics that far exceed human cognitive capabilities, but it does not undermine the importance of human observation within contexts. Big data is commonly too thin to describe many interesting aspects of people. Recent studies have highlighted the need for integrating big data with human insights to achieve more comprehensive and accurate models. For example, Blank [23] emphasized the limitations of big data in capturing the richness of human experiences and the necessity of contextual understanding. Moreover, integrating big data with machine learning techniques has shown promising results in various domains. Guerrero et al. [89] demonstrated the use of in-vehicle and roadside sensors for intelligent transportation systems, showcasing the potential of big data to improve traffic management and safety. Additionally, Sheikh et al. [169] explored the application of wearable sensors for health monitoring, highlighting the role of big data in personalized healthcare. In summary, while big data offers powerful tools for analysis and prediction, it must be complemented with human observation and contextual understanding to fully capture the complexity of human behaviors and experiences.

2.2.2 Thick data

Differently, thick data is derived from observations of the contexts in which human behaviors occur, such as people's interviews and self-reports. Thick data enables researchers to understand and reflect upon the scenarios where people are [27]. This approach, often referred to as thick description, was first introduced by Gilbert Ryle [161] and further developed by Clifford Geertz [72]. Thick description involves providing detailed narratives and interpretations of observed situations, capturing the cultural, social, and emotional contexts that influence human behavior.

Recent studies have highlighted the value of thick data in complementing big data. For instance, Blank [139] emphasized the limitations of big data in capturing the richness of human experiences and the necessity of contextual understanding. Thick data, collected through ethnographic research methods such as participant observation, in-depth interviews, and focus groups, provides insights into the nuances of human behavior that cannot be captured by big data alone. Recently, Mortati et al. [136] revealed the core role of thick data in the definition of fuzzy problems through case studies of 8 design thinking projects. Moreover, thick data is particularly useful in fields such as user experience design, where understanding the context of user interactions is crucial for creating intuitive and effective interfaces. By combining the quantitative insights from big data with the qualitative richness of thick data, researchers can develop a more comprehensive understanding of human behavior and design better solutions to meet user needs.

2.2.3 Big-thick data

Up to this point, we have known big data and thick data; this section focuses on the integration of these two into big-thick data.

Figure 2.1, which is adapted from [27, 84], illustrates the Cartesian plane on

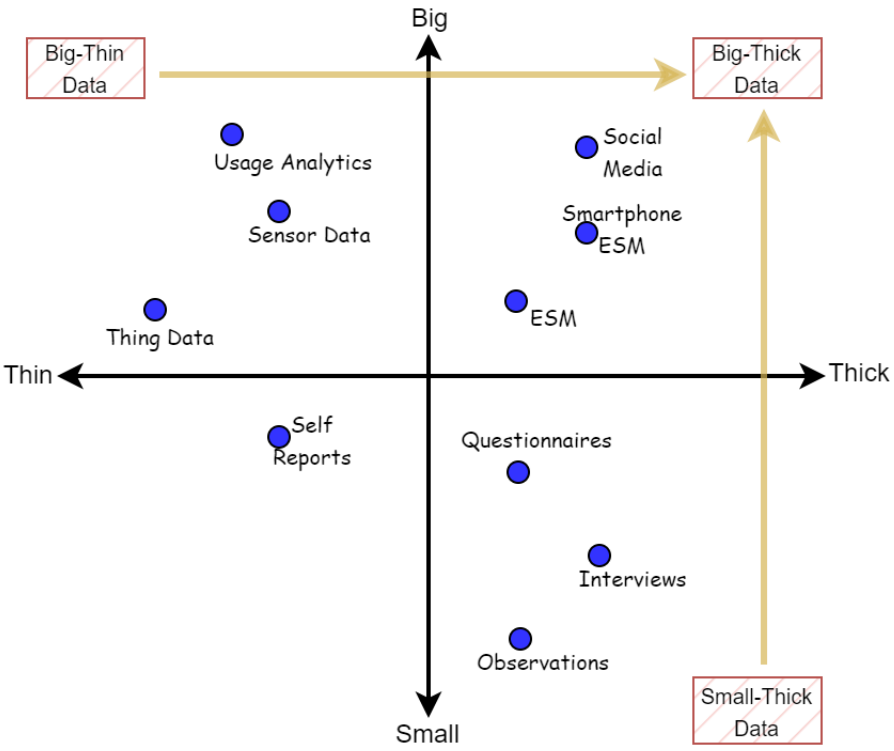


Figure 2.1: Big-thick data explanation.

which various data and data sources pertaining to individuals are plotted. The plane is divided into four quadrants: thin and thick, big and small. Big-thin data sources occupy the top-left corner of the coordinate system, characterized by their vast quantity but limited contextual depth. Examples include datasets generated from sensors, such as GPS location data, which are plentiful but lack the perspective of the individual. In contrast, at the bottom-right corner, small-thick data are collected through traditional social science methods and tools, including participant observation, ethnography, interviews, and questionnaires (arranged from the thickest to the thinnest). However, these data types may lack objective information to corroborate subjective insights.

The top-right quadrant introduces big-thick data, which focuses on integrating ethnographically collected thick observational data. This blending approach merges the extensive coverage of big data with the contextual richness of thick data, providing a more comprehensive and nuanced understanding of human

behavior and interactions.

The integration of big data and thick data into big-thick data represents a comprehensive approach to understanding human behaviors and societal trends. This merging leverages the vast, quantitative insights derived from big data: characterized by its volume, velocity, and variety: with the deep, qualitative insights obtained from thick data, which involves contextual, ethnographic observations and narratives. Bornakke and Due [27] proposed the "Big–Thick Blending" method, which combines analytical insights from both big and thick data sources to provide a richer understanding of human behaviors. Hong et al. [97] demonstrated the value of this approach in urban science and smart city governance, while Smets and Lievens [173] explored human sensemaking in smart cities by integrating big and thick data.

2.2.4 Big-thick data collection platform

Currently there are various research platforms exist that collect and process sensor and human input data from smart devices. For example, Ranjan et al. [152], using wearable devices and smartphone technologies, developed RADAR-base, a platform for mHealth with data aggregation, study management, and real-time visualization capabilities for monitoring major depressive disorder, epilepsy, and multiple sclerosis. Thanos et al. [176] introduced a knowledge-driven framework to enhance the efficacy and efficiency of healthcare through remote and intelligent assessment. It comprises a rule-based strategy to detect health-related problems from wearable lifestyle sensor data, adding clinical value to follow-up and intervention decisions. DemaWare2 [175] is another context-aware fusion system that employs OWL 2 as its knowledge representation and reasoning language. Raw sensor data, audio, and images are analyzed and processed, represented semantically, and interpreted.

Authors in [171] use a combination of sensors from mobile and wearable smart devices to detect bad habits. They use sensor data from wearable devices

and the position of the smartphone to recognize the user's actions and context, and to determine when to provide feedback. Aware [63], built on top of Beiwe [141], is an additional open-source mobile platform for generating user contexts from sensor data from smart devices and human digital questionnaires. Karl et al. [3] track a user's digital footprint by combining virtual and physical sensors from mobile devices to infer digital memories of the user's activities and then constructing an episodic memory by applying semantic reasoning to aggregate these activities.

Within our KnowDive group from the University of Trento, Italy, we have pioneered the development of an advanced application termed i-Log [212]. This innovative tool is engineered to amass substantial, high-dimensional data, and detailed data collection was upon user consent. i-Log harnesses the smartphone's intrinsic sensors to gather data and deploy context-sensitive inquiries to users. The overarching objective is to analyze user habits, enabling adaptive personalized services and generating valuable datasets for subsequent research. Publications [81, 80] elucidate the deployment and data acquisition efficacy of i-Log across diverse studies. Presently, i-Log is compatible with both Android and iOS platforms, albeit with varying sensor data collection capabilities. The unique ability to integrate user responses with sensor data delineates i-Log from existing tools (refer to [193, 159, 112] for a comparative analysis). This dual functionality significantly augments contemporary methodologies in time-diary studies, particularly those that are structured [80, 29].

2.3 The quality of big-thick data

2.3.1 Sensor data quality

Sensor data quality is a critical factor in the performance and reliability of Internet of Things (IoT) applications, Human-Computer Interaction (HCI), and AI. High-quality sensor data ensures accurate and reliable decision-making,

whereas poor data quality can lead to incorrect conclusions and potentially harmful outcomes. Recent advancements in sensor data quality research have focused on identifying, quantifying, and correcting various types of sensor data errors, such as missing data, outliers, bias, drift, and uncertainty.

One of the key contributions to the IoT field is the systematic review by Teh et al. [185], which provides a comprehensive overview of the different types of physical sensor data errors and the methods used to detect and correct them. The review highlights the use of Principal Component Analysis (PCA) and Artificial Neural Networks (ANN) as common techniques for error detection and correction. Another significant work is the paper by Zhang et al. [213], which surveys data quality frameworks and methodologies for IoT data. The paper compares various data quality definitions, dimensions, metrics, and assessment techniques, aiming to help practitioners choose the appropriate method for their specific IoT system. The authors also emphasize the importance of international standards in ensuring consistent and reliable data quality management. Shehu Yalli et al. [204] discuss the elements necessary to ensure quality of data, including outlier detection, data-cleaning techniques, and quality assurance methods. Their comprehensive taxonomy provides a structured approach to maintaining high data quality in IoT systems.

Recent research in HCI has focused on improving sensor data quality to enhance user experience and interaction. For instance, Rechy-Ramirez et al. [154] reviewed the impact of commercial sensors in HCI, highlighting the importance of sensor specifications and their applications in creating human-computer interfaces for complex tasks. Another significant contribution explores novel sensing and interfacing technologies that encourage new ways of interaction between humans and computers. It covers intelligent sensors for HCI, such as speech, gesture, and gaze recognition, and adaptive HCI that deals with cognitive and affective levels of user activity [8].

2.3.2 Human input data quality

Mobile self-reports have emerged as a prevalent technique for in-situ data collection. This approach, predicated on its capacity to document participants' behaviors directly, enables the establishment of ground-truth data, where the data's intrinsic meaning is ascribed by the user. A significant challenge, however, lies in the inability to ascertain the precise causes of errors, primarily due to the difficulty in observing respondents' behaviors in natural settings (e.g., identifying specific causes and conditions) [200]. Another complication arises from researchers' limited control over the surrounding environment. The user remains unsupervised by a human interviewer who could dictate the timing and context of responses. Consequently, the accuracy of self-reported data has been insufficiently scrutinized in the literature, with a prevalent but erroneous assumption that responses are inherently accurate [192].

In the domain of human input data collection, the Experience Sampling Method (ESM) [118] and Ecological Momentary Assessment (EMA) [170] are extensively used to dispatch inquiries to participants. These methodologies enable researchers to design intensive, repeated measure questionnaires that yield comprehensive contextual information about respondents. The ongoing evolution of smartphones has culminated in the advent of smartphone-based ESM [133], a technology that facilitates the documentation of human behaviors and experiences through real-time user queries and time-use analysis. Participants can be queried about various behaviors, daily events, and moods via their smartphones [43].

The issue of low-quality responses has been a focal point of EMA/ESM research [170, 118], with findings subsequently integrated into more complex research protocols [118, 50, 91]. Nonetheless, existing literature has predominantly concentrated on enhancing the response rate of participants. For instance, Boukhechba et al. [28] reported higher response rates when ESM ques-

tions were sent post-phone calls compared to social media usage. Similarly, Berkel et al. [191] observed that active phone usage preceding an ESM query could elevate response rates. Contextual influences on response rates have also been examined, with studies investigating the impact of location [184], activities [135], various times of the day [146], and social interactions [28]. Focusing on response accuracy, Berkel et al. [192] found that the highest accuracy levels were achieved when participants' screens were off upon notification arrival, indicating non-usage of the phone. They also noted that prolonged question completion times corresponded with lower accuracy, with the primary limitation of their work being the exclusive consideration of smartphone usage as the contextual factor.

2.3.3 Total survey error paradigm

The Total Survey Error paradigm, as articulated in Sociology, offers a theoretical framework for enhancing survey accuracy while optimizing resource allocation [168, 9]. Within this framework, four principal sources of error, though independent, correlate with the studied phenomena. These sources are delineated as follows:

1. The *situational and temporal context* in which the respondent inputs information into the smartphone [120, 68, 200].

2. The *cognitive task* intrinsic to the response process [199], particularly in temporal inquiries within multi-component approaches [127] and respondent motivation theories [39, 113, 114].

3. The *conscious or unconscious factors*, such as personality, attitudes, and habits, that influence respondent behavior [127, 153].

4. The *technical problems* pertaining to the functionality of the technology, including the smartphone and its applications [86, 164, 64, 162, 15, 181].

These four sources are posited to engender errors in any question-answering process, thereby extending to mobility self-reports. This hypothesis constitutes the primary theoretical underpinning of the work presented in this thesis. To our knowledge, this hypothesis has not been previously applied to mobility self-reports, particularly in the context of contextual inquiries.

2.3.4 Measure answer quality

In the realm of Artificial Intelligence, there has been substantial research focused on understanding the quality of users' responses. However, current investigations are primarily concerned with reaction times and predicting the likelihood of a user timely responding to a notification. For instance, Iqbal et al. [104] demonstrated that users are more receptive to interruptions from messages that convey valuable information. Similarly, Fischer et al. [66] found that the content of a message is more significant than the timing of the notification, indicating that participants prefer interruptions from content that piques their interest. Furthermore, these authors [65] explored the dependency of response times on prior interactions, such as phone calls or SMS messages. Ho and Intille [140, 94] developed an application predicated on the notion that transitions between physical activities (e.g., from sitting to standing) present opportune moments for notifications, including phone calls, messages, and reminders. In another study, Mehrotra et al. [132] advocated for the use of context information and application usage patterns to forecast optimal notification times. Additionally, Aminikhanghahi et al. [10] integrated Experience Sampling Method (ESM) technology with knowledge of user activities to enhance the prediction of notification timing.

Chapter 3

Datasets

3.1 Introduction

Daily behavior and mood are crucial for university students' academic performance. Obtaining a university degree can significantly improve a student's job prospects [12], while student drop-out represents a loss of economic and social resources for society [119]. Secondly, mood is a crucial element that impacts mental health, affecting individuals across a range of contexts, for instance, academic workload [144]. Poor mood can lead to severe consequences, see, e.g., substance abuse and suicidal thoughts [71].

University students' daily activities and behaviors, which involve balancing academic activities and other daily tasks, play a vital role in their academic success. Traditionally, sociological researchers use surveys to understand students' daily activities by asking them to report the duration and the time they spend. However, these traditional tools often suffer from data quality issues due to the reliability of participants' responses.

To address the above problem, we conducted the Smart Unitn Two and WeNet experiments in 2018 and 2021, respectively, to collect time-use data from students at the University of Trento. The primary objective of these experiments was to understand how students manage their time and how it impacts their academic performance. Instead of relying on traditional surveys, we devel-

oped the i-Log app [1], which was installed on students' mobile phones to collect their responses and sensor data embedded in the devices. In this chapter, we will mainly simply introduce the Smart Unitn Two and WeNet experiments and the datasets generated by these experiments, because the work of this thesis is not about generating the datasets, but rather to use them to evaluate our system and algorithms in subsequent work.

3.2 The i-Log

According to the final goals set for the Smart Unitn and WeNet projects, we define the following two types of data that we are going to collect via an smartphone application named i-Log [212], which can be installed by the student's smartphone:

- **Objective Data:** This includes sensor data collected from participants' mobile phones, which collects sensor data from multiple sensors on smartphones simultaneously, from hardware sensors (e.g., Location, Accelerometer, Gyroscope, etc.) to software sensors (e.g., Airplane Mode, Bluetooth Devices, etc.) and even the Question-answer sensors (e.g., Time Diary Question, Task Question, etc.). This type of data provides an accurate and unbiased record of the participants' movements and activities.

- **Subjective Data:** This includes responses to context questions collected through the i-Log app. Which also pushes time diaries to the app users. Each time diary consists of questions about the activities, locations, and social relations of students. These time diaries are sent every 30 minutes, and the student users' answers provide details on how they allocated their time during the day. Each time diary can be answered within 150 minutes, with a maximum of 5-time diaries stacked in a queue. These responses

[1] https://datascientia.disi.unitn.it/ilog

offer insights into participants' personal experiences, moods, and motivations, adding depth and context to the objective data.

By leveraging both hardware and software sensors, along with strategically timed and unobtrusive time diaries, i-Log ensures comprehensive and high-quality data collection without disrupting the students' daily routines.

3.3 Smart Unitn 2

Based on the i-Log app, we have conducted multiple experiments; the first experiment we introduced was named SU2 (*Smart University Two*), which involved gathering data from students at the University of Trento, Italy. This experiment received preliminary approval from the Ethical Committee and GDPR Committee of the University of Trento. The project spanned four weeks (28 days), from early May to early June 2018. The experimental protocols, participant recruitment strategies, and data cleaning methods of this dataset have been comprehensively described in [76]. In addition, the full description of the collection of the SU2 dataset includes the design and execution of this experiment. Also include the GDPR compliant, which can be found at: https://datascientia.disi.unitn.it/projects/su2. This link also provides a way to download the dataset.

At the outset, a short questionnaire was sent to a random sub-sample of 10,006 students, inquiring whether they regularly attended classes and possessed an Android smartphone. This sample was drawn from the entire student population of the University of Trento. Those who completed the questionnaire and responded positively to both questions were then invited to participate in the survey. The invitation detailed the study's aims, explaining that participants could choose to engage for either two or four weeks. During the first two weeks, participants would receive notifications every half-hour, and during the second

two weeks, notifications would be sent every two hours. Additionally, students were told which types of sensor data (see Figure 3.1) would be collected.

Sensor	Frequency	Sensor	Frequency	Sensor	Frequency
Acceleration	20Hz	Screen Status	On change	Proximity	On change
Linear Acceleration	20Hz	Flight Mode	On change	Incoming Calls	On change
Gyroscope	20Hz	Audio Mode	On change	Outgoing Calls	On change
Gravity	20Hz	Battery Charge	On change	Incoming Sms	On change
Rotation Vector	20Hz	Battery Level	On change	Outgoing Sms	On change
Magnetic Field	20Hz	Doze Modality	On change	Notifications	On change
Orientation	20Hz	Headset plugged in	On change	Bluetooth Device Available	Once every minute
Temperature	20Hz	Music Playback	On change	Bluetooth Device Available (Low Energy)	Once every minute
Atmospheric Pressure	20Hz	WIFI Networks Available	Once every minute	Running Application	Once every 5 seconds
Humidity	20Hz	WIFI Network Connected to	On change	Location	Once every minute

Figure 3.1: Hardware and software sensor data collected together with their sampling rate. “On change” means that the value of the sensor is collected only when it changes its value.

As noted in the literature [109, 10], willingness to participate in mobile data collection is strongly influenced by the promised incentives. A reward of 20 euros was offered for each of the first two weeks of participation. Participants were also informed that, upon survey completion, a lottery would be held among those who responded to more than 75% of the notifications. The lottery would award three prizes of 100 euros for the first two weeks and three prizes of 150 euros for the second two weeks. The invitation included a second questionnaire seeking additional information about participants’ general characteristics, university experience, and profiling of their procrastination syndrome. At this stage, personal email addresses and GDPR-compliant consent forms were also collected.

The recruitment process yielded 1,042 applicants, a response rate consistent with other web surveys without reminders. From this group, students over the age of 25 were excluded to limit the sample to those with regular academic careers. A stratified sample of 318 students, proportional to the student population of each department, was drawn from the remaining 860 applicants. These se-

lected students received a second questionnaire to explore their university life, habits, and routines. Students who completed the questionnaire and signed a second informed consent form were provided a password to install i-Log. In total, 275 students completed the questionnaire and installed i-Log.

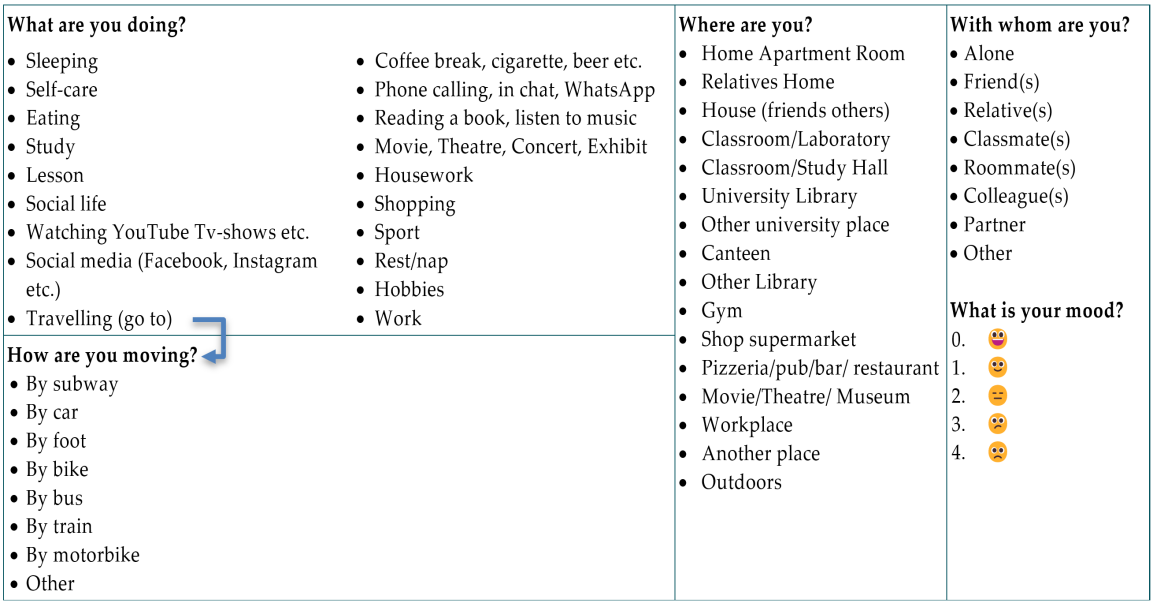


Figure 3.2: Time diary.

The data collection was divided into two stages. In the first stage, over 14 days, students received notifications with questions (see Figure 3.2) every half-hour. Participants had 720 minutes (12 hours) to respond to each notification, after which the question would be dropped and marked as missing. To reduce the burden, users could pause data collection for 6 hours while sleeping. Of the 275 students, 237 responded to the notifications at least once. During the initial days, 25 students withdrew, citing technical issues or loss of interest. By the end of the first two weeks, 184 students had provided valid responses for at least 13 out of 14 days.

In the second stage, which also spanned 14 days, 202 of the initial 237 students expressed willingness to continue. During this period, notifications were sent every hour. Of these students, 113 completed the task for at least 12 days

and provided over 100 valid answers. This dataset was used in [214] to predict individual behaviors and in [82] to examine the impact of social media usage on students' academic performance.

3.4 WeNet

The second experiment is also collected from the University of Trento, Italy. But from the WeNet EC project, also known as *WeNet - The Internet of us*, see https://www.internetofus.eu/ for a detailed description of the project plus the possibility of downloading the dataset. The detailed experiment is described in [75]. To begin with, all students at the University of Trento were invited via email to participate in a survey. From the 5,000+ respondents, a representative sample of 370 students was selected based on their fields of study and socio-demographic characteristics. To mitigate bias, we focused on the data of 170 students. This selection was based on considerations related to the number of answers provided and demographic characteristics, including gender, study degree, and department (see Table 3.1). 170 participants during this period had a total size of 110 gigabytes. It contained 20 billion sensor readings together with answers to time diary questions, which were deemed annotations to the sensor data. The total number of answers generated during the experiment was 155,040 out of 290,016 sent. Detailed information on this experiment can be found in [75].

Table 3.1: Selected participants' demographic information.

	Sex		**Degree**		**Department**			
Feature	Female	Male	BSc	MA+PhD	Information Science	Industrial	Business	Sociology
Number	101	69	108	62	50	25	44	51
Percent	59.41%	40.59%	63.53%	36.47%	29.41%	14.71%	25.88%	30.00%

The first part of the closed-ended questionnaires was administered using the

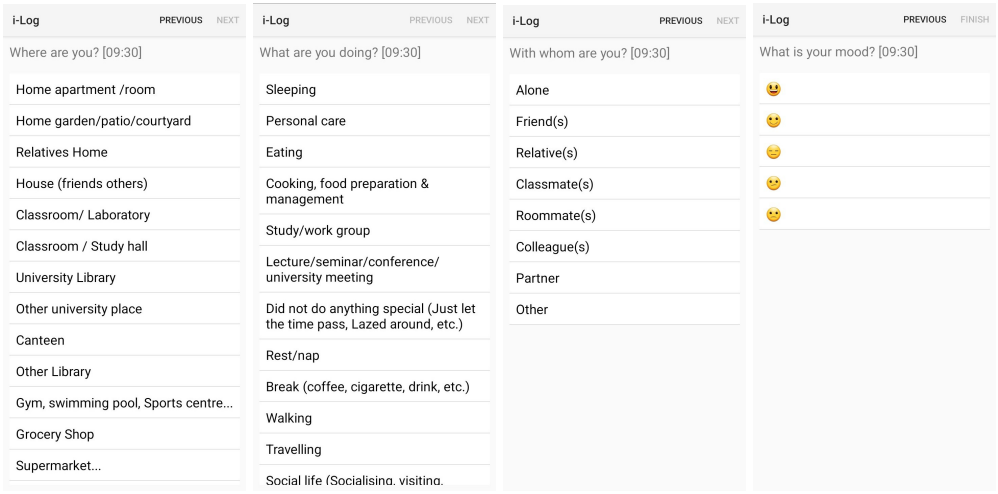
Limesurvey [2] platform. An invitation to participate in our project was emailed to all enrolled students at the Italian university. From 5000+ applicants, we randomly selected 370 students, with the sample weighted to reflect the student population of each department. The average age of these students was 24.1. All 370 students installed the i-Log [212] at the commencement of the survey. Other tools for this purpose are available, e.g., [112, 193]. Furthermore, the data collection process adhered to a protocol compliant with the EU General Data Protection Regulation (GDPR).

Secondly, we obtained participant self-reports via their smartphones through Time Use Diaries [29], which are studies often based on the Harmonised European Time Use Survey (HETUS) [3] standard. In our diary study, the data collection spanned a period of four weeks. During the first two weeks, participants received questions every half hour and were given 12 hours to respond. In the following two weeks, the frequency of questions was reduced to once per hour, but participants still had a 12-hour window to respond. In total, we collected more than 110,000 self-reports from all participants. Our time diary included special sections on the main activities. Participants received a notification on their smartphone with four questions, the first three of them were based on the HETUS standard. Figure 3.3 shows some of the questions (and pre-compiled answers) asked during the data collection.

- Their activity "What are you doing?" offering the participant 34 answer categories such as eating, driving, studying, etc.;
- The current location "Where are you?" offering the participant 26 categories such as home, classroom, canteen, restaurant, etc.;
- The individuals accompanying the participants at the time of the question "Who is with you?" offering the participant 8 categories such as alone;

[2] https://www.limesurvey.org/

[3] https://ec.europa.eu/eurostat/web/time-use-surveys



(a) (b) (c) (d)

Figure 3.3: Sample questions captured in the WeNet project

partner, friends, etc; and

- The mood "What is your mood?" offering the participant a scale of 5 levels ranging from happy to sad.

Thirdly, we collected data from a pre-selected list of sensors in the background without user intervention. The data was generated as time series, consisting of tuples composed of a timestamp and one or more values. In this thesis, we only consider the time component of the data.

As indicated in the literature [109], to foster participation, participants were remunerated with 20 euros for every two weeks of participation. In addition, two economic incentives were introduced to promote better results. The first comprised five daily awards of 5 euros each, bestowed upon those who most effectively adhered to daily reporting. The second involved six final prizes, three of 100 euros for the initial two weeks and three of 150 euros for the subsequent two weeks, granted to six randomly selected participants who best complied with the entire research protocol and responded to more than 75% of notifications.

Chapter 4

Answer Quality Assurance

4.1 Introduction

To learn context-awareness information from participants, we decided to ask them questions; however, the quality of human input data often does not meet the requirements of current AI. Additionally, there is no direct method to assess the quality of participant feedback. Thus, the goal of this Chapter is to conduct an in-depth study of the factors influencing the quality of answers provided by users in natural settings. By quality, we refer to a low incidence of errors in the answers.

The primary challenge is to identify key factors that affect answer quality. These factors must be operationalizable as part of a strategy that enables a machine to derive the best possible results from user responses. The starting point for this analysis is the *Total Survey Error paradigm* discussed in Section 2.3.3.

The *Total Survey Error paradigm*, as elaborated in Sociology, provides a theoretical framework for optimizing surveys by maximizing data quality within budgetary constraints [168, 9]. According to this framework, there are four main sources of error that are independent of one another while correlating with the phenomena under study. These factors can be summarized as follows:

1. The *situational and temporal context* in which the user inputs information

into the smartphone [120, 68, 200].

2. The *cognitive task* involved in the answer process [199], especially in time-related questions within multi-component approaches [127] and respondent motivation theories [39, 113, 114].

3. The *conscious or unconscious factors*, such as personality, attitudes, and habits, that influence the user's behavior [127, 153].

4. The *technical problems* related to the functioning of the technology, e.g., the phone and phone applications [86, 164, 64, 162, 15, 181].

These four sources are assumed to generate errors in any question-answering process and, therefore, also in mobility self-reports. This hypothesis serves as the primary theoretical foundation for the work presented in this Chapter. To our knowledge, this hypothesis has not previously been applied to mobility self-reports, particularly in the context of inquiries about surroundings. In this regard, the work in this section aims to validate and elucidate the specific factors that instantiate the hypotheses of the *Total Survey Error paradigm* within this scenario.

4.2 Answer correctness chain

Based on the above work from the *Total Survey Error paradigm* in Section 2.3.3, we assume that the existence of a causal chain of events influences the quality of answers and that we abstractly model according to the schema sketched in Figure 4.1.
Proceeding from the beginning left to the ending right, we distinguish between *endogenous* and *exogenous* factors, which play a significant role and impact the quality of answers. These two sources of error can be detailed as follows:

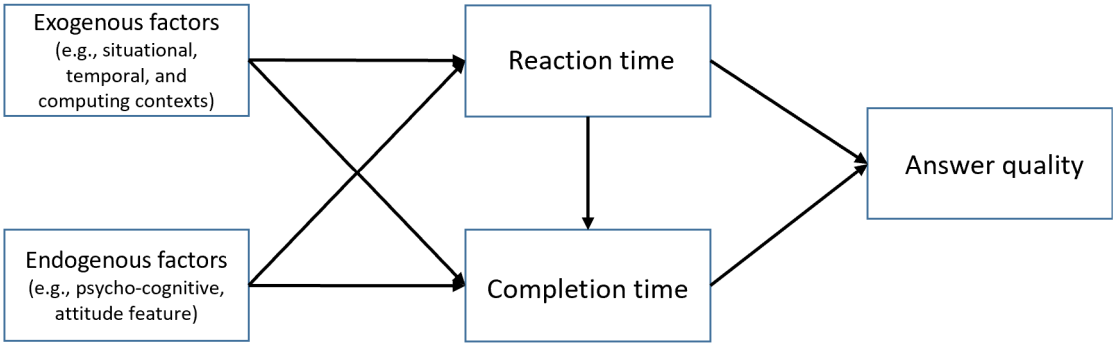


Figure 4.1: The causal chain of the factors impacting the answer quality.

- *Endogenous factors* are organized along two main dimensions. The first dimension is the user's willingness and ability to follow and complete a task on a device [88, 127, 142]. This factor encompasses a range of psycho-physical attitudes, personality traits, behaviors, and habits—both conscious and unconscious—that influence the individual [127, 153]. Cognitive performance varies among individuals [48] and is also time-variant. Note that mental state [18, 47], as well as certain exogenous factors, can influence cognitive performance. Examples of such exogenous factors include the time of day [166, 201] and smartphone usage [101, 115]. And the second is the user's characteristics and behavior, such as familiarity or comfort with the device, and how the respondent uses the device (i.e., frequency of use, duration, and frequency of browsing sessions) [123, 88, 116, 150, 127, 153, 105].

- *Exogenous factors*, primarily related to the context of use, are organized along three dimensions: the *physical* and *social situational context*, the *temporal context*, and the *computing context*. The first two dimensions pertain to factors such as the degree and type of distractions, the presence of others, or multitasking behaviors—whether on the same device, a different device, or a different medium [120, 121, 127, 123, 153, 200]. The third dimension, the computing context, relates to technical issues associ-

ated with the functioning of i-Log, accounting for all possible malfunctions [93].

We structure our analysis around two critical factors that influence answer quality: reaction time and completion time, which aligns with the study in [192]. These factors enable us to model the chain of effects of how exogenous and endogenous variables affect the quality of responses. The factors are defined as follows:

Reaction time: The purpose of analyzing reaction time is to identify *which endogenous and exogenous factors affect reaction time*, where reaction time refers to the interval between receiving a notification and starting a response. We formalize *reaction time*, denoted as t_r, as:

$$t_r = t_a - t_n \tag{4.1}$$

where t_a is *answer time*, which is the time when a participant starts to answer the questions. The t_n is *notification time*, which is the time when a participant receives the questions that are sent by machines. If the participant did not answer a question (e.g., $t_a = `null'$), we set $t_r = \infty$. In real-world datasets, the notification time and answer time are always recorded by *timestamp* datatype, which stores the year, month, day, hour, minute and second values, e.g., '2024-07-01 10:11:20' in MySQL Database. To facilitate the representation and application of reaction time, we invariably set the unit of reaction time to minutes, e.g., '28' minutes.

Completion time: The purpose of analyzing completion time is to identify *which endogenous and exogenous factors affect completion time*, where completion time refers to the duration taken to complete a response, starting from the end of the reaction time. We formalize *complete time*, denoted as t_c, as:

$$t_c = t_f - t_a \tag{4.2}$$

where t_f is *finish time*, which is the time when a participant finish to answer the questions. The finish time and is always recorded by *timestamp* datatype, which

stores the year, month, day, hour, minute and second values, e.g., '2024-07-01 10:13:46' in MySQL Database. To facilitate the representation and application of reaction time, we invariably set the unit of reaction time to minutes, e.g., '1.2' minutes.

Answer quality: The goal of analyzing answer quality is to validate the causal chain, as depicted in Figure 4.1, by quantitatively evaluating *how reaction and completion times jointly impact answer quality*. Here, by answer quality, we mean the number of questions that get correct answers. To quantify answer quality, we binary classify answer quality as high-quality or low-quality responses.

$$Answer \quad Quality = \begin{cases} High, & \text{if } t_r < RT \\ Low, & \text{if } t_r \geq RT \end{cases} \tag{4.3}$$

where RT is a cut-off point as the time threshold that distinguishes answer quality. If the reaction time of an answer t_r is less than the reaction time cut-off point RT, the quality of the answer is considered 'High'. On the other hand, if the reaction time t_r is greater or equal to RT, the quality of the answer is considered 'Low'. Of course, if the participant did not answer a question, we have $t_r = \infty$. Hence, this answer is considered a low-quality answer.

In the analysis of reaction time and completion time, we employed a multilevel discrete-time event history model (MLM) [179, 186] and a Cox regression model [7, 179, 49]. Firstly, the MLM is a generalized linear model wherein repeated notifications over time (level 1) are nested within users (level 2) [83, 172], modeling variability across upper-level units using random effects. MLMs also estimate parameter attributes at multiple distinct levels. A common example of an MLM is a school setting, where a student's competence in a class is influenced partly by the student (first level, e.g., their ability) and partly by the class (second level, e.g., the attitude of the teachers).

Secondly, the Cox regression model, also known as the Cox Proportional Hazards model, is a statistical technique used primarily in survival analysis

to explore the relationship between the time until an event occurs and one or more predictor variables. Developed by Sir David Cox in 1972, this model does not require the specification of a baseline hazard function, making it a semi-parametric method. The primary assumption of the Cox model is the proportional hazards assumption, which states that the hazard ratios between different levels of the predictor variables are constant over time. This model estimates the hazard function for an individual, given their covariates, by combining a baseline hazard function with the exponential effect of the covariates. This approach allows for the estimation of the impact of covariates on the hazard or risk of the event occurring, while leaving the baseline hazard function unspecified. One of the strengths of the Cox regression model is its ability to handle censored data, which is common in survival analysis when the exact time of the event is not known for all subjects. By incorporating both right-censored and left-censored data, the model provides a robust framework for analyzing survival times and identifying significant predictors of the event of interest.

For assessing answer quality, we used both a Multilevel Structural Equation path analysis Model (MSEM) [95, 99, 107] and a standard SEM path analysis. These models integrate evidence from the initial analyses into a cohesive chain of causal events. In this approach, reaction and completion times are influenced by user characteristics, while response errors are modeled as random variables. Consequently, an MSEM was chosen, formalizing reaction and completion times as a two-level model (responses nested within respondents).

4.3 The reaction time

The reaction time refers to the interval between a user receiving a questionnaire and beginning to provide an answer. Ideally, users should respond immediately upon receiving the notification to minimize the risk of memory errors (primarily forgetting), which increases with time and leads to higher chances of incorrect

answers [131]. In this context, we focus on three factors that may influence reaction time:

- *Context history* [40] (an exogenous factor): This includes a blend of user context (social situation), physical context, computing context (network connectivity, etc.), and time context. The *social and physical context* are addressed through self-reported time diary entries: (a) "**What are you doing?**" captures potential distractions from current activities; (b) "**Where are you?**" identifies environmental factors and associated distractions; (c) "**With whom are you?**" notes social distractions. The specifics of these questions and possible responses are detailed in Figure 3.2. The question "**How are you moving?**" is asked whenever the user indicates they are traveling. The *computing context* is considered by measuring the **time elapsed**, in seconds, from when the server sends the notification to when the smartphone receives it. Lastly, the *time context* differentiates between weekdays and weekends to account for varying weekly activities.

- *Motivation and burden* (an endogenous factor): This pertains to the effort (time and resources) and difficulty level. In the model, it is represented as a quadratic function of the **day of study**, ranging from 0 (first day) to 13 (fourteenth day).

- *User characteristics* (an endogenous factor): This factor encompasses psychological traits and daily emotional status. Mood states and personality traits significantly affect the processing of emotion-congruent information in various cognitive tasks [160]. This aspect is evaluated by asking about the user's **procrastination** syndrome and **emotional mood state**. The procrastination question uses the *Irrational Procrastination Scale (IPS)* [177, 178]. The emotional mood state is assessed on a 5-point Likert scale, with options ranging from happy (0) to sad (4) [125, 110], as shown in Figure 3.2.

Table 4.1: Cox Regression model and Multilevel discrete time model with random intercept on the reaction time.

Features	Cox-regression Coef B	Multilevel Coef B
(Spatial context): Where are you?		
Home/Room	Reference	Reference
Relatives Home	-0.061***	-0.100***
House(friends/others)	-0.021	-0.029
University	0.188***	0.277***
Shop/Pub/Theatre	0.028	0.013
Work place	-0.234***	-0.331***
Other place	-0.071***	-0.128***
Outdoors	-0.218***	-0.324***
(Event context): What are you doing?		
Personal care	Reference	Reference
Eating	-0.141***	-0.193***
Study alone or with others	-0.174***	-0.277***
Class room lecture	-0.130***	-0.175***
Social life/Break	-0.137***	-0.167***
Watching YouTube TV, etc.	-0.080***	-0.126***
Social media, Phone call, chat	0.214***	0.311***
Free time	-0.298***	-0.406***
Work	-0.195***	-0.271***
Travel	0.007	0.043
(Social Context): Who are you with?		
Alone	Reference	Reference
Friend(s)	-0.188***	-0.292***
Relative(s)	-0.046***	-0.079***
Classmate(s)	-0.222***	-0.329***
Roommate(s)	0.025	0.018
Colleague(s)	-0.238***	-0.396***
Partner	-0.291***	-0.432***
Other	-0.575***	-0.834***
(Time Context): When questions sent		
Sunday	Reference	Reference
Monday	0.089***	0.122***
Tuesday	0.115***	0.154***
Wednesday	0.068***	0.076***
Thursday	0.131***	0.172***
Friday	0.032**	0.049**
Saturday	-0.022	-0.032
Study day	-0.076***	-0.123***
Study day^2	0.003***	0.005***
Question delivery delay time (sec.)	0.0002***	0.0003***
User Characteristics		
Mood	0.060***	0.092***
Procrastination	-0.010**	-0.021**
var(cons[user])		0.643***
var(cons[user>notid])		0.604***
Observations	58340	2,565,159
Number of groups	158	158

To evaluate reaction time in a real-world experiment, we present the results in Table 4.1. In this table "*" means p<0.1, namely, relatively weak (statistical) significance; "**" is p<0.05, a strong significance; "***" is p<0.01, a very strong significance, and p means "P-value", which providing a measure of the statistical significance of that feature compared to the reference feature. The median survival time for reaction time is 20 minutes, indicating that fifty percent of notifications receive a response within this time frame. As hypothesized, reaction time is influenced by both historical context and user characteristics. For instance, the median reaction time varies from a minimum of 10 minutes for users engaged in social media/phone/chat activities to a maximum of 27 minutes for those involved in leisure activities (Log-rank test. $\chi^2(9) = 1046.69$, p<0.05). In a social context, median reaction time ranges from 16 minutes when the user is alone to 33 minutes when in a social setting (Log-rank test. $\chi^2(7) = 635.38$, p<0.05). Regarding location, the median reaction time varies from 15 minutes at the university to 32 minutes when outdoors (Log-rank test. $\chi^2(8) = 708.50$, p<0.05).

To assess the net effects of all the factors influencing reaction time, we conducted two regression models (Table 4.1): a Cox regression model with user notification shared probability and a multilevel discrete-time model. For both models, we performed a likelihood ratio test against the null model [25]. Results indicate that both models are statistically significant ($\chi^2(34) = 3437.34$, p<0.05; $\chi^2(47) = 5819.65$, p<0.05) and capture the same interpretation of the probability of reacting to a notification at time t. While parameter estimates differ slightly, the sign and interpretation are consistent. Additionally, the Cox regression model estimates that it explains 5.0% of the variance in accuracy ($R^2_D = 0.0462$) [158], indicating that the independent variables listed in Table 4.1 account for 5.0% of the variability in reaction time.

4.4 The completion time

Completion time refers to the interval from when a user begins answering a questionnaire to when they complete the response process. As noted earlier, delayed responses can cause memory errors, affecting both reaction and completion times. Specifically, the cognitive processes involved in response formation can result in extended completion times, potentially impacting response accuracy [131].

Additionally, according to multi-component approaches [188], a respondent's ability to concentrate on specific tasks while ignoring other stimuli [108] influences answer quality. Generally, shorter completion times are expected to yield higher-quality answers by reducing memory error risks [183]. However, this is not always the case. Rapid completion can increase the likelihood of inaccuracies or typing errors [38, 131]. Furthermore, as discussed in [187, 188], the answer completion process involves four steps:

1. comprehending the question,
2. retrieving relevant information from memory,
3. making the judgment required by the question, and
4. selecting an answer.

Extended completion times might indicate more time spent retrieving information from memory, positively affecting answer quality. While the negative impact of increased reaction time is well-established, the same cannot be conclusively stated for completion time. In the following, we consider the following four factors as the possible sources of changes in the completion time:

- *Competence in the usage of i-Log*: This competence improves over time, resulting in a swift reduction of completion time, even in the early stages.

We model this learning process using two components. The first component is the *date/time of use*, measured as the **day of study**. The second component is the *memory of response alternatives* as they appear in the list of answers. The more frequently a response option is used, the more likely a respondent will recall its exact location, leading to faster responses. We model the **list of activities** in the time diary as a pseudo-continuous variable, with each activity associated with its percentage of appearance. For example, "studying" is represented with an occurrence percentage of 19.04%, "eating" with 4.61%, and so forth [41].

- *Reaction time*: Unanswered notifications may accumulate, forming blocks that the user can quickly address in a single response session. This factor is modeled as the **total number of pending notifications** at the start of the completion process.

- *Social context*: This factor accounts for the disturbance or multitasking effects caused by the presence of other people during the completion process [108]. We model this by noting whether the respondent is **alone** when responding (see Figure 3.2).

- *Psycho-social aspects*: Literature suggests that both **mood** and **procrastination syndrome** significantly impact memory and motivation [70, 177], and consequently, the completion process.

We also examined completion time in a real-world experiment, with results detailed in Table 4.2. The median survival time for completion time is 9 seconds, meaning that fifty percent of all notifications are completed in 9 seconds or less. As expected, completion time is influenced by both historical context and user characteristics. For instance, the median completion time ranges from a minimum of 7 seconds when the user is engaged in studying (alone or with others) or attending a lesson (classroom lecture) to a maximum of 11 seconds

Table 4.2: Cox regression model and Multilevel model with random intercept on the completion time.

Features	Cox-regression Coef B	Multilevel Coef B
Event context: Activity	0.038***	0.104***
Social context: Alone vs not alone	0.267***	0.637***
User Characteristics: Mood	0.039***	0.095***
User Characteristics: Procrastination	-0.006*	-0.015*
Study time	0.048***	0.131***
Study time^2	-0.002***	-0.005***
Reaction time (min.)	-0.0002***	-0.001***
Pending notification (count)	0.044***	0.093***
Theta/constant	0.0731***	-14.704***
Observations	58,340	2,565,159
Number of groups	158	158

during leisure activities (Log-rank test. $\chi^2(9) = 5478.37$, p<0.05). In a social context, the median completion time varies from 8 seconds when the user is alone to 10 seconds when with a partner (Log-rank test. $\chi^2(7) = 1682.50$, p<0.05). Regarding location, the median completion time spans from 7 seconds at the university to 12 seconds when at a shop, pub, or theater (Log-rank test. $\chi^2(8) = 2072.78$, p<0.05).

To disentangle the effects of different contexts on completion time, we also ran a Cox regression model and a multilevel discrete-time model (Table 4.2). The likelihood ratio test against the null model indicates that both models are statistically significant ($\chi^2(8) = 3683.55$, p<0.05; $\chi^2(19) = 8457.39$, p<0.05) and convey the same interpretation. Similar to the results for reaction time, the parameter estimates differ only slightly, but the sign remains consistent. Additionally, the Cox regression model estimates that 4.1% of the variance in completion time is explained by the model ($R^2_D = 0.0407$).

4.5 Evaluation

To test the chain of events leading to errors, we continue to use the Smart Unitn 2 dataset, which was introduced in Section 3.3. We use the GPS-detected position of the smartphone when the user claims to be "at home." As shown in Figure 3.2, "at home" is one of the possible answers to the question "**Where are you?**," while GPS data is collected every minute, as detailed in the list of sensor data in Figure 3.1 (with GPS labeled as "location"). We selected the variable "at home" for three reasons. First, we know the GPS position of the respondent's home in the SU2 dataset. Second, a home typically covers a very small area, unlike other locations such as "at the university." Third, this type of question is virtually error-free since it requires no cognitive effort or reasoning and is perceptually and habitually straightforward. Being at home is arguably the most familiar situation for everyone.

The variables used here are the same four used for completion time in the previous subsection, namely competence with i-Log, reaction time, social context, mood, and procrastination, plus the accuracy of the GPS (considered as an additional effect of the computing context). The resulting number of incorrect answers was substantial, thus justifying the research hypothesis that not all user-provided answers are correct, despite a minor number of local mistakes. Specifically, we observed 730 errors out of a total of 7624 questions, corresponding to an overall error rate of 9.6%.

Finally, we test the quality of answers in a real-world experiment to determine whether reaction time and completion time impact answer quality during the answering process.

Figure 4.2 presents an initial exploratory analysis of the effects of reaction and completion times on answer quality. This analysis indicates that both reaction time (Figure 4.2) and completion time (Figure 4.3) have a statistically significant negative impact on response quality. While the negative impact of

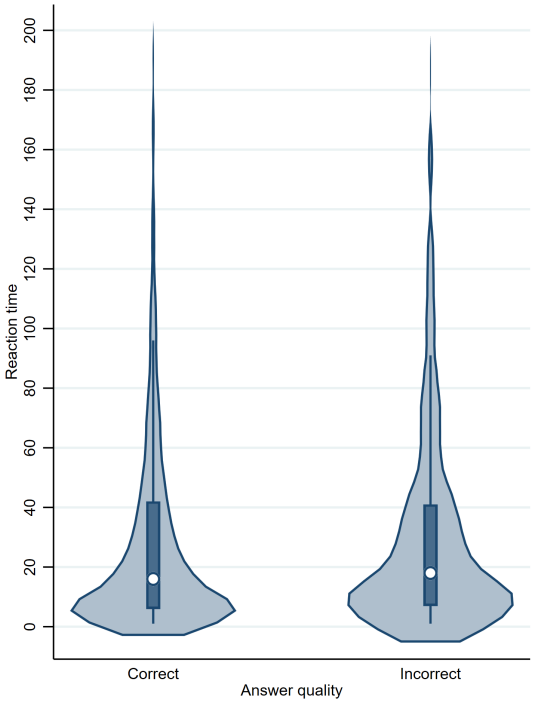


Figure 4.2: How does the reaction time impact the answer quality.

reaction time was anticipated, the analysis reveals that the adverse effects of extended completion times (e.g., memory errors) outweigh the positive effects (e.g., increased time for computing the answer), as discussed in the previous section. Specifically, a comparison between the two groups of correct and incorrect answers, where a distance greater than 50 meters is considered an incorrect answer, yields the following values:

- *Reaction time*
  - Mean : Correct (38 minutes), Incorrect (43 minutes) (Fisher $F = 5.02$, p<0.05);
  - Median survival reaction time: Correct (17 minutes), Incorrect (19 minutes) (Log rank test: $\chi^2(1) = 4.93$, p<0.05; Non-parametric equality-of-medians test: $\chi^2(1) = 3.73$, p<0.05) [129].
- *Completion time*

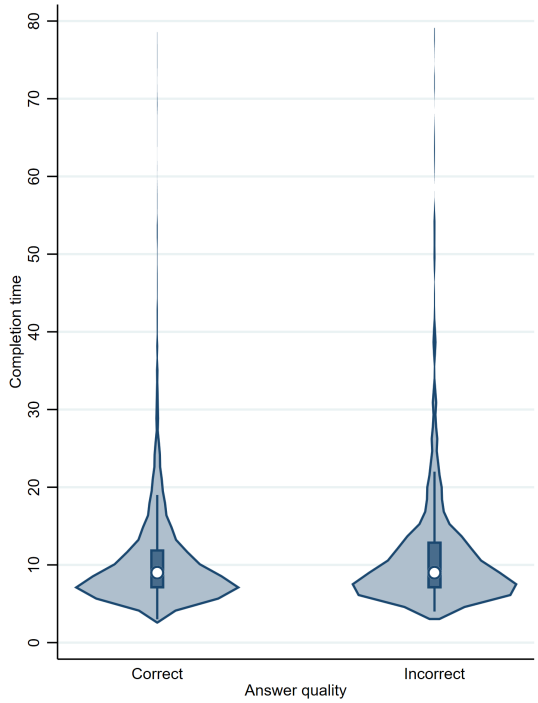


Figure 4.3: How does the completion time impact the answer quality.

- Mean : Correct (11.0 seconds), Incorrect (11.8 seconds) (Fisher $F = 4.73$, p<0.05);
- Median survival completion time: Correct (9 seconds), Incorrect (9 seconds) (Log-rank test: $\chi^2(1) = 4.36$, p<0.05; Non-parametric equality-of-medians test: $\chi^2(1) = 5.17$, p<0.05).

For both reaction and completion time the mean time for correct answers is lower than that of incorrect answers. This applies also to the median survival reaction time, while the median survival completion time is the same for correct and incorrect answers.

However, reaction time and completion time are not independent; they are interlinked components of the same process. As demonstrated in Table 4.3, our analysis of answer quality, using a Multilevel Structural Equation Model (MSEM) [149] and a Structural Equation Model (SEM), supports the theoreti-

Table 4.3: Chain of errors: Multilevel structural equation model and a Structural Equation model.

Features	ML-SEM Model (Coef B)	SEM Model (Coef B)
Home phone distance (meter)		
Completion time	0.2401***	0.0863***
Reaction time	0.4725**	0.0238**
GPS accuracy	0.3439**	0.0237**
Constant	7.3587	0.0457
Unanswered Questions (count)		
Reaction time	0.0252***	0.8809***
Constant	-0.1165*	-0.0691
Reaction time (minutes)		
User Characteristics: Procrastination	1.0325***	0.1216***
User Characteristics: Mood	1.4515***	0.0168*
Event context: Activity	0.5991***	0.0530***
Study time	1.6248***	0.1127***
Study time^2	-0.3105***	-0.0716***
Social context: Alone vs not alone	-13.5724***	-0.1170***
Constant	18.7450***	0.3596***
Completion time (second)		
Pending notification (count)	-0.4689***	-0.0946***
Question delivery delay time (sec.)	-0.0012**	-0.0212*
Event context: Activity	-0.2697***	-0.1784***
Social context: Alone vs not alone	-2.521***	-2.521***
Study time	-0.1818***	-0.0844***
Study time^2	0.0203***	0.0357***
User Characteristics: Procrastination	0.0766***	0.0644***
Constant	13.7792***	1.7036***

cal model depicted in Figure 4.1 and the concept of a chain of events impacting answer quality [153], also due to the correlation between reaction time and the number of pending notifications, the covariance between the errors of these two variables has been defined in both models. The input variables are predominantly relevant, and the fit indices indicate a good fit of the model to the observed data (RMSEA = 0.022; SRMR = 0.013; $\chi^2 = 95.41(20)$; $R^2 = 0.127$). The model explains 13.0% of the variance in answer quality. However, multilevel models pose challenges in constructing fit indices due to the multiple levels of hierarchy involved, resulting in a lack of consensus on suitable fit indices for such models [44].

As shown in Table 4.3, factors such as context history, burden, technology, and personality aspects play different roles in determining "when" the user decides to respond and "how" the user completes the notification. In a causal chain, reaction time (when) and completion time (how) directly affect the quality of data collected and associated errors. In the causal model presented in Table 4.3, the "when" influences both the question-answer process and the data quality through the number of notifications pending when the user begins to respond. This, in turn, is influenced by the burden, activity repetitiveness, and compliance with the research protocol, which varies according to the user's personality characteristics, such as procrastination, daily mood, and social context. The "how" is influenced by burden, distractions due to the presence of others, activity repetitiveness, and the user's level of procrastination, directly affecting response errors. In summary, burden, activities, personality idiosyncrasies, and cognitive aspects directly influence user behavior, which in turn directly impacts errors.

The overall conclusion is that *reaction time is the key variable with the highest impact on answer quality*. First, as shown in Figure 4.2 and Figure 4.3, reaction time is significantly longer than completion time, with a mean of 38 minutes for correct answers and 43 minutes for incorrect answers, compared to

11 seconds for correct answers and 11.8 seconds for incorrect answers. Second, reaction times on the order of half-hours or longer significantly increase the likelihood of memory errors, leading to incorrect answers. It is also notable that the difference between the median completion times for correct and incorrect answers is relatively small (11 to 11.8 seconds).

4.6 Description of the results

The purpose of this description is to present conclusive remarks and general take-away lessons. In the first part, we examine the impact of key exogenous factors (i.e., situational and temporal context) and endogenous factors (i.e., general user characteristics) on answer quality. In the second part, we analyze the chain effect from reaction time to completion time. Lastly, in this section, we provide a set of general recommendations, based on the findings from this project and the literature, on how to conduct future ESM data collections and experiments effectively.

Although we trust the reliability of our memories, as has previous work on EMA/ESM compliance, research on autobiographical memory reveals that memory can be unreliable [31, 188, 180]. Our recollections are not only inaccurate but are often systematically biased [170]. The longer the time elapsed from the event we want to recall, the higher the risk of inaccuracies. Everything hinges on the effects that different contexts have on time and, subsequently, on memory and the cognitive process related to response quality. Based on the current state of the art and the analyses we have conducted, we present a detailed view of how these aspects exert a direct and/or indirect significant effect on the entire response process and, consequently, how they collectively influence answer quality.

The situational context: Location, activity, and social context (Table 4.1) influence reaction time in varied ways. For instance, being “at university”,

"alone", and connected with "social media" significantly increases the likelihood of an early response to notifications. Conversely, being "outdoors", engaged in "leisure" activities, and with a "partner" can significantly delay the user's answer. This model highlights that merely detecting whether the phone is *on/off* [195] or if the subject is changing activities [146, 95] is insufficient. The interplay between location, activity, social relation, and time significantly impacts reaction time.

Activities and social context also influence completion time on two levels (Table 4.2). The first level is the expected reduction in compilation time due to event repetition. This reduction is driven by cognitive training, where users learn to classify different activities within a limited list of alternatives through repeated notifications. The higher the probability of an event occurring, the more likely the user will remember its position. The second level involves the impact of external distractions, where being alone decreases completion time (Table 4.2). When alone, users find it easier to concentrate on their answers.

The temporal context: Time context significantly affects user behavior. Over time, users tend to increase their reaction time, with this delay rising rapidly in the initial days before reaching a balance between the task and the time it is performed.

Conversely, app usage impacts completion time by reducing it over time. This reduction can be attributed to the learning process, which predominantly occurs in the initial observation period. The indirect effect on answer quality appears to stabilize after one week (Table 4.3). Over time, users find a balance among the frequency of server notifications (every half-hour), task burden, research compliance, daily routine, and cognitive ability. For long-term observations, it is necessary to replace the survey day with a more sensitive answer routine.

Additionally, there is a third type of time—social time [143], referring to students' daily activities and their variations throughout the week. Students do

not attend classes every day, nor do they follow a strict routine. Contrary to expectations, longer reaction times do not occur during peak academic activity (weekdays) but rather during weekends (Saturdays and Sundays). This suggests that personal time and relaxation reduce task attention.

User characteristics: Sex and age were found to have no statistically significant or even moderated effects. Procrastination increases the delay for both reaction time and completion time, while a sad mood reduces it. This aligns with the literature [70], which indicates that a lower mood enhances attention and memory. Notably, from the answer quality model in the previous section, mood has a significant direct effect on reaction time but not on completion time.

However, Steel [177] suggests a relationship between mood and procrastination. Procrastination temporarily alleviates anxiety, improving mood in the short term, but has negative long-term effects due to an accumulation of unfinished tasks. This may lead to a depressive spiral where depression leads to procrastination, resulting in a bad mood. Thus, in an interactive survey process with a smartphone, the mood may influence memory, accuracy, motivation, and, consequently, completion time.

4.7 The chain effect of reaction and completion time on the answer quality

Completion and reaction times correlate positively with the distance between the user's home and their phone, referred to as the variable "Home $\Leftrightarrow$ phone". This distance is crucial in this analysis, as the accuracy of user responses is measured against their home location. This correlation has several negative consequences on answer quality, as illustrated in Figures 4.2 and 4.3. Specifically, a longer reaction time (Table 4.3) has two effects. First, it directly increases the distance "Home $\Leftrightarrow$ phone" (the longer the user waits, the further they move away from home). Second, it indirectly affects the completion time by increas-

ing the number of "pending notifications" (the longer the user waits, the more unanswered questions accumulate). This effect, in turn (Tables 4.2 and 4.3), shortens the completion time, theoretically decreasing errors as previously analyzed. This result aligns with our analysis but illustrates two ways in which errors can occur. With long completion times, memory errors may arise. However, repeated activities (e.g., a student quickly and "automatically" filling in a sequence of four notifications with the same answer at the end of a two-hour class) increase the risk of reduced attention and typing errors.

4.8 Discussion

The key lesson learned from this section is that *reaction time is the key factor to be controlled to improve human answer quality*. Controlling completion time is considerably more challenging due to the numerous conflicting factors influencing it and the small difference between the time taken to complete correct and incorrect answers. To improve reaction time, it is essential to have accurate information about the user's location, activity, and social context—namely, the current context in its various dimensions (e.g., situational, social, temporal). This model demonstrates that simply noting whether the phone is *on/off* [195] or if the subject is changing activity [146, 95] is insufficient. The combination of different contextual dimensions, together with time, defines more precisely when a user is most likely to respond and provide a correct answer.

Evidence suggests that if we aim to study human activities in the wild, limiting observation windows to a couple of weeks, as suggested in [195], or to a single historical moment is inadequate. Human activities vary according to social time activities (e.g., weekdays and weekends), seasons, and social relations. There is no single best practice for conducting ESM research. The approach depends on (a) the research question and design (e.g., event-based sampling, time-based sampling, or a combination of both); (b) the phenomenon under in-

vestigation, its frequency and regularity, complexity, and interaction with other factors that need to be controlled (e.g., social context); and (c) the survey duration. As with traditional survey research, we must continue to evaluate and report the limitations and our strategies. The goal is to minimize errors and their consequences.

Based on these considerations and the experience gained from the work described in this thesis, we provide the following recommendations for organizing the various aspects of an EMA/ESM experiment.

(a) Avoid voluntary participation. As several research studies have shown, participants must be compensated [10, 109]. Without payment, user dropout and non-cooperation increase.

(b) Avoid short time limits when responding to notifications. Over time, users establish their own response routines. While a longer reaction time reduces workload, it can increase memory errors. However, it is preferable to have more responses, which can be excluded from analysis if necessary, than to have no information at all.

(c) Avoid complex cognitive tasks and limit questions to simple, factual information, especially during long observation periods (two weeks or more).

(d) Evaluate the duration of data collection and notification frequencies based on the topics under study. If studying habits or specific daily routines, design the observation time and notification frequencies according to the temporal evolution of the phenomenon (e.g., days or weeks).

(e) Consider both short and excessively long completion times as part of the response set. These responses are useful for computing median completion times and the diversity of answers (e.g., when compared with other users).

(f) Use smartphone sensors to assess the probability that an answer is correct. For example, sensors can be used to detect distractions or predict reaction times.

(g) Prioritize sending notifications when the phone is in active use (preferably

when the user is on social media) and when the user is alone (preferably in situations where the mood is likely lower or the user is bored, such as during travel or self-care).

Three conclusive remarks are presented. First, assuming the focus should be on improving reaction time, the next research question to address is: **What is the best time to ask a question?** This raises the issue of how to determine the optimal timing. Our recommendation is to minimize exogenous effects due to context history. Future research should develop a holistic approach, enabling systems to learn the best times to ask questions by considering optimal situational contexts and, where possible, the subject's endogenous factors. In this context, certain factors, such as procrastination syndrome and personality, can be computed once and used as input parameters for the machine learning algorithm; this research question will be handled in Chapter 6.

Second, the analysis in this part is based on data collected from a student population. The sample selection is motivated by the accessibility of students and their propensity for innovation and participation in research experiments. This choice aligns with a longstanding tradition of studies focusing on this demographic, as seen in [194, 195, 214, 82, 19]. To achieve broader generalizability of the results, future studies should extend this type of research to other populations.

Third, in this study, we send questions according to a fixed schedule and administer numerous questions to each participant daily. Given individual differences in procrastination syndrome and personality, some participants may become annoyed by the frequency of questions, potentially compromising the quality of their responses. To address this issue in future studies, we should minimize the number of questions, paying particular attention to reducing the number of disruptive online questions. These questions should be prioritized based on the risk of forgetting or the urgency of decision-making. Section 6.5 will address this aspect of our work.

In this work, we have investigated the effects of various factors on the correctness of answers. Our study, based on an empirical analysis of a large dataset of daily behaviors, captures a rich and multifaceted picture of individual behaviors and potential interaction errors.

Our results suggest that, while subjective annotations are useful in predicting certain contextual patterns, they present a degree of error due to both exogenous and endogenous factors affecting response quality. Context history, cognitive ability, attention, effort, motivation, burden, procrastination, mood, and technical problems can increase the likelihood of stopping interaction with the machine, non-compliance with the interaction protocol, reduced attention, and consequently, decreased answer accuracy.

In research studies where users provide data to a third party, these problems add to many known issues. Examples include the social desirability effect, which may prevent participants from reporting certain socially disapproved activities [46]; unreported activities when perceived as privacy intrusions [38]; and the incorrect design of data collection instruments, such as a lack of an exhaustive list of response alternatives allowed to the respondent [22, 88]. Furthermore, in practical applications, collecting self-reported annotations is not always feasible.

Chapter 5

High-Quality Answers Collection Methodology

5.1 Introduction

As we discussed in Chapter 1, many works are working on collecting human-generated data. However, most of these studies have only concentrated on the use of sensor data. This, in turn, has generated two types of errors, from the first perspective, while providing a good objective representation of reality, sensors are unaware of the subjective motivations behind a given individual's activities. From the second perspective, most noticeably sensor data validity, namely the accuracy of the indicator of the phenomenon being measured, and data completeness; besides that, we also pay attention to the data quality of human input, namely the time diary data collected directly from humans. In this section, we use the WeNet dataset to explain and test our method, The WeNet dataset has been introduced in Section 3.4.

To collect high-quality human input data, we present a methodology via three key functionalities, where each feature builds upon and extends the features offered by the previous one.

- *Context modeling: representing the situational context*: Understanding and exploiting the personal context, including the user's internal physical

and psychological states, is crucial for the concept of thick data [72]. This relevance has been highlighted in numerous personal mobile applications (see, e.g., [103, 159, 193, 100, 214]). Knowledge of the user's physical, social, emotional, and informational states allows for better interpretation of the vast amounts of big sensor data collected [42], enabling the creation of context-aware systems that adapt to users' real-life conditions and improving the relevance and quality of the collected data [53, 30]. The key innovation is focusing on the context of the data collection itself. Our methodology operates within a context whose sole purpose is to enhance the awareness of the machine, participant, and researcher regarding the data collection process, as a first step toward increasing control, quality, and richness of the collected data. Our ultimate aim is to generate big-thick data about the process of generating big-thick data, the former being a key enabler for the latter. This part will be introduced in the Section 5.2.

- *Planning language: representing the temporal context*: An experiment is modeled as a plan where each action—such as a human answer to a machine question, sensor data collection, or a machine answer to a human question—is associated with a set of *scheduling constraints* and *execution annotations* that encode information about past, present, and future actions. Examples of scheduling constraints include the time frame within which a question can be asked and the condition that it should only be asked when the participant is at home. Examples of execution annotations include noting if an answer was not provided or was delayed by half an hour. To this end, we have developed a representation language called planning language, which allows for representing all context dimensions, both temporal and situational, and using them to condition the activation of user questions and sensor data collection. Any researcher designing a data collection experiment can declaratively specify the questions to be asked and the conditions under which they can be asked via a user-friendly

graphical interface (e.g., calendar). This plan can be shared with experiment participants before, during, and after data collection to build consensus or obtain approval. This structured approach ensures that the context in which data is collected is well understood and leveraged, enhancing the overall quality and relevance of the collected data. This part will be introduced in the Section 5.3.

- *pbgc-Forest: learning human answer behavior*: The data collection plan can be revised in real-time by individual participants within bounds set by the researcher, either for themselves or for a group of colleagues, or by the researcher for one or more participants. The plan can also be revised by the platform itself, based on a machine learning algorithm that has learned the optimal and suboptimal times for obtaining high-quality responses. We propose a novel algorithm for answer quality prediction, the PCA-based multi-grained cascade Forest (pbgc-Forest), leveraging the concept of deep forest (also known as multi-grained cascade forest, or gcForest) [215, 216]. In the case of conflicting suggestions, the researcher can establish a variation period (possibly null) within which they can modify the plan. The decision of the machine can always be overridden by the participant and, subsequently, by the researcher. The control hierarchy proceeds from the researcher to the participants and then to the platform. This hierarchical control system ensures flexibility and adaptability in the data collection process, enhancing the quality and relevance of the collected data. This part will be introduced in the Section 5.4.

5.2 Context modeling

To let machines understand human daily life, it is necessary to define human personal context. The complexity in defining and understanding the notion of personal life led to the development of a comprehensive concept of the personal

situation context grounded on the practical aspects of everyday life. Taking a sociological approach and focusing on individuals' subjective perceptions of reality, together with some objective information, gives a more accurate portrayal of their reality. Undoubtedly, this gives rise to numerous concerns such as difficulties related to perception, the formation of more complex conceptions, and the dissemination of information. The objective component of our approach enables us to comprehend reality by using sensor data, hence enhancing our representation of the everyday life constructed within this system. From the subjective point of view, we decided to pose questions to the humans to let them provide the information.

Several research predominantly collects contextual information from objective sources [193, 100]. However, we consider both quantifiable parameters, such as time measured using sensors, and qualitative factors, such as moods given by self-reports (questionnaires) as the experiments we have done in Section 3. These together constitute what we refer to as the situational context.

The concept of context used in this section is built upon prior research in [74, 82, 203, 73]. To explain how we use the notion of context in this setting, we present a motivating example: an evening where the activities of a participant, *(me)*, are captured as in Figure 5.1:

- During a first period of time T1 (yellow box), *me* is at a pizzeria having lunch with the friend John. They are having pizza and *me* is happy;
- Then, in the following period of time T2 (gray box), *me* is driving to work alone and is in a scared;
- Finally, during T3 (pink box), *me* is in a meeting in office with the colleague Li and *me* is in a neutral mood.

The above work is a set of daily scenarios, and we define the situational context, which is the annotated context that can used to represent these scenarios

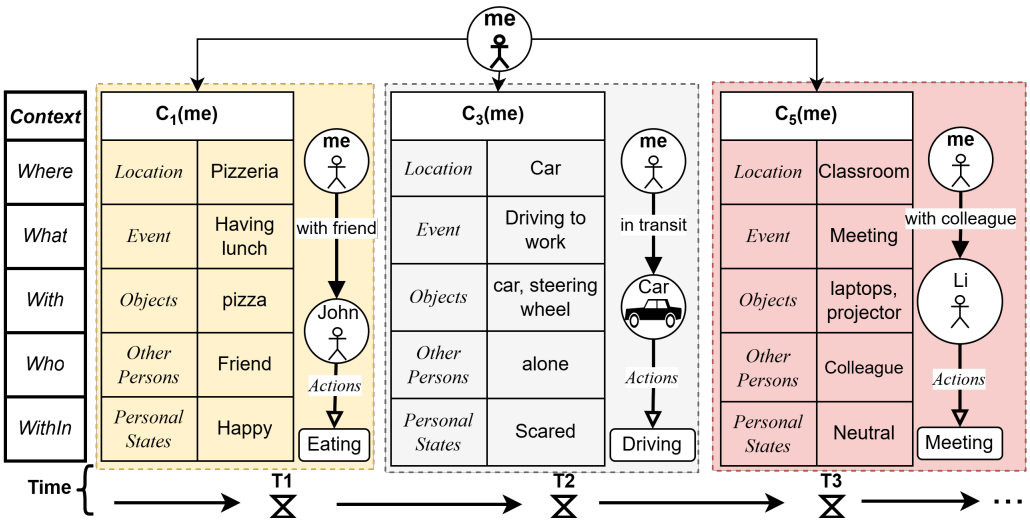


Figure 5.1: A motivating example for everyday life.

in detail, for example, taking the scenario in T3, where *me* was having a meeting at the office with Li.

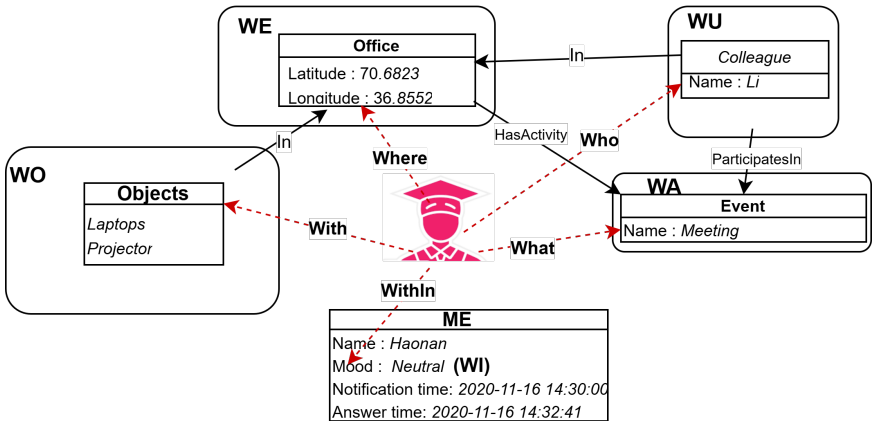


Figure 5.2: A motivating example of situational context model.

Figure 5.2 shows the above scenario as a knowledge graph, representing the personal context of the participant *(me)*. The components of the graph are explained as follows:

$$MyWorld = <Cxt,\ me>$$

where: *me* is the subject and *Cxt* is the context (real world) of the subject, including elements in his local immediate surrounding. The red arrows in Figure

5.2 show the relation between *me* and *Cxt* thus;

$$Cxt = < WA, WE, WI, WO, WU >$$

where:

- *WA* is the temporal context from answering the question; "**W**h**A**t are you doing?". In Figure 5.2, *WA* is the main event, i.e., *meeting*.
- *WE* is the spatial context generated from the question "**W**h**E**re are you?". In Figure 5.2, *WE* shows the most relevant location, i.e., the *office*.
- *WI* is the internal context generated from the question "**W**hat mood are you **I**n?". As shown in Figure 5.2, it is the emotional state (mood) of *me*, i.e., *Neutral*.
- *WO* is the object context generated from the question "**W**hich **O**bject are you with?". In Figure 5.2, *WO* includes a few objects, e.g., the *projector*, and the *laptops*.
- *WU* is the social context generated from the question "**W**ho is with you (**U**)?". As shown in Figure 5.2, *WU* focuses on *me*'s colleague *Li*.

The obtained sensor data is used to integrate supplementary attributes into the context. As an illustration, in Figure 5.2, the *latitude* and *longitude* attribute values in the *classroom* entity of the *WE context* are sensor readings whereas the *mood*, *notification time* and *answer time* attributes in *ME* are got from the questions sent to *ME* (participant). Additional gathered sensor data includes; accelerometer, social media apps, bluetooth devices among others, as detailed in [75].

5.3 Planning language

We model the temporal context using the planning language, a scheduling language developed on top of iCal. the iCalendar/RFC5545 (iCal) standard. Using

the iCal standard provides three primary benefits, namely;

1. Researchers can use the Recurrence Rule[1] (RRule) to simplify the formulation of repeating questions and sensor collections.

2. Participants can easily modify their schedule by eliminating questions or adjusting response times in a calendar, providing flexibility and convenience, and

3. It enables us to convert our scheduling language into an open-source calendar application, e.g., *Fantastical Calendar*[2].

Our planning language is based on the iCalendar (iCal) standard, as defined by the Internet Engineering Task Force (IETF) in RFC 5545, which is a widely adopted format for representing and exchanging calendaring and scheduling information. It facilitates the sharing of events, to-dos, journal entries, and free/busy information across various systems and applications, including Microsoft Outlook, Google Calendar, and Apple Calendar. iCalendar is extensively used for personal, professional, and organizational calendaring, providing a flexible and comprehensive solution for managing schedules and tasks across different platforms. This can also provide the flexibility to allow the participant to change the time to provide feedback via the calendar component.

Key features of iCalendar include its interoperability and support for various data components such as VEVENT (events), VTODO (to-dos), VJOURNAL (journal entries), VFREEBUSY (free/busy time), and VTIMEZONE (time zones). Each component contains properties that define its details, including start and end times, summaries, locations, and descriptions. In our work, we define the novel VEVENT to collect questions and sensor data. The iCalendar format also supports recurrence rules (RRULE) for defining repeating events. The RRULE property in the iCalendar standard is a powerful feature that allows

[1] https://icalendar.org/iCalendar-RFC-5545/3-8-5-3-recurrence-rule.html
[2] https://flexibits.com/fantastical

users to define recurring events, to-dos, journal entries, or time zone definitions, making it easier to manage and schedule recurring activities.

The planning language allows participants to choose when they are available to answer questions or permit sensor collections (modify the plan), thereby enhancing their attention and improving collaboration between participants, researchers, and AI. In our methodology, the planning language organizes the specification of an experiment in three main components, as follows:

1. the general schedule aggregating all the different components;
2. The question-answering component;
3. The sensor data component.

We introduce below a snapshot of these three components using the Bachus-Naur Forms (BNFs) notation and describe their main features.

Experiment General schedule:

```
<User>                  := <Calendar>;
<Calendar>              := <calendar id>, <Context collection>;
<calendar id>           := Integer ;
<Context collection>    := <Question collection>   |   <Sensor
                           collection>;
```

Figure 5.3: Experiment general schedule.

A key observation is that a user may be associated with multiple calendars, each containing multiple question collections as well as multiple sensor collections. This design allows a user to participate in multiple experiments simultaneously and facilitates diverse data collection within the same experiment. The `calendar id` distinguishes the different calendars used by the user in the same experiment.

<Question collection> := <Qid>, <time>, <status>, <RRule>, <question> ;
<Qid> := Integer ;
<time> := <notification time>, <answer time>, <end time> ;
<notification time> := DateTime ;
<answer time> := DateTime ;
<end time> := DateTime ;
<status> := Boolean ;
<RRule> := <Interval>, <Count>, <Frequency> ;
<Interval> := Integer ;
<Count> := Integer ;
<Frequency> := Daily | Weekly | Monthly | Yearly ;
<question> := <question id>, <dtstart>, <dtend>, <question category>, <content>, <question type> ;
<question id> := Integer ;
<dtstart> := DateTime ;
<dtend> := DateTime ;
<question category> := WA | WE | WU | WO | WI;
<content> := <question content>, <answer content> ;
<question content> := String ;
<answer content> := String ;
<question type> := Dichotomous Question | Multiple choice question | Single choice question | Free Text Question | Image Question ;

Figure 5.4: Question collection.

Planning Language for Questions:

`Question collection` is a way to collect context data via the calendar, with each question collection composed of a unique id `Qid` to identify the questions, a Recurrence Rule `(RRule)`, a `time` range about the questions, a questionnaire, and a `status`, here we especially add the last three factors to make the planning language become more concrete with the questions, then we explain these factors in detail. Time includes the `notification time`, which is the time the participant's smartphone received questions; the `answer time`, which is the time the participant started to give answers; and the `end time`, which is the time the participant finishes giving the answer and submits. The `status` field indicates whether the event has been accepted by the user, with a value of 1 indicating acceptance and 0 indicating rejection.

The `RRule` specifies the values used to determine each recurrence and how the event should be repeated. The RRule includes three elements: `interval`, `count`, and `frequency`. The interval specifies the number of frequency units that must elapse before the next occurrence of the event. The frequency defines the unit of time for repetition, such as daily, weekly, or yearly, whereas the count specifies the number of times the event will be repeated. A `question` is composed with a start time `(dtstart)` and end time `(dtend)`. For the definition of the `question category`, we base on research in [73, 74] and our context modeling methodology in Section 5.2. For the `content` part, we have the `<question content>`, which means the text of the query we intend to pose to the participant, e.g., "*What are you doing?*", the potential answers would be included in the `<answer content>`, e.g., "*Meeting*". For the definition of `<question type>`, we follow the inspiration in [32], which shows the participants how they can answer the questions.

Example of planning language for questions: Table 5.1 provides a real-world example of our planning language for questions based on the WeNet experiment, which we described in Section 3.4. Our aim is to comprehend the user's

situational context, for which we pose several questions to a user. For instance, the question in the first column is designed to gather information about the temporal context (WA). Questions in the second and third columns are intended to collect spatial context (WE) and social context (WU) information, respectively. To provide the user with greater flexibility in their responses, we opted for multiple-choice as the question type. The plan was to dispatch the question to the participants 48 times a day (once every 30 minutes) over a span of 28 days (four weeks). This was achieved by setting an RRule with a daily frequency, an interval of 48, and a count of 28. In addition, from the last column of Table 5.1; the question was designed to understand the participant's internal context. To encourage them to respond, we employed 5 images to represent different moods. Our plan was to send this question to the participants 24 times a day (once every hour) for 28 days (four weeks), by setting an RRule with a daily frequency, an interval of 24, and a count of 28.

Table 5.1: Sample questions used in the WeNet project

Qid	**question content**	**question category**	**RRule**			**question type**	**answer content**
			Interval	Count	Frequency		
1	What are you doing?	WA	48	28	Daily	Multiple choice question	The participant was provided 34 answer categories such as sleeping, eating, etc.
2	Where are you?	WE	48	28	Daily	Multiple choice question	The participant was provided 26 categories such as home, university, restaurant, etc.
3	Who is with you?	WU	48	28	Daily	Multiple choice question	The participant was provided 8 categories such as alone, partner, friends, etc.
4	What is your mood?	WI	24	28	Daily	Image question	The participant was provided a scale of 5 levels ranging from happy to sad.

Note: The columns in Table 5.1 correspond with the BNF depicted in Fig. 5.3. Specifically, the **Qid** and **RRule** align with the parameters under the `Question Collection` category. Other parameters within the `Question Collection` category, namely `time` and `status`, are associated with the participant's answer process and are therefore not displayed in the table. Similarly, under the `Question` category, the `dtstart` and `dtend` parameters are not shown in the table due to the same considerations. In addition, the **question type** has been shown in the Table 5.1 because this should be designed by the researchers

before sending the questions to participants. Furthermore, the **question content** and **answer content** fall under the `content` category within the `question` part, as with **question type**.

Planning Language for Sensor Data:

`<Sensor Collection>`	:=	`<Sid>`, `<dtstart>`, `<dtend>`, `<sensorRRule>`, `<sensor>` ;
`<Sid>`	:=	Integer ;
`<dtstart>`	:=	DateTime ;
`<dtend>`	:=	DateTime ;
`<sensorRRule>`	:=	`<Interval>`, `<Count>`, `<Sfrequency>` ;
`<Interval>`	:=	Integer ;
`<Count>`	:=	Integer ;
`<Sfrequency>`	:=	Millisecondly | Secondly | Minutely | Hourly ;
`<sensor>`	:=	`<Name>`, `<Description>`, `<Sensor type>` ;
`<Name>`	:=	String ;
`<Description>`	:=	String ;
`<Sensor type>`	:=	Social | Motion | Location | Inertial | Device | Ambient;

Figure 5.5: Sensor data collection.

`Sensor collection` refers to the method of gathering context data via the calendar, identified by a unique sensor id `(Sid)`. Since the i-Cal standard's RRule cannot handle recurring events at the minute or second level for sensor collection, we had to define a new entity type, the `sensorRRule`, which is similar to the RRule but handles sensor frequency `(Sfrequency)` at the millisecond, second, minute, and hour levels. A `sensor` is specified by a `name`, which indicates the sensor's identifier; a `description`, which explains how the sensor can be used; and a `sensor type`, which informs the user of the purpose for collecting this type of sensor data.

Example of planning language for sensor data: Table 5.2 serves as a real-

world example of our planning language for sensor data based on the WeNet experiment, which we described in Section 3.4. Our objective was to understand the participant's situational context based on the sensor type. For instance, sensors in the first, second, and third columns were used to collect social, ambient, and location types of sensor data, respectively. We planned to gather this sensor data every minute for a duration of 14 days. This was achieved by setting a sensorRRule with a minutely, an interval of 1, and a count of 20160. Additionally, for sensors in the fourth and fifth columns, which are used to collect device and inertial data, we planned to collect their data every 5 seconds for a period of 14 days. This was facilitated by setting a sensorRRule with a frequency set to secondly, an interval of 5, and a count of 1209600.

Table 5.2: Sample sensor data collected in the WeNet project

Sid	**Name**	**Sensor type**	**sensorRRule**			**Description**
			Interval	Count	Sfrequency	
1	Bluetooth Device Available	social	1	20160	Minutely	Bluetooth low energy devices detected
2	WIFI Networks Available	ambient	1	20160	Minutely	List of WIFI networks detected by the smartphone
3	Location	location	1	20160	Minutely	Location information using GPS connections
4	Applications	device	5	1209600	Secondly	Name of the applications that are in the front end
5	Acceleration	inertial	5	1209600	Secondly	3D vector of the acceleration

Note: The columns in Table 5.2 correspond with the BNF depicted in Fig. 5.5. Specifically, the **Sid** and **sensorRRule** align with the parameters under the `Sensor Collection` category. Other parameters within the `Sensor Collection` category, namely `dtstart` and `dtend`, are associated with the user's answer process and are therefore not displayed in the table. Furthermore, the **Name**, **Description** and **Sensor type** fall under the `sensor` within the `Sensor Collection` scope.

In summary, our planning language offers researchers enhanced flexibility in formulating questions with varying frequencies using the RRules. Our planning language further empowers researchers to craft questions in their preferred format, such as free text or multiple choice. Researchers are also accorded the flexibility to design the recurrence for sensor data collection and categorize sen-

sors based on their intended usage. Moreover, we provide participants with the capability to track feedback time via their calendars. On top of extending the iCal standard, our novel planning language maintains compatibility with various open-source packages that visualize calendars. The calendar enables users to synchronize data collection events with their real-life schedules and evaluate whether these events disrupt their daily activities.

5.4 The pbgc-Forest algorithm

As our goal is to enable the machine to automatically identify the optimal moments when participants can provide high-quality answers, the current machine learning algorithms, while capable, do not achieve high accuracy, rendering their results somewhat unreasonable. To address this issue, we introduce a novel methodology for answer quality prediction: the PCA-based multi-grained cascade Forest (pbgc-Forest), which leverages the concept of deep forest (also known as multi-grained cascade forest or gcForest) [215, 216]. This approach is characterized by three distinct features:

1. **Cascaded Structure:** In the context of deep neural networks, representation learning primarily relies on the layer-by-layer processing of raw data. Drawing inspiration from this principle, the cascaded structure serves as a crucial mechanism to emulate deep models. Each layer in the cascade processes feature information transferred from the preceding layer and forwards the current processing result to the subsequent layer. Each layer within this cascaded structure constitutes an ensemble of decision tree forests, effectively creating an ensemble of ensembles.

2. **Multi-Grained Scanning:** The scanning process can be described as follows: Initially, a sliding window of length K scans the complete P-dimensional sample, yielding $S = (P-K)+1$ subsample feature vectors (K-dimensional).

Each subsample is then used for training completely-random tree forests and random forests. Each forest generates a probability vector of length C, resulting in each forest having a representation vector of length $S \times C$. The forests in each layer are concatenated to produce the output layer. The sliding window mechanism effectively leverages various sizes of samples, capturing more characteristics of the sample and truly realizing multi-granularity.

3. **Tree-Based Integrated Model:** By employing deep forest, we obtain an effective learning scheme that capitalizes on an interpretable and reliable model for future mood prediction. This scheme requires significantly fewer parameters and is less sensitive to parameter settings.

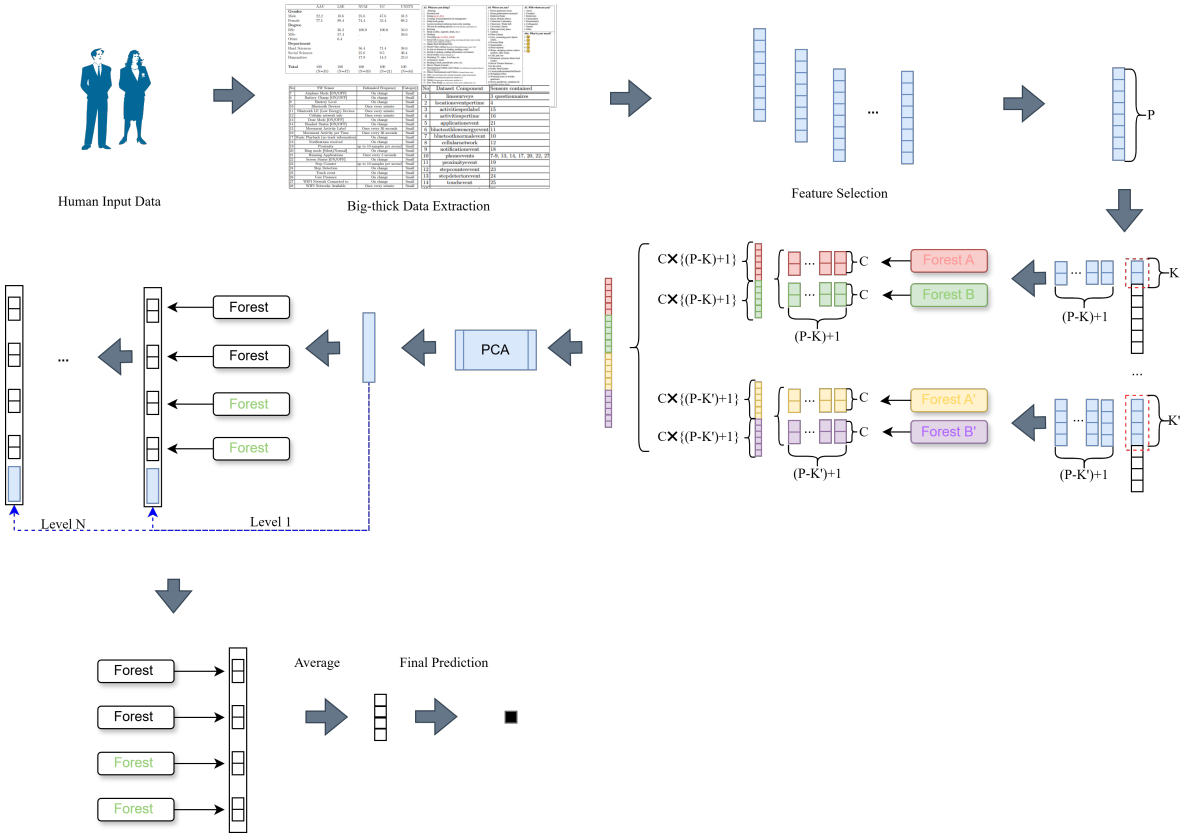


Figure 5.6: The architecture of pbgc-Forest model

To illustrate how the pbgc-Forest identifies optimal moments for participants to provide high-quality answers, we present the architecture of the pbgc-Forest in Figure 5.6. Initially, data collection is performed using our novel platform, which gathers big-thick data from participants, followed by normalization to obtain original feature data. This dataset is then split into training and test sets, with the training set's answer quality features scanned through a multi-dimensional window and input into the random forest for feature transformation, yielding high-dimensional features. Next, PCA feature extraction is conducted on the transformed features, where they are normalized, the optimal number of PCA principal components is selected via singular value decomposition, and the feature space is transformed to obtain reduced-dimensional features for input into the cascade forest. These PCA-extracted features are first fed into the initial layer of the cascade forest for training, after which enhanced features from each layer are combined with the PCA-extracted features and passed to subsequent cascade forest levels. Finally, the test dataset is input into the trained model to make classification predictions and output diagnostic results, ensuring the pbgc-Forest can effectively identify the optimal times for participants to provide high-quality answers using advanced machine learning techniques to enhance data collection and analysis.

In essence, our approach presents an innovative and efficient solution for answer quality prediction by harnessing the power of deep forest. As we conducted the data collection process ourselves, and there are no other studies using the same data, the effects of our experimental methods are not compared with others'. This is because different methods of collecting experimental data and different characteristics for prediction can produce vastly different results. Therefore, we employed multiple machine learning methods to confirm the effectiveness and feasibility of our proposed method. The real-world experimental data used for classification encompasses both numerical and categorical values, it is crucial to consider which types of classifiers would perform well on

such data, as it is not immediately apparent without evaluation. For instance, tree-based classifiers (e.g., Deep forest and Random Forest) generally perform well and are robust on such types of data. The features of our data may not be independent, with examples including the interaction between activity and location (e.g., a participant in the classroom is always studying and not eating). Domingos and Pazzani [57] demonstrated that Bayesian classifiers can be applied to such data and can achieve good performance. Additionally, tree, rule, and Bayesian-based classifiers are widely adopted in the domain of interruptibility prediction, such as Random Forest, ANN, and Naive Bayes.

Chapter 6

Answer Quality Prediction

6.1 Introduction

After we know the reaction time is the key factor that can impact human answer quality in Chapter 4 and the high-quality answers collection methodology in Chapter 5, the following step is to find the optimal moment to pose questions to a participant, with the aim of enhancing the quality of the responses. In this context, the quality of a response is defined as the minimal number of errors made when answering machine-generated questions. To the best of our knowledge, no existing study has specifically focused on the quality of human answers. Current research in the field of human-in-the-loop AI primarily concentrates on predicting the most suitable time to send notifications, see, e.g., [132]. The researchers consider the objective factors that can influence the time of checking notifications, rather than taking into account the subjective perspective. Variations in individual personality traits [58], human values [167], and experiences can result in differing response behaviors among distinct individuals [130].

We already know that the quality of responses is influenced by reaction time, which is defined as the time interval between receiving the notification and providing an answer; the less reaction time, the higher the quality of answers. In this Chapter, we will test which parameters impact reaction time and, therefore, influence human answer quality. We consider three factors that can impact

human answer quality:

1. **Situational Context:** As described in Chapter 5, human daily life can be considered a set of situational contexts. If a participant receives questions at an uninterrupted moment (e.g., while driving or in a meeting), they cannot respond promptly, leading to longer reaction times and, consequently, lower answer quality.

2. **Human Differences:** Individual variances in personality traits, lifestyles, and experiences can lead to different behaviors in distinct individuals, even in the same situational context. For example, introverts might experience discomfort at social events, while extroverts might derive enjoyment [130].

3. **Number of Questions:** Frequent questions from smartphones can be intrusive, resulting in annoyance and a reduction in work and study efficiency [155]. This can lead to survey fatigue, where participants become weary or lose interest in the survey, resulting in lower answer quality [106].

To solve the above issues, we propose three research questions(RQs):

- *RQ1: How does situational context influence the answer quality?*
- *RQ2: Which subjective or objective factors significantly impact the prediction accuracy of answer quality?*
- *RQ3: How many questions can be answered with high-quality answers per day?*

Answers can be annotated with labels based on their quality as 'High' or 'Low'. This involves first calculating the reaction time of these answers using Equation (4.1) (reaction time (t_r) values of unanswered records are set as ∞). Subsequently, we use Equation (4.3) to indicate the quality of the answers. There is a relationship between reaction time and answer quality. For different

datasets, the cut-off point of reaction time is challenging to determine, depending on the proportion of good-quality answers in the databases and the expected quality from answer collectors, etc. In Chapter 4, we found that the mean reaction time of the correct answers is 38 minutes, considering the time interval of each question (30 minutes), we propose the cut-off point of reaction time (RT) can be set at **30 minutes** to differentiate between high-quality and low-quality answers. Based on the answer quality, we annotated answers with labels (e.g., 'High' or 'Low'; '1' or '0'). In other words, we regard the participant's interruptibility as a binary predicted variable, which signifies whether the participant could respond to the questions within 30 minutes or not, and consequently, the quality of the responses.

To address the issue of data consistency, we use a different dataset, the WeNet dataset, than the one used in the Chapter 4. The WeNet dataset is introduced in Section 3.4. We present the features employed to address three research questions (RQs) in Section 6.2. Sections 6.3, 6.4 and 6.5 address Research Questions RQ1, RQ2, and RQ3 respectively. Finally, Section 6.6 presents the conclusions drawn from our work in this Chapter.

6.2 Features

Table 6.1: Data types used in this survey.

Feature	Description
Sex	Gender at birth: Male / Female
Department	Recoded department
Degree	Recoded degree
What	What are you doing? when answer
With whom	With whom are you? when answer
Where	Where are you? when answer
Mood	Mood at the time of answering: Sad – Happy (1-5)
Week	Day of the week (1-7)
Cohort	Classification of participants based on gender and degree
Day of survey	The number of day of the survey
Hour	Time of answer in biotime clock (1-24, 05:00=1 at 5am)
seqtot_mean	Mean frequencies of answers in each hour
Personality traits	Big Five personality traits scale [58]: Extroversion, Agreeableness, Conscientiousness, Neuroticism, and Openness.
Human values	Schwartz scale [167]: Benevolence, Conformity, Security, Tradition, Universalism, Self-direction, Stimulation, Hedonism, Achievement, and Power.

We introduce a series of metrics derived from quantifying users' profile information and situational context. These metrics underpin our explanatory and predictive models, which we will detail in the subsequent section. Some of these metrics have been extensively used in the ubiquitous computing community [206], while others, such as human values, are novel. We present all features in Table 6.1. And we showed the detailed information as follows:

Time context: Chen and Kotz [40] propose the inclusion of Time Context in the interruptability model, as time is a fundamental and natural context in which human, physical, and computing contexts are recorded over time to compile a context history. Our experiment spanned four weeks. We segmented the time context into three categories: Hour, Week, and Day of the survey. We consider the hour to commence at 5:00 and conclude at 4:00 the next day, ranging from 1 to 24. The week extends from 1 to 7, representing Monday to Sunday. Lastly, on the day of the survey, our experiment lasted for 28 days. We would like to know whether the participant can provide different answer behavior during the different experiment periods, e.g., the beginning and the end of the survey. A prolonged experiment can impose a burden on humans. As [67] defines, burdens as being associated with the content and presentation of the questionnaire and the modes of data collection. In other words, the longer the experiment endures, the greater the burden on the participants.

Social demographic: Social Demographic is a feature consistently used to characterize different individuals through a set of ascriptive (e.g., sex) and acquisitive (e.g., degree and department) traits [24], as well as the cohort age feature to depict the groups of participants and ascertain how social demographic information can impact the time taken to answer questions.

Situational context: From a smartphone perspective, context is a multidimensional concept where environmental, social, and technical conditions interact. To comprehend the participants' situational context, we posed three multiple-choice questions: *What are you doing?*, *Where are you?*, and *Who are you*

with?. The responses to these three questions enabled us to gather information about what we term the *event context*, the *spatial context*, and the *social context*, the three primary dimensions of the situational context as defined in [74].

Personality Traits: In smartphone data collection experiments, researchers have observed that some individuals are willing to participate in extensive surveys, experiencing little to no burden, while others discontinue participation within a minute [200]. It is evident that these factors can vary differently at different times, even for the same individual, and can differ from one person to another. It is known that the psychological state of the interviewer prior to contact may influence their behavior during the interaction [87]. Methodologists have begun to pay attention to these aspects. Empirical evidence indicates that the response style factor is strongly correlated with personality, notably the "Big Five" factor [58].

Human Values: Besides the Big Five, from the perspective of personality trait theory, we hypothesize that there is an alternative approach to comprehensively understanding humans. From this hypothesis, we discovered Human Values (Schwartz scale [167]), which categorizes humans into ten dimensions. We posit that human values can provide a more comprehensive explanation of people's behavior than the big five personality traits. This is the first instance of introducing human values into the interruptibility prediction model.

Mood: While personality traits and human values do not frequently change over time, a respondent's mood can often fluctuate within the same day. Recently, researchers have shifted their focus to respondents' moods [69]. Steel [177] posits that mood can induce procrastination, causing participants to delay answering questions. Moreover, it is probable that mood may influence the memory, motivation, and accuracy of the answers.

6.3 Answer quality variation by context

Firstly, we employ a multilevel logistic regression model to address RQ1. We employ a multilevel logistic regression approach to model it with an underlying continuous variable and assume a normal distribution for the underlying continuous value. Hence, the interruptibility binary is generated by a thresholded cumulative-normal model. We treat both the categorical variables (e.g., location, activity, relations) and the continuous variable (e.g., time) as numeric variables. Subsequently, the result is divided into six macro areas. The first is the time context. The second, third, and fourth are, respectively, the spatial, event, and social contexts. The fifth and sixth pertain to the respondent's characteristics, which are moods and human values.

Impact of Time Context on Answer Quality: Figure 6.1 illustrates the percentage of questions answered out of the total messages received during the first two weeks. The x-axis represents the hour of the day, and the y-axis signifies the probability of participants providing high-quality answers. The maximum contribution interval is between 7:00 and 8:00. This is primarily because participants start to wake up and check the questions. The time axis is based on a biological rhythm rather than a calendar basis, i.e., the time window commences at 5:00 and concludes at 4:00 the following day. Indeed, a participant's day does not necessarily begin at midnight and end at midnight on the same day. We selected 5:00 as the starting point because more than 98% of participants are asleep at that time and begin to awaken.

Impact of Spatial Context on Answer Quality: We conducted an analysis of nine commonly visited place categories, such as University, Home, and outdoors. The results are depicted in Figure. 6.2. It can be observed that when participants are in the university, they exhibit a higher likelihood of responding to questions promptly (58%). This finding contradicts our initial expectations, which had assumed a lower likelihood of answering questions within 30 min-

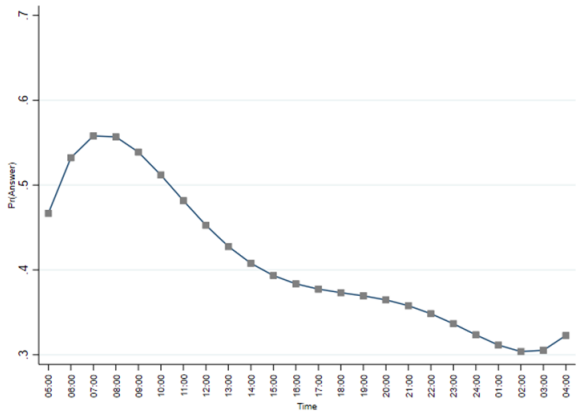


Figure 6.1: Responses filled out on messages received during the day (in %).

utes at a university. Conversely, when participants are at the workplace, they demonstrate the lowest likelihood (32%) of responding to questions promptly. In these results, we only consider the location as a variable in this analysis. However, in real life, individuals often inhabit different places with different people or engage in various activities. The reason we only consider one feature in a subsection is to avoid collinearity between the three dimensions of the time diary.

Impact of Event Context on Answer Quality: In this analysis, we have categorized the activities into 10 distinct types, such as Traveling, Working, Eating, and Using social media. The results are presented in Figure 6.3. It is evident that activities can influence the likelihood of answering questions within 30 minutes, and consequently, the quality of answers. For instance, when participants engage in social media activities, they exhibit a 72% likelihood of answering questions within 30 minutes. Conversely, when participants are enjoying free time, they demonstrate a lower likelihood of promptly answering questions, as they are preoccupied with other tasks and may not have the time to attend to smartphone notifications.

Impact of Social Context on Answer Quality: We categorized different interactive individuals into eight groups: Alone, Friends, Relatives, Classmates, Roommates, Partners, Colleagues, and Others. The results are presented in

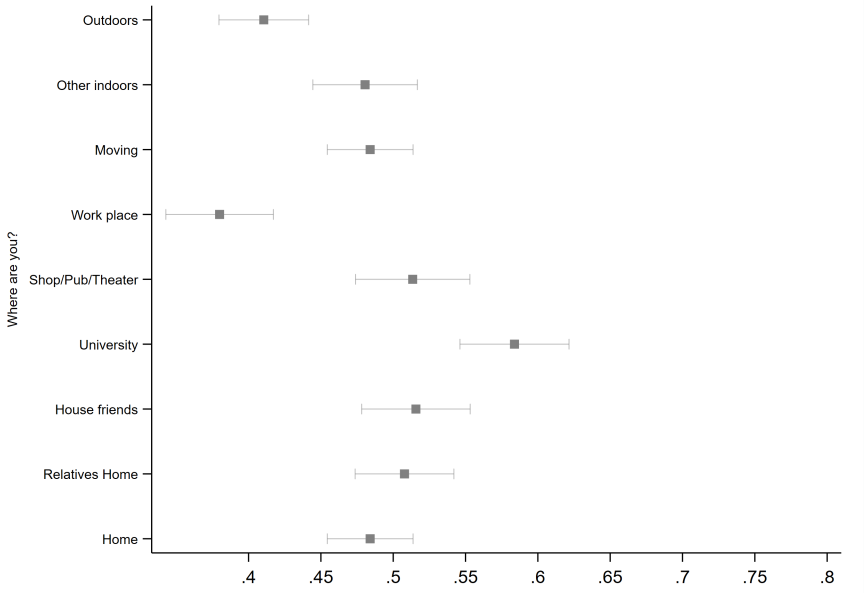


Figure 6.2: The statistical result of answering questions within 30 minutes at different places.

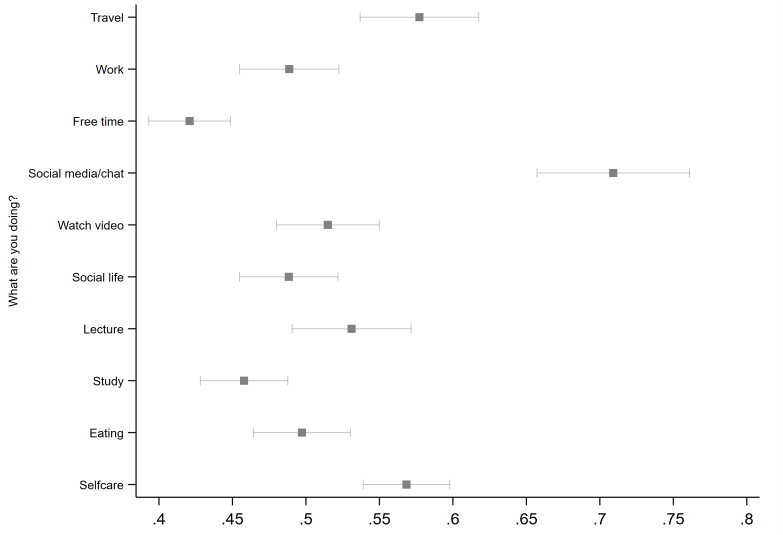


Figure 6.3: The statistical result of answering questions within 30 minutes at different activities.

Figure. 6.4. It is evident that personal relationships can influence the quality of answers. For instance, when participants stay Alone, they exhibit a higher likelihood (55%) of promptly answering questions. Conversely, when they are with a colleague, they demonstrate a lower likelihood (42%) of providing high-quality answers.

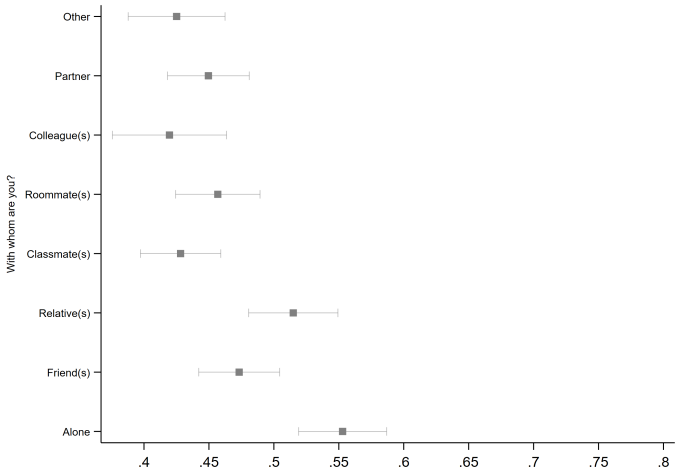


Figure 6.4: The statistical result of answering questions within 30 minutes at different interact persons.

Impact of Mood state on Answer Quality: Mood states were conceptualized as a five-tier scale, i.e., sad, a bit sad, neutral, a bit happy, and happy. The outcomes are depicted in Figure. 6.5. Our observations indicate a detrimental effect of emotional state on the mean probability of responding to queries within a 30-minute timeframe. This insinuates that individuals exhibit heightened attentiveness towards notifications when experiencing negative emotions. A contemporaneous study [192] also discovered an augmentation in reaction time as the user's emotional state ameliorates, thereby diminishing the quality of responses. The findings from this study corroborate this observation and concurrently underscore the existence of a causal relationship.

Impact of Individual difference on Answer Quality: Several studies have incorporated personality traits into the interruptibility model (see, e.g., [206]), yet

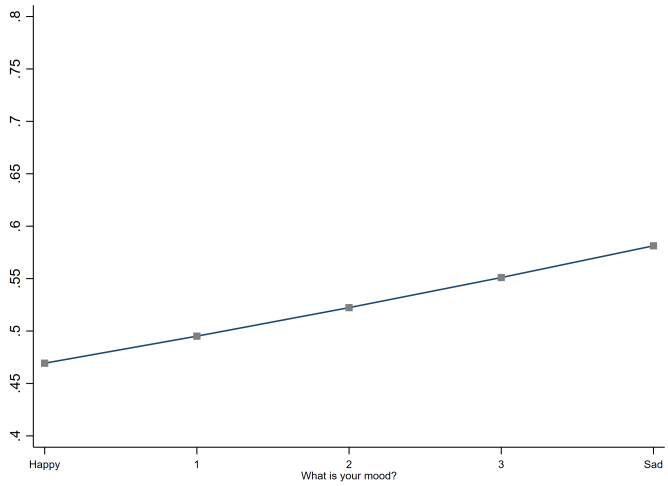


Figure 6.5: The statistical result of answering questions within 30 minutes at different moods.

Table 6.2: Multilevel logistic regression prediction results of personality traits.

Personality Traits	**W 1**	**W 2**
Extraversion	-0.004**	-0.006
Neuroticism	0.006**	0.005
Openness	-0.006**	0.000

they do not account for the influence of distinct personality traits on the quality of answers. In this part, we not only incorporate personality traits into the interruptibility model but also take into account human values, which offer a more granular categorization of individual differences. The results are presented in Table 6.2 and 6.3. As we treat personality traits and human values as continuous variables rather than discrete categories, they are differentiated by their high and low values. Upon evaluating the best fit of the interruptibility model, two personality traits and one value variable were omitted from the two tables due to their negligible impact on the likelihood of responding to questions within a 30-minute window.

Subsequently, we elucidate the significance of the parameters depicted in the following two tables. Here, "*" denotes p $<$0.1, indicating relatively weak statistical significance; "**" denotes p $<$0.05, indicating moderate statistical

Table 6.3: Multilevel logistic regression prediction results of human values.

Human Values	W 1	W 2
Conformity	0.172*	0.304**
Tradition	-0.080	-0.046
Benevolence	0.104	-0.166
Universalism	-0.259***	-0.328**
Self-direction	0.280***	0.423**
Stimulation	-0.078	-0.080
Hedonism	0.283***	0.361***
Achievement	-0.210***	-0.200*
Power	-0.190***	-0.269**

significance; "***" denotes p $<$0.001, indicating strong statistical significance. The absence of stars denotes p $>$0.1, indicating no statistical difference from the reference. W1 refers to the first two weeks and W2 refers to the second two weeks.

We introduce human values as an example to illustrate how to interpret these tables. As the human values are derived from a set of questions, we obtain precise values for each human values feature. Here, we demonstrate the improvement from the high value to the low value. Let us consider the coefficient of "Hedonism". Given that it is 0.283 in the first two weeks, we infer that at any time t, the probability of answering questions within 30 minutes at t at high Hedonism value is 1.327 times the low Hedonism value, with the odds value being $e^{0.283}$ = 1.327; Similarly, in the second two weeks, the high value of Hedonism is 1.435 times the likelihood of answering questions within 30 minutes compared to a person with a low value of Hedonism, implying that individuals with a high value of Hedonism consistently provide higher quality answers than those with a low value of Hedonism.

Answer to RQ1: Our analysis corroborates that contextual factors influence the quality of answers, aligning with state-of-the-art research on smartphone notifications [132]. For instance, our dataset reveals that individuals typically exhibit

high-quality answers during a negative mood. Between the hours of 7 AM and 9 AM. This is particularly evident when they are in a university setting, library, or at home, while using social media applications or during public transportation. Individuals are more inclined to respond to questions when spending time with family or when alone. Furthermore, our analysis results indicate that individual differences, specifically personality traits and human values, can significantly impact the answer behavior across different participants.

6.4 Answer quality variation by human difference

Next, we evaluate various models to identify the most significant features for prediction, thereby providing an answer to RQ2. In this part, we initially predict whether participants can respond to questions within a 30-minute window, which is indicative of their ability to provide high-quality answers. We opt for the most frequently employed classifiers in the Human and AI domain to extract information from the interruptibility model [11]. Specifically, we assess Random Forests (RF), K Nearest Neighbors (KNN), Logistic Regression, Support Vector Machines (SVM), Gaussian Naive Bayes, and the pbgc-Forest we introduced in Section 5.4. The classifiers were trained to use 5-fold cross-validation [111] on the comprehensive training and testing sets for all participants. We allocated 80% of the data to the training set and the remaining 20% to the testing set. We executed all experiments 5 times independently, adopting the median of the 5 outcomes as the final results presented in this section, to decide which classifier could be used for the further experiment. Subsequently, our objective is to identify which features can significantly influence the likelihood of promptly responding to a question.

In the above section, we conducted an analysis of how individual features influence the quality of answers. Subsequently, based on the outcomes derived from the explanatory model, we amalgamated all the features and aimed

to predict whether the participant could furnish high-quality responses under the given circumstances. To facilitate a comparison of classifiers, we opted to binary encode the quality of responses. Specifically, if the participant responds to questions within a 30-minute window, we designate this situation as conducive for posing questions as it enables us to procure high-quality responses. As depicted in Table 6.4, pbgc-Forest surpasses the other four classifiers and can attain a prediction accuracy of 0.764. Accuracy is defined as the ratio of correct predictions to the total number of predictions.

Due to the accuracy paradox [217], we also reported precision, recall, F1 score, Area Under the Curve (AUC), and kappa values. Precision, which measures the exactness of the classifier, is the ratio of true positive predictions to all predicted positive predictions. Recall measures the completeness of the classifier and is the ratio of true positive predictions to all positive class values in the data. The F1 score is a weighted average of precision and recall, measuring their balance. The Area Under the Curve (AUC), a performance metric for classification models, is derived from the Receiver Operating Characteristic (ROC) curve. The ROC curve is constructed by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The AUC value, which ranges between 0 and 1, serves as an indicator of the classifier's performance. An AUC value of 1 denotes a perfect classifier, while an AUC value of 0.5 implies that the classifier's performance is equivalent to random guessing. Kappa is a measure of agreement between a pair of variables and is commonly used as a metric of interrater agreement in a dataset. The findings suggest that the pbgc-Forest classifier outperforms other models, thereby making it the preferred choice for subsequent experiments. It achieves the highest accuracy and exhibits superior performance in terms of the AUC metric.

Next, in the subsequent phase, our objective is to evaluate the contribution of different features, e.g., situational context, personality traits, and human values. Consequently, as depicted in Table 6.5, we have six distinct models to ascertain

Table 6.4: Prediction results of different machine learning classifiers.

Classifier	Accuracy	Kappa	Precision	Recall	F1 score	AUC
pbgc-Forest	0.7637	0.4681	0.7215	0.7360	0.7147	0.7440
Random Forest	0.7283	0.4472	0.7092	0.7283	0.7200	0.7105
KNN	0.7111	0.4389	0.7023	0.7254	0.7229	0.6894
Logistic Regression	0.7017	0.3914	0.6595	0.7003	0.7002	0.5121
Gaussian Naive Bayes	0.6687	0.3405	0.6081	0.6687	0.6761	0.5143
SVM	0.7082	0.4097	0.6712	0.7082	0.7104	0.5485

Table 6.5: Set different model to test the most important features.

Model	Vars description
Model 1	This is the basic model. The variables used are: Weekday; Day of survey ; Hour of the day.
Model 2	The second model add to the Model 1 the following social demographic variables: Sex; Department; Degree, and Cohort.
Model 3	The third model add to the Model 2 the following variables from the time diary:What are you doing, Where are you, With whom are you, Mood.
Model 4	The fourth model add to Model 2 the personality traits and human values.
Model 5	Model fifth model combine Model 3 and Model 4 together.
Model 6	The last model, add to Model 5 the answer profile (i.e., the average of answer on massages).

the result. The initial model, Model 1, is the fundamental model, where we use basic variables, i.e., the day of the week, the day of the survey, and the hour of the day. The second model is predicated on Model 1 and incorporates social demographic values, that are, i.e., sex, department, degree, and user age cohort. Subsequently, Models 3 and 4 are the most crucial models. They are predicated on Model 2. Model 3 supplements Model 2 with situational context, inclusive of users' responses to time diary queries. Model 4 is also predicated on Model 2 and incorporates personality traits and human values as context. These two models will exhibit clear contribution results for understanding the situational context or personality with human values, which will be more beneficial in predicting whether participants will respond to questions within 30 minutes and, therefore, furnish high-quality answers. Subsequently, we amalgamate the different features in Models 3 and 4 into Model 5 to ascertain whether utilizing both situational context, personality, and human values will be more advantageous for prediction. Finally, Model 6 is predicated on Model 5 and incorporates the average response times of questions in each hour. We present the

Table 6.6: Prediction results of pbgc-Forest classifier in the train set.

Model	Accuracy train	Precision train	Recall train	F1 score train
Model 1	0.6513	0.6323	0.7776	0.6975
Model 2	0.6853	0.6726	0.7502	0.7093
Model 3	0.7388	0.7236	0.7749	0.7484
Model 4	0.7575	0.7355	0.7989	0.7659
Model 5	0.7746	0.7518	0.8084	0.7791
Model 6	0.7741	0.7527	0.8054	0.7782

Table 6.7: Prediction results of pbgc-Forest classifier in the test set.

Model	Accuracy test	Precision test	Recall test	F1 score test
Model 1	0.6486	0.6311	0.7723	0.6946
Model 2	0.6876	0.6816	0.7300	0.7050
Model 3	0.7435	0.7372	0.7571	0.7470
Model 4	0.7616	0.7426	0.7941	0.7675
Model 5	0.7810	0.7649	0.7983	0.7812
Model 6	0.7802	0.7632	0.7995	0.7809

results in Table 6.6 for the training set and Table 6.7 for the testing set. It is evident that the accuracy of Model 4 (0.7616) surpasses that of Model 3 (0.7435), which implies that personality traits coupled with human values exhibit superior prediction performance than situational context. Additionally, Model 5 (0.7810) demonstrates that utilizing situational context, personality traits, and human values collectively can achieve the optimal performance in predicting the quality of responses.

In summary, we used the pbgc-Forest model to achieve a prediction accuracy of 78.1%. When employing features such as personality traits, human values, social demographics, and time context, the model achieved a prediction accuracy of 76.2%. This surpasses the prediction accuracy of 74.3% achieved when employing features such as situational context, time context, and social demo-

graphic information. Our results suggest that human values and personality traits exert a significant influence on response behaviors and the reaction time of responding to questions and consequently impact the quality of responses.
Answer to RQ2: Our observations indicate that individual differences, specifically personality traits and human values, can exert a more pronounced influence on the prediction of answer quality than attributing significant weight to the situational context. Furthermore, we can procure optimal outcomes by amalgamating individual differences and situational context.

6.5 Answer quality variation by the number of daily questions

The number of questions is always designed by the experiment manager before the beginning of an experiment for better questioning, e.g., considering how many answers are needed from the participants. Experiment managers are willing to get plenty of answers from participants, so they would like to ask many questions per day [61]. Jeong et al. [106] indicate that an excessive quantity of questions can invariably precipitate survey fatigue, a state wherein participants experience weariness or diminished interest in the survey, thereby compromising the accuracy of answers. Consequently, the number of questions emerges as a factor influencing answer quality. Our aim is to send fewer questions to participants but get a higher percentage of high-quality answers. And we propose that the number of daily questions should be the number of high-quality answers that could be provided by most participants.

The number of daily questions is designed to correspond to the number of high-quality answers that most participants can provide each day. First, we calculate the average number of high-quality answers each participant provides. Next, participants are grouped based on their average number of high-quality answers, meaning two participants are in one group only if they have the same

average number of high-quality answers. Finally, we identify the group with the most participants. The number of daily questions is an integer closest to the average number of high-quality answers in this group.

Finding the number of daily questions: To clarify the annotation process of answer quality. We show the part of the pre-processed and anonymized (e.g., '1970-01-08') dataset, as shown in the *Day*, *Activity Answer*, *Notification Time* and *Answer Time* columns in Table 6.8, where values of *Activity Answer* are participants' answers to the question of 'What are you doing?'. Based on our workflow, we aim to calculate the reaction time and then annotate answer quality for each record in our WeNet dataset, as examples of *Reaction Time* and *Answer Quality* columns in Table 6.8.

Each record includes the value of *Notification Time*, namely the time when the participant's smartphone receives this question, and the value of *Answer Time*, namely when the participant starts to answer this question. Based on the Equation 4.1, we can calculate the reaction time for each record that has been answered. If a participant gives no answer to the question, we set *Reaction Time* values ∞. Secondly, after knowing the reaction time for all of the records in the WeNet dataset, based on Equation 4.3 and cut-off point $RT = 15\ minutes$, we can annotate the values of the *Answer Quality* column as '1' or '0', where '1' represents the high-quality answer, and '0' represents the low-quality answer.

Table 6.8: Example for a part of the annotated dataset.

Day	Activity Answer	Notification Time	Answer Time	**Reaction Time**	**Answer Quality**
1970-01-18	{"Eating"}	1970-01-18 10:05:09	1970-01-18 10:10:09	**5**	**1**
1970-01-18	null	1970-01-18 12:06:01	null	∞	**0**
1970-01-18	{"Sport"}	1970-01-18 14:03:08	1970-01-18 14:12:22	**9.14**	**1**
1970-01-19	null	1970-01-19 12:06:01	null	∞	**0**
1970-01-19	{"Studying"}	1970-01-19 13:06:06	1970-01-19 13:43:16	**37.1**	**0**

Result: After labeling answer quality into the WeNet dataset, we can con-

duct the statistical analysis on the quality of answers in the original WeNet dataset (which is the baseline of our evaluation in Section 6.5, as shown in Table 6.9). The 34.26% of questions were answered by high quality. In addition, on Monday and Saturday, participants give more high-quality answers (occupying 38.56% and 39.17%), while on Friday, participants give less high-quality answers (occupying 31.24%). It means we find that, at the beginning of weekdays/weekends, people always provide more high-quality answers.

Daily questions number: For each participant in the WeNet dataset, we calculate the average number of high-quality answers that she/he provided per day. These participants are grouped according to the average number. The results show that the biggest group includes 20 participants. They provided an average of 9.6 high-quality answers every day. The second biggest group includes 14 participants, who gave an average of 17.8 answers with high quality per day. However, for the smallest group, 4 participants averagely answered 2.4 questions with high-quality answers. Therefore, the daily question number should be the integer around the average number of high-quality answers of the biggest group, i.e., 10.

Result: Upon analyzing the number of high-quality answers each participant can contribute daily, we have elected to present ten questions per day. This decision aims to minimize any potential disruption to the participants' daily routines and maximize high-quality answers.

Evaluation: We conducted evaluation machine learning experiments using the most common classifiers: Random Forests, Decision Trees, Artificial Neural Networks, Logistic Regression, and Gaussian Naive Bayes. We proposed that by learning from participants' real context information and answer quality in the first two weeks, we could predict optimal times to ask questions in the following two weeks to obtain high-quality answers. During the second two weeks, we asked participants 10 questions daily (as determined in above) at the times suggested by the machine learning models.

Table 6.9 presents the percentage of high-quality answers from the machine learning experiments. The baseline percentage is calculated from the high-quality answers in the original WeNet dataset. We also show the percentages of high-quality answers based on the machine learning results for the second two weeks.

Table 6.9: Comparation of percentages of high-quality answers.

Weekday	BaseLine	Predictions Via Our Methodology				
		Random Forest	Decision Tree	Gaussian Naive Bayes	Artificial Neural Networks	Logistic Regression
Monday	38.56%	**64.03%**	63.78%	60.90%	63.30%	58.39%
Tuesday	36.80%	62.27%	**62.57%**	58.82%	61.91%	57.47%
Wednesday	36.74%	**63.08%**	60.08%	58.65%	61.18%	56.47%
Thursday	33.53%	60.17%	**61.46%**	57.35%	58.67%	56.02%
Friday	31.24%	**60.82%**	60.65%	57.51%	59.60%	56.86%
Saturday	39.17%	63.35%	**63.40%**	60.42%	61.56%	58.30%
Sunday	37.64%	**61.21%**	60.71%	56.30%	58.89%	57.45%
Overall	34.26%	**62.28%**	61.87%	58.56%	60.73%	57.28%

The baseline overall percentage is 34.26%. The overall percentage of high-quality answers predicted by the Random Forest classifier is the highest at 62.28%. The other rows break down these percentages by each day of the week. For each day, the percentages of high-quality answers predicted by the classifiers are higher than the baseline percentage, indicating improvement across all days. The highest percentage of high-quality answers is predicted by the Random Forest classifier on Monday (64.03%), while the lowest is predicted by the Logistic Regression classifier on Thursday (56.02%). All the algorithms we tested show significant improvement over the original baseline, supporting our methodology. This indicates that sending a suitable number of questions (fewer than before) based on participant context information can enhance answer quality. We infer that fewer questions and more specific contexts make participants feel less annoyed, leading them to provide higher-quality answers.

6.6 Discussion

The results of this study highlight the substantial influence of situational context, human differences, and the daily number of questions on the quality of responses to machine-generated inquiries. First, our analysis demonstrated that factors like time of day, location, activity, and social context are crucial in determining the likelihood of obtaining high-quality answers. For example, participants were more inclined to offer prompt and accurate responses during early morning hours, especially when they were at home or in a university setting, and engaged in activities such as using social media or public transportation.

Secondly, this study underscores the critical role of individual differences in predicting the quality of responses. Traits such as extraversion and neuroticism, alongside human values like hedonism and self-direction, were identified as key determinants of response behaviors. These findings indicate that integrating these individual differences into predictive models can improve the accuracy of forecasting high-quality responses. The predictive model developed in this study achieved an accuracy rate of 78.1%, illustrating the potential of combining situational context with individual differences to optimize the timing of question prompts. This method not only improves the quality of responses but also reduces the intrusiveness of notifications, thereby enhancing the user experience.

Finally, by analyzing a real-world dataset, we have identified that reducing the number of daily questions and considering contextual information can enhance the quality of data collected. Our methodology, which incorporates statistical analysis and machine learning techniques, offers a promising approach to optimizing mobile surveys for human-centered AI systems. The findings suggest that a thoughtful balance between the quantity of questions and the timing of their delivery can lead to more accurate and reliable data. Future research may explore the integration of additional contextual factors and the refinement

of predictive models to further improve the efficacy of data collection in AI applications. The key findings of this section can be drawn as:

- **Number of Questions:** The study finds that an excessive number of questions can lead to survey fatigue, where participants become weary or lose interest, resulting in low-quality answers. By determining an optimal number of daily questions, the researchers were able to increase the percentage of high-quality answers by up to 28%.
- **Contextual Factors:** Context plays a crucial role in the quality of answers. The study considers various contextual factors such as time of day, location, and activity. For instance, participants provided higher-quality answers when they were not engaged in activities that required their full attention, such as driving or attending meetings.

The findings of this study have significant implications for the fields of Human-Computer Interaction (HCI) and AI. By optimizing the number of questions and considering contextual factors, systems can collect higher quality data from users, which is essential for training accurate and reliable AI models. This approach can be applied to various applications, including:

- **Mobile Health:** In mobile health applications, collecting high-quality data from users is crucial for monitoring and managing health conditions. By reducing the number of questions and considering contextual factors, mobile health systems can improve user engagement and data quality.
- **Personalized Recommendations:** Personalized recommendation systems rely on high-quality user data to provide accurate and relevant suggestions. By optimizing the number of questions and considering context, these systems can enhance the user experience and increase the accuracy of recommendations.

- **Context-Aware Computing:** Context-aware systems can benefit from the findings of this study by incorporating contextual factors into their data collection processes. This can lead to more accurate and reliable context-aware applications, such as smart home systems and location-based services.

Chapter 7

A Configurable High-quality Big-Thick Data Collection Platform

7.1 Introduction

Understanding human behavior through context is pivotal for enhancing AI feedback mechanisms. As highlighted in Chapter 2, smartphones serve as invaluable tools for gathering detailed context-aware data, combining sensor inputs with context self-reporting questionnaires [74, 103, 159]. However, the quality of collected data often falls short of the needs for big-thick data, largely due to the high volume of questions posed to users, which can lead to low-quality answers [133]. Existing approaches relying on fixed or random schedules may not align well with participants' availability or willingness to respond [214, 26], reducing the effectiveness of human-AI collaboration.

In response, our research objective in this Chapter is *to develop a configurable platform that enables the collection of high-quality big-thick data from participants while ensuring minimal disturbance from the AI system.* Our method involves collecting big-thick data from both objective and subjective perspectives and monitoring the data collection process. For configurable data collection, we strive to offer participants flexibility. We extend the iCal (Inter-

net Calendar)[1] standard as a basis for a planning language as we discussed in Section 5.3. This is not a programming language for researchers to design questions repeatedly but uses the recurrence rules feature in iCal, allowing people to schedule recurring events in a calendar. Thus, we extend iCal as a planning language to configure question scheduling in our big-thick data collection process. This language models the collection of big-thick data based on different types of questions and recurrence rules, providing flexibility for researchers to design questions and sensor collections. It also offers participants enhanced flexibility in responding to questions. For monitoring, our system enables participants to visualize their collected data via a dashboard, providing real-time feedback on their data collection performance, lifestyle, and behavior. Additionally, we integrate a machine learning component into the system, allowing the AI to learn about each participant and reschedule context questions at times when the participant is available to answer, thus enhancing human-AI collaboration.

The platform presented in this Chapter represents an enhancement and extension of i-Log [212], which we described in the Section 3.3. We assess our platform through a case study within the *WeNet - The Internet of us project* [75]. The case study is designed to illustrate how the planner component of our platform can help researchers collect high-quality big-thick data from participants. The WeNet dataset has been used in numerous case studies, including [13, 138, 163]. In this Chapter, we use the WeNet dataset as an example to show that our platform can work, the detailed explanation of this project has been shown in Section 3.4.

[1] https://icalendar.org/

7.2 Overall platform

7.2.1 Logical architecture of the overall platform

Figure 7.1 depicts the overall platform architecture. The yellow rectangles represent the key components of the software systems: the ISAC (i-Log Service for Autonomous Configuration), calendar, dashboard, and planner, which facilitate artificial and intelligent data processing. The blue cylinders symbolize two databases. The first is the *Short Term Memory (STM)*, where the experiment plan, participant and planner modifications (committed plan), and aggregated data from the plan execution are stored. The second is the *Long Term Memory (LTM)*, where data (sensors and question responses) gathered during the execution of the plan is stored. Large volumes of data collected from the participants are transmitted from their devices to the LTM, which acts as permanent storage for the collected data and the details of the executed plan. The data visualized in the dashboard, a visualization component, is directly retrieved from the STM database.

The humanoid figures represent two roles of the personas: *Researchers* and *Participants*. The researcher is the individual seeking to collect context information from participants, while the participant provides their knowledge through the calendar and the big-thick data collection app. The arrow lines depict the interactions between the different components. The mobile phone symbolizes our data collection application, which collects data in the form of answers to questions sent and sensor data for constructing the user's situational context. The planner and calendar components are used in the planning and modification of the experiment plan, whereas the dashboard enhances the quality and representation of the data, allowing both participants and the researcher to track how participants are contributing during the experiment.

The data collection process starts with the researcher's planning phase. The researcher uses the ISAC component to design the initial plan for the timing

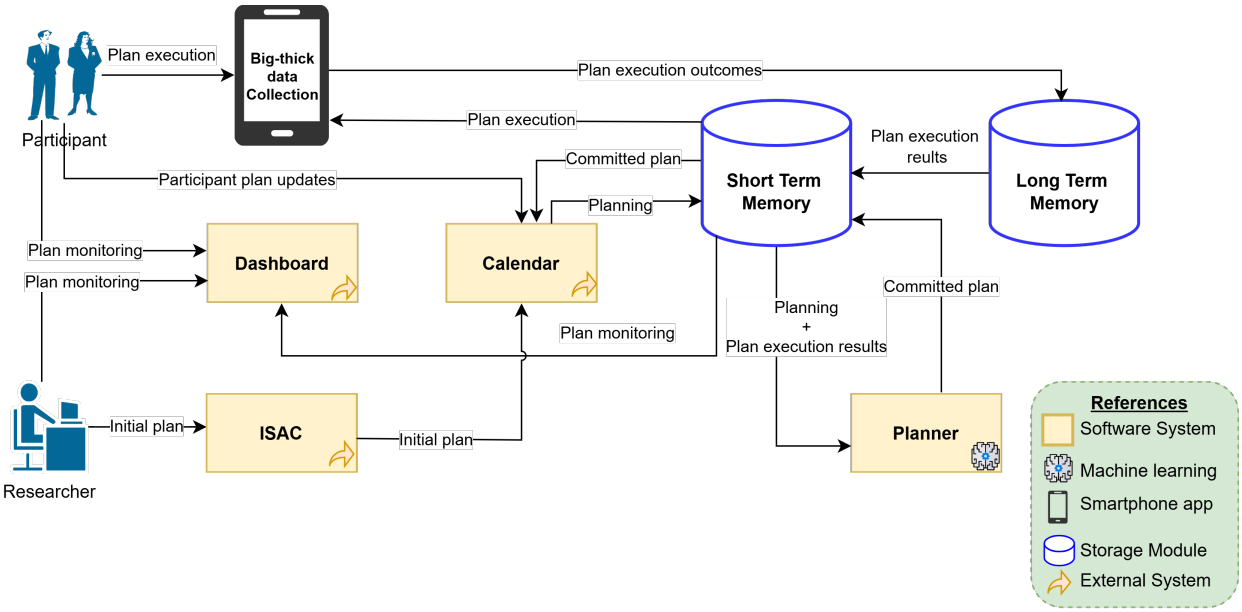


Figure 7.1: Overall platform architecture.

of context questions and schedule sensor data collection. This plan is first sent to the calendar, allowing participants to check when they should respond to questions. The plan is subsequently stored in the STM database as scheduled actions. When it is time to respond to questions, they are sent to the big-thick data collection application where participants can provide answers. If participants are unable to provide answers at the scheduled time, they can modify the time (change the plan) through the calendar component. Data collected during the execution of the plan is archived in the LTM database. Specific answers and sensor data are then sent to the STM database, aligning with the scheduled actions established by the researcher. Then, the plan execution results are sent to the planner component, which is a personalized question-answering system to learn from the plans from a participant and give the committed plan, which is a machine-generated plan that can help the researcher to get answers with high quality.

Furthermore, the planner component plays a crucial role in this architecture. It determines the optimal time to ask context questions based on the personal

context and re-plan modifications received from a specific participant. In the later section, we will explain all the key components in detail.

7.2.2 Description of components

In this section, we will explain the functioning of external systems, with running examples from the WeNet experiment for each component: the dashboard, ISAC, and calendar. For the planner, which represents the key contribution to our thesis, we will provide a detailed description in Section 7.3.

ISAC: The ISAC was designed to offer researchers services related to the configuration of data collections. This system aims to streamline communication between researchers conducting studies with i-Log and the i-Log experts responsible for deployment. As such, the system provides a set of guidelines and a structured pathway for researchers to convey the necessary information to the i-Log expert, ensuring that the i-Log configuration is accessible to them. Simultaneously, the ISAC seeks to enable researchers to design multiple questions and sensor collections automatically. This is particularly valuable because, in current ESM application-related work, researchers must design data collection methods one by one, leading to repetitive and time-consuming tasks. In ISAC, we are exploring the use of Recurrence Rules (RRules), as specified in the iCal standard, to simplify the design of repetitive tasks for researchers. The RRules can also help to make the configuration generate the visualization result in the calendar to let participants check when they should give feedback to the questions.

As depicted in Figure 7.2, the researcher uses the ISAC front-end module, an interface that enables the creation of a new plan for system deployment or a new plan template for repeated use in the future for the same data collection. The plan is then sent to the back-end for processing into the planning language compatible with the calendar system. The plan template is also stored in the

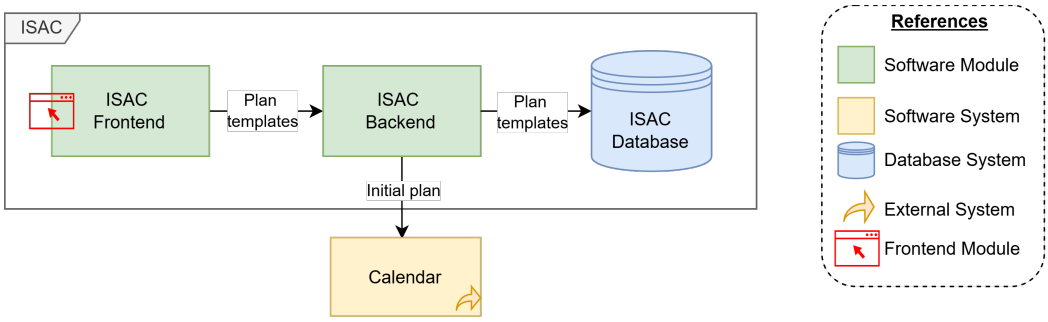


Figure 7.2: ISAC architecture.

ISAC database for future reuse.

Calendar: The calendar component is crucial for fostering interaction between researchers and participants. At the start of an experiment, the researcher creates the initial plan via the ISAC component for collecting context data and schedules sensor data collection. This plan is then sent to the calendar, allowing participants to see when they need to respond to questions. As the experiment progresses, participants' availability may change due to their busy schedules. To accommodate this, we enable participants to adjust their answer times and sensor data submissions via the calendar. This flexibility helps researchers better understand the participants, minimizes disruptions to their daily routines, and improves data quality. To support this flexibility, we designed a planning language by extending iCal, as we shown in Section 5.3, which is used in the calendar component to offer configurable options to both researchers and participants.

As illustrated in Figure 7.3, the calendar system comprises present and past modules that interact with the ISAC system to obtain the initial plan. This plan is displayed to the participant as the present plan, allowing them to modify the time if it is unsuitable for providing answers. The modified plan is then sent to the short term memory, where the planner uses it to learn the participant's answer behavior. The completed data collection process is shown as the past plan, enabling participants to review their responses, although no further changes can

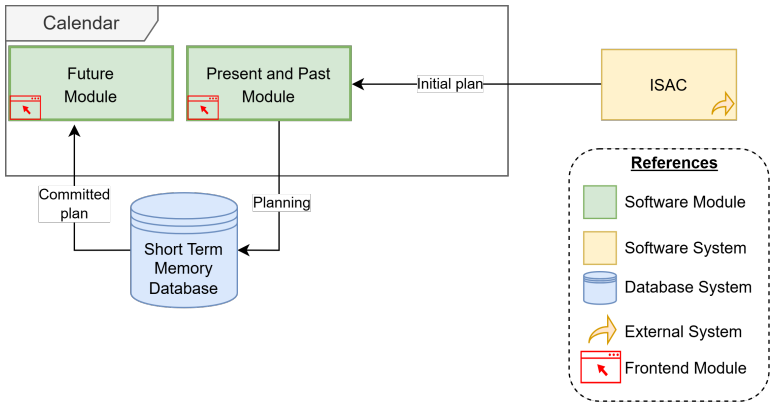


Figure 7.3: Calendar architecture.

be made. The future plan, derived from the short term memory database and informed by the planner, suggests optimal times for sending questions to the participant to obtain high-quality answers. Participants can also adjust the committed plan based on their real-world schedules.

Dashboard: The dashboard provides an interface for both researchers and participants to view and navigate the collected data. Users can filter the data using various queries, such as by sensor or date, to gain valuable insights about the data collected by the experiment. This not only helps participants become more aware of their lifestyle and potential areas for improvement (as their behavior is reflected in the data), but also allows them to track their progress in terms of data collection compared to other participants or benchmarks set by researchers. On the other hand, researchers gain a comprehensive understanding of the experimental data to ensure the data quality, including any issues that may have arisen, to ensure the smooth execution of the experiment.

As shown in Figure 7.4, the dashboard component comprises the dashboard database and a front-end module, known as the quality control module. This module allows the researcher to verify if the collected data meets the predefined quality standards set before the experiment. It also enables the identification of participants who consistently provide low-quality data, facilitating the diagnosis and resolution of any underlying issues. Additionally, the participant dash-

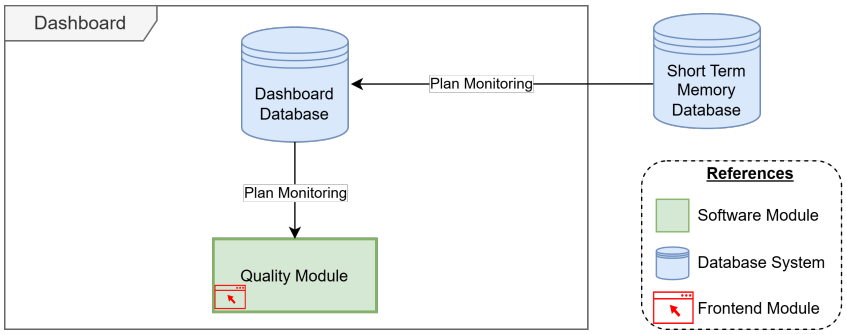


Figure 7.4: Dashboard architecture.

board has a distinct interaction interface compared to the researcher dashboard. While the researcher dashboard can assess the data quality for all participants, the participant dashboard is limited to viewing data specific to the individual participant.

7.3 Planner

The planner component is essential for maintaining effective human-AI collaboration. Researchers create a general experiment plan using the ISAC component, and participants can modify the plan according to their availability via the calendar. As outlined in Chapter 4, the time it takes for a participant to react to a question (reaction time) and the total time taken to answer the question (completion time) determine the quality of the answer. The planner component, supported by a machine learning module, predicts the most favorable moments to send questions to specific participants, ensuring the least reaction and response times and, consequently, high-quality information.

7.3.1 Logical architecture of the planner

As illustrated in Figure 7.5, the core component of the planner is a machine learning module designed to learn the answer behavior of individual participants. We developed a novel machine learning algorithm named pbgc-Forest,

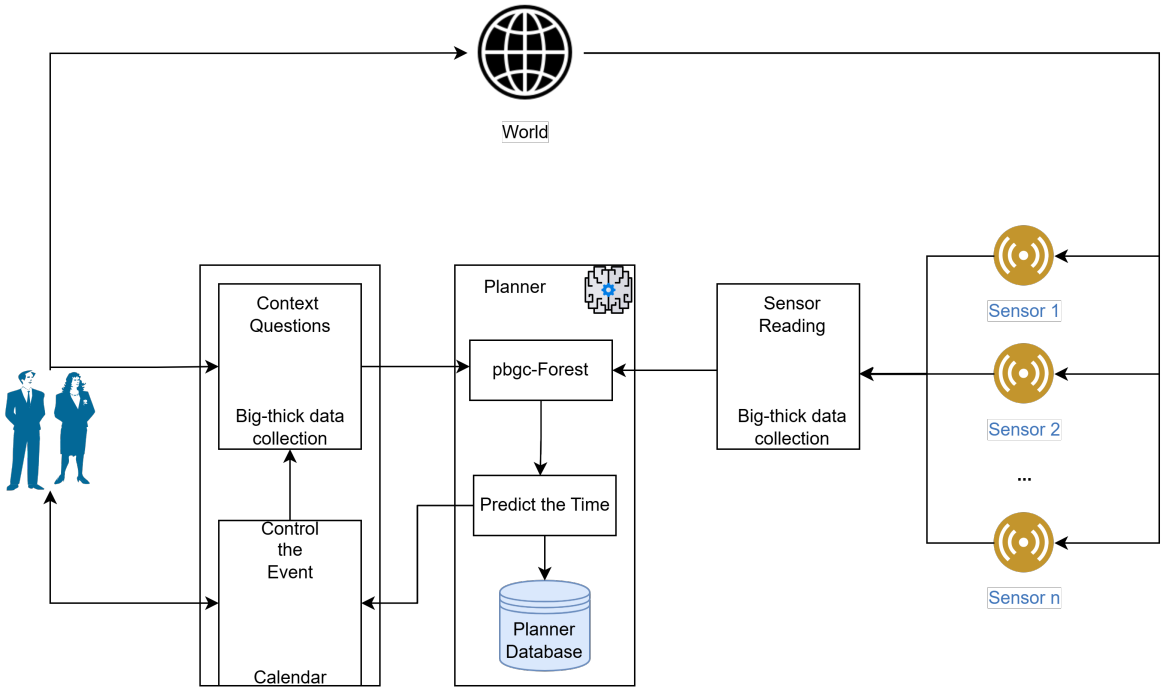


Figure 7.5: Planner architecture.

detailed in Section 5.4. The input data used to train the pbgc-Forest includes context information collected through our big-thick data collection application when participants answer our context questions. Additionally, we gather participant plan execution results, such as notification time, reaction time, and completion time for each question-answering process. The big-thick data collection application also collects sensor data, including time and situational context, as well as personal data such as demographic information, personality traits, and human values. The way participants handle events via the calendar, including how they change the time to answer questions, is also collected as input data to train the Planner. After learning the participant's answer behavior, the planner can predict suitable times when the participant is more likely to provide high-quality feedback. This new schedule is then shown on the calendar, allowing the participant to modify it as needed.

To ensure the planner minimizes participant annoyance, it can be config-

ured to send only 10 questions per day to each participant, as determined in Section 6.5. Furthermore, the planner will restrict question delivery to specific time frames—between 8:00 and 22:00—based on insights from Section 6.3. Within these parameters, the planner component can identify the 10 time intervals during which participants are most likely to provide prompt, high-quality responses. This approach enhances data quality by aligning question delivery with participants' optimal response times, minimizing disruptions, and maintaining participant engagement.

7.3.2 Evaluation

To analyze the data we collected from the WeNet experiment, which we introduced in Section 3.4, we constructed models using different classifiers: pbgcForest, Deep Forest, Random Forest, Logistic Regression, Gaussian Naive Bayes, and Artificial Neural Networks(ANN). We evaluated the model using a 5-fold cross-validation [111].

As part of our machine learning planner module, here we focus on a simple prediction label that can easily show the power of the planner component. That is, we send questions when the participant is engaged in activities other than studying or attending a lesson. We binary-encoded participants' activities, with studying alone or with others and attending a classroom lecture encoded as 1, while all other activities were encoded as 0. This activity information was retrieved from the time diaries questionnaire, specifically the question: "What are you doing?" as we described the situational context in Section 5.2. To train the machine learning classifiers, we used a variety of context factors collected from questionnaires (based on the questionnaire categories shown in Section 5.3, the BNF from Fig 5.4) and sensors (based on the time context). The detailed input features include: 1. Weekday, 2. Hour of the day, 3. Situational context, and 4. Mood, which is the same as the fractures we described in Section 6.2.

The result is shown in Table 7.1. The data were from 170 students who

Table 7.1: Prediction results of different machine learning classifiers.

Classifier	Accuracy	Kappa	Precision	Recall	F1 score	AUC
pbgc-Forest	0.7844	0.4236	0.5843	0.4002	0.4936	0.7762
Deep Forest	0.7629	0.4170	0.6102	0.4031	0.4843	0.7341
Random Forest	0.7292	0.3845	0.5900	0.3979	0.4752	0.7453
Decision Tree	0.6864	0.2458	0.5346	0.4053	0.4611	0.6695
Artificial Neural Networks	0.7003	0.2622	0.5702	0.3835	0.4546	0.7315
Logistic Regression	0.6648	0.3323	0.4545	0.6402	0.1122	0.6576
Gaussian Naive Bayes	0.6154	0.2161	0.4432	0.6323	0.5212	0.6548

answered all questions during the WeNet experiment, which was introduced in the Scetion 3.4. The pbgc-Forest classifier gave the highest accuracy of 78.4%, followed by the Deep Forest and Random Forest classifiers with accuracy of 76.3% and 72.9%, respectively. Due to the accuracy paradox [217], we also reported precision, recall, F1 score, Area Under the Curve (AUC), and kappa values. Generally speaking, the results show that the planner component can predict participants' academic time based on the different classifiers, and the pbgc-Forest classifier performs best. The planner component has the power to predict if the participant is present during the study activity and, in the future, to predict the human input data quality.

7.4 Discussion

The configurable platform for the collection of high-quality big-thick data presented in this Chapter demonstrates significant advancements in the field of data collection and human-machine collaboration. By integrating both big data and thick data, the platform offers a comprehensive approach to understanding human behavior and context. The fusion of objective sensor data with subjective human feedback provides a more nuanced and holistic view of user activities and motivations.

One of the key strengths of the platform is its flexibility and adaptability. The use of ISAC and calendar components allows researchers to design and schedule data collection events with ease. This flexibility is further enhanced by the re-planner component, which enables participants to adjust their response times, thereby minimizing disruptions to their daily routines and improving the quality of the collected data. The dashboard component provides valuable real-time feedback to both researchers and participants. Researchers can monitor the progress of the experiment, identify any issues, and make necessary adjustments to ensure data quality. Participants, on the other hand, can gain insights into their own behavior and performance, which can motivate them to engage more actively in the data collection process. The inclusion of a machine learning component in the planner is the key notable innovation. By predicting the most suitable times to send context questions based on participants' situational context and personal characteristics, the scheduler ensures that high-quality data is collected with minimal participant burden. The empirical evaluation of the scheduler within the WeNet project demonstrated its effectiveness in accurately predicting participants' study times, thereby avoiding interruptions during academic activities.

Despite these strengths, there are some limitations to the platform that need to be addressed in future work. The reliance on self-reported data for thick data collection can introduce biases and inaccuracies. Additionally, the platform's effectiveness in diverse real-world scenarios needs to be further evaluated. Future research should also explore the ethical and privacy implications of collecting and using big-thick data, ensuring that participants' rights and confidentiality are protected.

In conclusion, the configurable platform for high-quality big-thick data collection represents a significant step forward in the field of context-aware data collection and human-AI collaboration. By combining the strengths of big data and thick data, and leveraging advanced machine learning techniques, the plat-

form offers a powerful tool for researchers to gain deeper insights into human behavior and context.

Chapter 8

Conclusion

8.1 Summary

In this thesis, the primary objective is to design a platform that can automatically collect high-quality big-thick data. To achieve this, our work is divided into three parts for different personas: 1) Identifying key factors and data collection methodology for researchers: Firstly, we identify the key factors that can be easily observed during the data collection process to ensure the quality of human input data, and we design a personal high-quality data collection methodology that can get the context information from the participants. 2) Finding the suitable situational context and number of questions: Secondly, we demonstrate the three case studies that illustrate how our work can be used in real-world settings and provide valuable insights during the human input data collection process, contributing to the design of human-in-the-loop AI. 3) Platform Design: Finally, we design the platform using four key components to interact with participants for their context information. We show that the planner component can automatically find the moment that lets the participants ask questions with high-quality answers.

In Chapter 3, we demonstrate the two experiments conducted to collect students' big-thick data via smartphone sensors and questionnaires. For the smartphone application, we designed an app named i-Log, which allows us to collect

both objective sensor data and subjective human input data simultaneously. We used the i-Log app to conduct multiple experiments, with a focus on two in this thesis. The first experiment, named Smart Unitn 2, collected human input data from over 150 students over a 4-week period at the University of Trento, and 30 sensors were used in this experiment. The second experiment, named WeNet, also used data from the University of Trento, where we collected data from over 170 students over 4 weeks and 34 sensors were collected in this experiment.

In Chapter 4, where we solved the *problem 1* defined in Section 1.2, we determine the factor impacting human answer quality that can be easily observed during the data collection process, we focused on the participant's location as the way to determine the human input data quality, particularly regarding participants' homes. By knowing each participant's home GPS coordinates and the coordinates of the smartphone used, we calculated the distance between the smartphone and the home. If a participant claimed to be at home, but the distance exceeded 50 meters, we considered the answer incorrect. Using the total survey error paradigm from sociology, we hypothesized that reaction time and completion time could be key factors impacting human answer quality and are easily observable during the data collection process. Our key finding is the strong, direct, and indirect influence of reaction time on answer quality. Future work will focus on minimizing reaction time by identifying the optimal moments for asking questions.

In Chapter 5, where we solved the *problem 2* defined in Section 1.2, we identified the key factor influencing human answer data quality through their behavior. To achieve this, we first designed a situational context model encompassing individuals' daily lives, which introduces participants by five dimensions: 1. WHAT, the temporal context, known from the question: "What are you doing?" 2. WHERE, the spatial context, known from the question: "Where are you?" 3. WHO, the social context, known from the question: "Who is with you?" 4. WITHIN, the internal context, known from the question: "What mood are you

in?" 5. OBJECT, the object context, known from the question: "Which object are you with?". After identifying the situational context information required from participants, we focus on the temporal context, where we define a planning language; this language can offer researchers the flexibility to design recurring data collection processes and allow participants to choose convenient times for providing feedback. Finally, based on the context information and the answer behavior we learned from the participant, a novel machine learning algorithm named pbgc-Forest is used to predict the optimal times to send questions to participants, ensuring high-quality answers based on past context information.

In Chapter 6, where we solved the *problem 3* defined in Section 1.2, we provided the three case studies to demonstrate our ability to identify the situational contexts in which participants can provide high-quality answers. Drawing from the findings in Chapter 5, we used reaction time as a measure of answer quality, selecting a 30-minute cut-off point to distinguish between high-quality and low-quality answers. This decision was based on our analysis, which showed that the mean correct answer time was 38 minutes (see Figure 4.2), and to ensure that participants respond before receiving the next question, we chose 30 minutes as the cut-off. Our results for the first case study indicate that situational context significantly impacts answer quality. Participants' likelihood of providing high-quality answers varies across different contexts. For example, our dataset reveals that individuals typically provide high-quality answers when they are in a negative mood, between 7 AM and 9 AM. This is particularly evident when they are in a university setting, library, or at home, while using social media applications, during public transportation, or when spending time with family or alone. Secondly, our analysis indicates that individual differences, such as personality traits and human values, significantly influence answer behavior across participants. Specifically, these individual differences have a more pronounced impact on predicting answer quality than situational context alone. This case study highlights the importance of considering both situational con-

text and individual differences to improve the quality of human input data and enhance the overall effectiveness of our data collection process. In Section 6.5, we provide the third case study, demonstrating how to reduce the number of questions to improve answer quality. From the literature [14], we know that when a participant's main activity is disrupted by other factors, such as smartphone notifications, they need more time to complete the main task. Hence, we aim to avoid sending questions when participants are focused on their main tasks, while maximizing the number of questions sent to gather more data. Our findings show that participants can answer up to 10 questions with high quality per day. Additionally, reducing the number of questions sent from 48 to 10 per day improves the percentage of high-quality answers by 28%.

In Chapter 7, where we solved the *problem 4* defined in Section 1.2, we design a platform that can automatically collect high-quality big-thick data from participants. To achieve this, we apply the knowledge learned from Chapter 5, which shows that lower reaction times lead to higher human answer quality. Thus, our aim is to enable participants to respond quickly. Our novel platform's architecture comprises four main components. The ISAC component is geared towards researchers, allowing them to design the initial plan for the data collection process. Calendar Component: This component is primarily for participants, enabling them to check when they need to provide answers to context questions. It also allows them to reschedule their response times, thus minimizing disruptions to their daily lives. Dashboard Component: This visualization tool enables both researchers and participants to monitor data quality during the data collection process. Planner Component: This crucial component understands human daily routines and answers behaviors. Also, the planner component is the core of this thesis, which shows that our platform has the ability to predict the suitable moment when the participant can answer the questions with high-quality answers. By integrating these components, our platform aims to efficiently collect high-quality big-thick data while minimizing the participant's

feeling annoyed.

8.2 Discussion and limitation

This thesis primarily addresses the key problem of designing a novel platform that can automatically collect high-quality big-thick data. The first contribution lies in identifying the key factor that can be easily observed during the data collection process: reaction time. This factor helps researchers gain an initial understanding of answer quality. During our research, we used the distance between the user's smartphone and home when they answered questions and claimed to be at home. If the distance exceeded 50 meters, the answer was considered incorrect. However, we did not account for potential technological issues, such as the accuracy of the provided GPS coordinates, which may not cover the entire home, leading to mistakes. Additionally, we only used a single dataset — Smart Unitn 2 — which may result in inaccuracies due to data collinearity.

After establishing that reaction time is the key factor influencing human answer quality, the second contribution is that we designed a platform to automatically collect sensor and questionnaire data from participants. During this process, we introduced a planning language that provides participants with the flexibility to change the times they provide data via a calendar component. However, we only used existing datasets rather than conducting new in-person experiments to collect fresh data. The planning language, an extension of the iCalendar standard, may also require a further extension for future use.

In the case study section, our third contribution is that we show the way how to find a suitable situational context to avoid annoying the participant and show a possible number of questions to send to a participant. We conducted two case studies using existing datasets. In future work, we should conduct new experiments to test the platform's capabilities and the methodology's effectiveness.

Additionally, we used common machine learning algorithms for the predictive experiments, rather than designing novel algorithms specifically for our analysis process. Developing tailored algorithms may yield better performance and enhance our ability to predict human daily routines and answer behavior.

8.3 Future work

Looking ahead, an important aspect that influences the reliability of the data collected is the incentives for participation, as we discussed at the end of Chapter 5. Mechanisms that simplify participants' activities while enhancing their involvement can be the first step toward improving the quality of the data we collect. Secondly, to ensure data quality, we can incorporate attention checker questions and straightlining detection methods during the data collection process. These techniques can provide us with new ways to ensure the data quality and then assess whether reaction time and completion time remain key factors impacting human answer data quality.

From the platform perspective, we should conduct real-world experiments to verify whether the planner can help identify the optimal moments when participants are willing to provide high-quality answers and whether the planning language can effectively bridge the ISAC and calendar components. Additionally, considering the Planning Domain Definition Language (PDDL) as an updated language can enhance the planning language's flexibility, making research more reusable and easily comparable across different AI areas. We should also explore novel ways to improve the pbgc-Forest algorithm, making the planner more adept at predicting human answer behavior quickly and accurately.

By implementing these improvements, we can analyze the collected data through explainable and predictable methods to understand why participants may not always provide data. Subsequently, we can refine our approach using active learning and reinforcement learning methods to achieve better per-

formance, ultimately leading to a lifelong human-in-the-loop study. This will enable machines to assist humans more effectively.

Bibliography

[1] Zahraa S Abdallah, Mohamed Medhat Gaber, Bala Srinivasan, and Shonali Krishnaswamy. Activity recognition with evolving data streams: A review. *ACM Computing Surveys (CSUR)*, 51(4):1–36, 2018.

[2] Marwa Abdulhai, Dong-Ki Kim, Matthew Riemer, Miao Liu, Gerald Tesauro, and Jonathan P How. Context-specific representation abstraction for deep option learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5959–5967, 2022.

[3] Karl Aberer, Michele Catasta, Horia Radu, Jean-Eudes Ranvier, Matteo Vasirani, and Zhixian Yan. Memorysense: Reconstructing and ranking user memories on mobile devices. In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*, pages 195–198. Ieee, 2014.

[4] Gregory D Abowd, Anind K Dey, Peter J Brown, Nigel Davies, Mark Smith, and Pete Steggles. Towards a better understanding of context and context-awareness. In *International symposium on handheld and ubiquitous computing*, pages 304–307. Springer, 1999.

[5] Laura Alessandretti, Ulf Aslak, and Sune Lehmann. The scales of human mobility. *Nature*, 587(7834):402–407, 2020.

[6] Laura Alessandretti, Piotr Sapiezynski, Vedran Sekara, Sune Lehmann, and Andrea Baronchelli. Evidence for a conserved quantity in human mobility. *Nature human behaviour*, 2(7):485–491, 2018.

[7] Paul D Allison. Event history and survival analysis. In *The reviewer's guide to quantitative methods in the social sciences*, pages 86–97. Routledge, 2018.

[8] Omri Alon, Sharon Rabinovich, Chana Fyodorov, and Jessica R Cauchard. First step toward gestural recognition in harsh environments. *Sensors*, 21(12):3997, 2021.

[9] Ashley Amaya, Paul P Biemer, and David Kinyon. Total error in a big data world: adapting the tse framework to big data. *Journal of Survey Statistics and Methodology*, 8(1):89–119, 2020.

[10] Samaneh Aminikhanghahi, Maureen Schmitter-Edgecombe, and Diane J Cook. Context-aware delivery of ecological momentary assessment. *IEEE journal of biomedical and health informatics*, 24(4):1206–1214, 2019.

[11] Christoph Anderson, Judith Simone Heinisch, Shohreh Deldari, Flora Salim, Sandra Ohly, Klaus David, and Veljko Pejović. Toward social role-based interruptibility management. *IEEE Pervasive Computing*, 22(1):59–68, 2023.

[12] Jaime Arellano-Bover, Carolina Bussotti, John M Nunley, and Alan Seals. Unbundling the effects of college on first-job search: Returns to majors, minors, internships, study abroad, and computer skills. *Minors, Internships, Study Abroad, and Computer Skills*, 2024.

[13] Karim Assi, Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Miriam Bidoglia, Sally Stares, George Gaskell, Altangerel

Chagnaa, Amarsanaa Ganbold, et al. Complex daily activities, country-level diversity, and smartphone sensing: A study in denmark, italy, mongolia, paraguay, and uk. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–23, 2023.

[14] Brian P Bailey, Joseph A Konstan, and John V Carlis. Measuring the effects of interruptions on task performance in the user interface. In *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics.'cybernetics evolving to systems, humans, organizations, and their complex interactions'(cat. no. 0*, volume 2, pages 757–762. IEEE, 2000.

[15] Julia M Balto, Dominique L Kinnett-Hopkins, and Robert W Motl. Accuracy and precision of smartphone applications and commercially available motion sensors in multiple sclerosis. *Multiple Sclerosis Journal–Experimental, Translational and Clinical*, 2:2055217316634754, 2016.

[16] Wageesha Bangamuarachchi, Anju Chamantha, Lakmal Meegahapola, Salvador Ruiz-Correa, Indika Perera, and Daniel Gatica-Perez. Sensing eating events in context: A smartphone-only approach. *IEEE Access*, 10:61249–61264, 2022.

[17] Gianni Barlacchi, Christos Perentis, Abhinav Mehrotra, Mirco Musolesi, and Bruno Lepri. Are you getting sick? predicting influenza-like symptoms using human mobility behaviors. *EPJ data science*, 6:1–15, 2017.

[18] Sian L Beilock and Marci S DeCaro. From poor performance to success under stress: working memory, strategy selection, and mathematical problem solving under pressure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(6):983, 2007.

[19] Dror Ben-Zeev, Rachel Brian, Rui Wang, Weichen Wang, Andrew T Campbell, Min SH Aung, Michael Merrill, Vincent WS Tseng, Tanzeem

Choudhury, Marta Hauser, et al. Crosscheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. *Psychiatric rehabilitation journal*, 40(3):266, 2017.

[20] Ethan M Berke, Tanzeem Choudhury, Shahid Ali, and Mashfiqui Rabbi. Objective measurement of sociability and activity: mobile sensing in the community. *The Annals of Family Medicine*, 9(4):344–350, 2011.

[21] Claudio Bettini, Oliver Brdiczka, Karen Henricksen, Jadwiga Indulska, Daniela Nicklas, Anand Ranganathan, and Daniele Riboni. A survey of context modelling and reasoning techniques. *Pervasive and mobile computing*, 6(2):161–180, 2010.

[22] Norman Blaikie and Jan Priest. *Designing social research: The logic of anticipation*. John Wiley & Sons, 2019.

[23] G etc. Blank. Online research methods and social theory. *The SAGE Handbook of Online Research Methods*, page 537–549, 2008.

[24] Peter M Blau and Otis Dudley Duncan. The american occupational structure. 1967.

[25] Benjamin M Bolker, Mollie E Brooks, Connie J Clark, Shane W Geange, John R Poulsen, M Henry H Stevens, and Jada-Simone S White. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3):127–135, 2009.

[26] Andrea Bontempelli, Marcelo Rodas Britez, Xiaoyue Li, Haonan Zhao, Luca Erculiani, Stefano Teso, Andrea Passerini, and Fausto Giunchiglia. Lifelong personal context recognition, 2022.

[27] Tobias Bornakke and Brian L Due. Big–thick blending: A method for mixing analytical insights from big and thick data sources. *Big Data & Society*, 5(1):2053951718765026, 2018.

[28] Mehdi Boukhechba, Lihua Cai, Philip I Chow, Karl Fua, Matthew S Gerber, Bethany A Teachman, and Laura E Barnes. Contextual analysis to understand compliance with smartphone-based ecological momentary assessment. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 232–238, 2018.

[29] Raymond V Bowers. Time budgets of human behavior., 1939.

[30] Danah Boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.

[31] Norman M Bradburn, Lance J Rips, and Steven K Shevell. Answering autobiographical questions: The impact of memory and inference on surveys. *Science*, 236(4798):157–161, 1987.

[32] Norman M Bradburn, Seymour Sudman, and Brian Wansink. *Asking questions: the definitive guide to questionnaire design–for market research, political polls, and social and health questionnaires*. John Wiley & Sons, 2004.

[33] Nicholas A Bradley and Mark D Dunlop. Toward a multidisciplinary model of context to support context-aware computing. *Human-Computer Interaction*, 20(4):403–446, 2005.

[34] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[35] Urie Bronfenbrenner. Toward an experimental ecology of human development. *American psychologist*, 32(7):513, 1977.

[36] Peter J Brown, John D Bovey, and Xian Chen. Context-aware applications: from the laboratory to the marketplace. *IEEE personal communications*, 4(5):58–64, 1997.

[37] Erhan Bulbul, Aydin Cetin, and Ibrahim Alper Dogru. Human activity recognition using smartphones. In *2018 2nd international symposium on multidisciplinary studies and innovative technologies (ismsit)*, pages 1–6. IEEE, 2018.

[38] M. Callegaro, K. L. Manfreda, and V. Vehovar. *Web survey methodology*. Sage, 2015.

[39] Charles F Cannell, Peter V Miller, and Lois Oksenberg. Research on interviewing techniques. *Sociological methodology*, 12:389–437, 1981.

[40] Guanling Chen and David Kotz. A survey of context-aware mobile computing research. 2000.

[41] Han-Ching Chen and Nae-Sheng Wang. The assignment of scores procedure for ordinal categorical data. *The Scientific World Journal*, 2014, 2014.

[42] Hsinchun Chen, Roger HL Chiang, and Veda C Storey. Business intelligence and analytics: From big data to big impact. *MIS quarterly*, pages 1165–1188, 2012.

[43] Lee Anna Clark and David Watson. Mood and the mundane: relations between daily life events and self-reported mood. *Journal of personality and social psychology*, 54(2):296, 1988.

[44] W Scott Comulada. Calculating level-specific sem fit indices for multilevel mediation analyses. *The Stata Journal*, 21(1):195–205, 2021.

[45] John H Connolly, Alan Chamberlain, and Iain W Phillips. An approach to context in human-computer interaction. 2008.

[46] P. Corbetta. *Social research: Theory, methods and techniques*. Sage, 2003.

[47] Nelson Cowan. The magical mystery four: How is working memory capacity limited, and why? *Current directions in psychological science*, 19(1):51–57, 2010.

[48] Nelson Cowan. *Working memory capacity*. Psychology press, 2012.

[49] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[50] M Csikszentmihalyi. *The experience of psychopathology: Investigating mental disorders in their natural settings*. Cambridge University Press, 1992.

[51] Andrea Cuttone, Sune Lehmann, and Marta C González. Understanding predictability and exploration in human mobility. *EPJ Data Science*, 7:1–17, 2018.

[52] Tushar Kant Das and P Mohan Kumar. Big data analytics: A framework for unstructured data analysis. *International Journal of Engineering Science & Technology*, 5(1):153, 2013.

[53] Thomas H Davenport and Jill Dyché. Big data in big companies. *International Institute for Analytics*, 3(1-31), 2013.

[54] Florenc Demrozi, Graziano Pravadelli, Azra Bihorac, and Parisa Rashidi. Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey. *IEEE access*, 8:210816–210836, 2020.

[55] Anind K Dey, Gregory D Abowd, and Andrew Wood. Cyberdesk: A framework for providing self-integrating context-aware services. *Knowledge-based systems*, 11(1):3–13, 1998.

[56] Trinh Minh Tri Do and Daniel Gatica-Perez. Contextual conditional models for smartphone-based human mobility prediction. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 163–172, 2012.

[57] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29(2):103–130, 1997.

[58] M Brent Donnellan, Frederick L Oswald, Brendan M Baird, and Richard E Lucas. The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, 18(2):192, 2006.

[59] Nathan Eagle, Alex Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 106(36):15274–15278, 2009.

[60] Nathan Eagle and Alex Sandy Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral ecology and sociobiology*, 63(7):1057–1066, 2009.

[61] Gudrun Eisele, Hugo Vachon, Ginette Lafit, Peter Kuppens, Marlies Houben, Inez Myin-Germeys, and Wolfgang Viechtbauer. The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, 29(2):136–151, 2022.

[62] Miikka Ermes, Juha Pärkkä, Jani Mäntyjärvi, and Ilkka Korhonen. Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. *IEEE transactions on information technology in biomedicine*, 12(1):20–26, 2008.

[63] Denzil Ferreira, Vassilis Kostakos, and Anind K Dey. Aware: mobile context instrumentation framework. *Frontiers in ICT*, 2:6, 2015.

[64] Nigel G Fielding, Grant Blank, and Raymond M Lee. The sage handbook of online research methods. 2016.

[65] Joel E Fischer, Chris Greenhalgh, and Steve Benford. Investigating episodes of mobile phone activity as indicators of opportune moments to deliver notifications. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*, pages 181–190, 2011.

[66] Joel E Fischer, Nick Yee, Victoria Bellotti, Nathan Good, Steve Benford, and Chris Greenhalgh. Effects of content and time of delivery on receptivity to mobile interruptions. In *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*, pages 103–112, 2010.

[67] S Fisher and L Kydoniefs. Using a theoretical model of response burden (rb) to identify sources of burden in surveys. In *12th International Workshop on Household Survey Nonresponse, Oslo, Norway, September*, pages 12–14, 2001.

[68] AAPOR Cell Phone Task Force. New considerations for survey researchers when planning and conducting rdd telephone surveys in the us with respondents reached via cell phone numbers. *Deerfield, IL: American Association for Public Opinion Research*, 2010.

[69] Joseph P Forgas. On being happy and mistaken: mood effects on the fundamental attribution error. *Journal of personality and social psychology*, 75(2):318, 1998.

[70] Joseph P Forgas. Don't worry, be sad! on the cognitive, motivational, and interpersonal benefits of negative mood. *Current Directions in Psychological Science*, 22(3):225–232, 2013.

[71] Joseph C Franklin, Jessica D Ribeiro, Kathryn R Fox, Kate H Bentley, Evan M Kleiman, Xieyining Huang, Katherine M Musacchio, Adam C Jaroszewski, Bernard P Chang, and Matthew K Nock. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological bulletin*, 143(2):187, 2017.

[72] Clifford Geertz. Thick description: Toward an interpretive theory of culture. In *The cultural geography reader*, pages 41–51. Routledge, 2008.

[73] F. Giunchiglia. Contextual reasoning. *Epistemologia, special issue: I Linguaggi e le Macchine*, 16:345–364, 1993.

[74] Fausto Giunchiglia, Enrico Bignotti, and Mattia Zeni. Personal context modelling and annotation. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 117–122. IEEE, 2017.

[75] Fausto Giunchiglia, Ivano Bison, Matteo Busso, Ronald Chenu-Abente, Marcelo Rodas, Mattia Zeni, Can Gunel, Giuseppe Veltri, Amalia De Götzen, Peter Kun, et al. A worldwide diversity pilot on daily routines and social practices (2020). 2021.

[76] Fausto Giunchiglia, Matteo Busso, Mattia Zeni, Ivano Bison, et al. A survey on students' daily routines and academic performance at the university of trento. 2022.

[77] Fausto Giunchiglia and Mattia Fumagalli. Teleologies: Objects, actions and functions. In *ER- International Conference on Conceptual Modeling*, pages 520–534. ICCM, 2017.

[78] Fausto Giunchiglia and Daqian Shi. Property-based entity type graph matching. *arXiv preprint arXiv:2109.09140*, 2021.

[79] Fausto Giunchiglia, Alessio Zamboni, Mayukh Bagchi, and Simone Bocca. Stratified data integration. In *2nd International Workshop On Knowledge Graph Construction (KGCW), Co-located with the ESWC 2021*, Hersonissos, Greece, 2021.

[80] Fausto Giunchiglia, Mattia Zeni, Enrico Bignotti, and Wanyi Zhang. Assessing annotation consistency in the wild. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 561–566. IEEE, 2018.

[81] Fausto Giunchiglia, Mattia Zeni, Elisa Gobbi, Enrico Bignotti, and Ivano Bison. Mobile social media and academic performance. In *International conference on social informatics*, pages 3–13. Springer, Cham, 2017.

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

[83] Harvey Goldstein. *Multilevel statistical models*. John Wiley, New York, 2011.

[84] Alejandra Gomez Ortega, Janne Van Kollenburg, Yvette Shen, Dave Murray-Rust, Dajana Nedić, Juan Carlos Jimenez, Wo Meijer, Pranshu Kumar Kumara Chaudhary, and Jacky Bourgeois. Sig on data as human-centered design material. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–4, 2022.

[85] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779–782, 2008.

[86] Andreas Grammenos, Cecilia Mascolo, and Jon Crowcroft. You are sensing, but are you biased? a user unaided sensor calibration approach for mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–26, 2018.

[87] Robert M Groves, Robert B Cialdini, and Mick P Couper. Understanding the decision to participate in a survey. *Public opinion quarterly*, 56(4):475–495, 1992.

[88] Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. *Survey methodology*. John Wiley & Sons, 2011.

[89] Juan Guerrero-Ibáñez, Sherali Zeadally, and Juan Contreras-Castillo. Sensor technologies for intelligent transportation systems. *Sensors*, 18(4):1212, 2018.

[90] Gabriella M Harari, Nicholas D Lane, Rui Wang, Benjamin S Crosier, Andrew T Campbell, and Samuel D Gosling. Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, 11(6):838–854, 2016.

[91] Joel M Hektner, Jennifer A Schmidt, and Mihaly Csikszentmihalyi. *Experience sampling method: Measuring the quality of everyday life*. Sage, 2007.

[92] Rim Helaoui, Daniele Riboni, and Heiner Stuckenschmidt. A probabilistic ontological framework for the recognition of multilevel human activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 345–354, 2013.

[93] Kristin E Heron, Robin S Everhart, Susan M McHale, and Joshua M Smyth. Using mobile-technology-based ecological momentary assessment (ema) methods with youth: A systematic review and recommendations. *Journal of pediatric psychology*, 42(10):1087–1107, 2017.

[94] Joyce Ho and Stephen S Intille. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 909–918, 2005.

[95] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[96] Karen Holtzblatt and Hugh Beyer. *Contextual design: defining customer-centered systems*. Elsevier, 1997.

[97] Andy Hong, Lucy Baker, Rafael Prieto Curiel, James Duminy, Bhawani Buswala, ChengHe Guan, and Divya Ravindranath. Reconciling big data and thick data to advance the new urban science and smart city governance. *Journal of Urban Affairs*, 45(10):1737–1761, 2023.

[98] Stefan E Hormuth. The sampling of experiences in situ. *Journal of personality*, 54(1):262–293, 1986.

[99] Joop J Hox et al. Multilevel regression and multilevel structural equation modeling. *The Oxford handbook of quantitative methods*, 2(1):281–294, 2013.

[100] Yu Huang, Haoyi Xiong, Kevin Leach, Yuyan Zhang, Philip Chow, Karl Fua, Bethany A Teachman, and Laura E Barnes. Assessing social anxiety using gps trajectories and point-of-interest data. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 898–903, 2016.

[101] Ira E Hyman Jr, S Matthew Boss, Breanne M Wise, Kira E McKenzie, and Jenna M Caggiano. Did you see the unicycling clown? inattentional blindness while walking and talking on a cell phone. *Applied Cognitive Psychology*, 24(5):597–607, 2010.

[102] Stephen Intille. The precision medicine initiative and pervasive health research. *IEEE Pervasive Computing*, 15(1):88–91, 2016.

[103] Stephen S Intille, John Rondoni, Charles Kukla, Isabel Ancona, and Ling Bao. A context-aware experience sampling tool. In *CHI'03 extended abstracts on Human factors in computing systems*, pages 972–973, 2003.

[104] Shamsi T Iqbal and Eric Horvitz. Notifications and awareness: a field study of alert usage and preferences. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 27–30, 2010.

[105] Annette Jäckle, Jonathan Burton, Mick P Couper, and Carli Lessof. Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: Response rates and response biases. *Institute for Social and Economic Research, University of Essex: Understanding Society Working Paper Series No*, 9, 2017.

[106] Dahyeon Jeong, Shilpa Aggarwal, Jonathan Robinson, Naresh Kumar, Alan Spearot, and David Sungho Park. Exhaustive or exhausting? evidence on respondent fatigue in long surveys. *Journal of Development Economics*, 161:102992, 2023.

[107] Karl G Joreskog, Sumorbom Dag, and Jay Magidson. *Advances in factor analysis and structural equation models*. Abt books, 1979.

[108] Ronald T Kellogg. *Fundamentals of cognitive psychology*. Sage Publications, 2015.

[109] Florian Keusch, Bella Struminskaya, Christopher Antoun, Mick P Couper, and Frauke Kreuter. Willingness to participate in passive mobile data collection. *Public opinion quarterly*, 83(S1):210–235, 2019.

[110] William D Scott Killgore. The visual analogue mood scale: can a single-item scale accurately classify depressive mood state? *Psychological reports*, 85(3_suppl):1238–1243, 1999.

[111] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[112] Frauke Kreuter, Georg-Christoph Haas, Florian Keusch, Sebastian Bähr, and Mark Trappmann. Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent. *Social Science Computer Review*, 38(5):533–549, 2020.

[113] Jon A Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3):213–236, 1991.

[114] Jon A Krosnick, Duane F Alwin, Charles Cannell, et al. Satisficing: A strategy for dealing with the demands of survey questions. 1987.

[115] Kostadin Kushlev, Jason Proulx, and Elizabeth W Dunn. " silence your phones" smartphone notifications increase inattention and hyperactivity symptoms. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 1011–1020, 2016.

[116] Jennie Lai, Lorelle Vanno, Michael Link, Jennie Pearson, Hala Makowska, Karen Benezra, and Mark Green. Life360: usability of mobile devices for time use surveys. In *American Association for Public*

Opinion Research annual conference, Hollywood, FL, pages 5582–5589, 2009.

[117] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. A survey of mobile phone sensing. *IEEE Communications magazine*, 48(9):140–150, 2010.

[118] Reed Larson and Mihaly Csikszentmihalyi. The experience sampling method. In *Flow and the foundations of positive psychology*, pages 21–34. Springer, 2014.

[119] A Latif, AI Choudhary, and AA Hammayun. Economic effects of student dropouts: A comparative study. *Journal of global economics*, 3(2):1–4, 2015.

[120] Paul J Lavrakas, Charles D Shuttles, Charlotte Steeh, and Howard Fienberg. The state of surveying cell phone numbers in the united states: 2007 and beyond. *Public Opinion Quarterly*, 71(5):840–854, 2007.

[121] Paul J Lavrakas, Trevor N Tompson, Robert Benford, and Christopher Fleury. Investigating data quality in cell phone surveying. In *annual American Association for Public Opinion Research conference, Chicago, Illinois*, pages 13–16, 2010.

[122] Matthew L Lee and Anind K Dey. Sensor-based observations of daily living for aging in place. *Personal and Ubiquitous Computing*, 19(1):27–43, 2015.

[123] Michael W Link, Joe Murphy, Michael F Schober, Trent D Buskirk, Jennifer Hunter Childs, and Casey Langer Tesfaye. Mobile technologies for conducting, augmenting and potentially replacing surveys: Executive summary of the aapor task force on emerging technologies in public opinion research. *Public Opinion Quarterly*, 78(4):779–787, 2014.

[124] Deryle Lonsdale, David W Embley, Yihong Ding, Li Xu, and Martin Hepp. Reusing ontologies and language components for ontology generation. *Data & Knowledge Engineering*, 69(4):318–330, 2010.

[125] MAURICE LORR. Chapter 2 - models and methods for measurement of mood. In Robert Plutchik and Henry Kellerman, editors, *The Measurement of Emotions*, pages 37–53. Academic Press, 1989.

[126] Manuel V Loureiro, Steven Derby, and Tri Kurniawan Wijaya. Topics as entity clusters: Entity-based topics from large language models and graph neural networks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16315–16330, 2024.

[127] Peter Lynn and Olena Kaminska. The impact of mobile phones on survey measurement error. *Public Opinion Quarterly*, 77(2):586–605, 2013.

[128] Jarmo Makkonen, Ivan Avdouevski, Riitta Kerminen, and Ari Visa. Context awareness in human-computer interaction. In *Human-Computer Interaction*. IntechOpen, 2009.

[129] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[130] Kira O McCabe and William Fleeson. What is extraversion for? integrating trait and motivational perspectives and identifying the purpose of extraversion. *Psychological science*, 23(12):1498–1505, 2012.

[131] Kira O McCabe, Lori Mack, and William Fleeson. A guide for data cleaning in experience sampling studies. 2012.

[132] Abhinav Mehrotra, Mirco Musolesi, Robert Hendley, and Veljko Pejovic. Designing content-driven intelligent notification mechanisms for mobile

applications. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 813–824, 2015.

[133] Abhinav Mehrotra, Jo Vermeulen, Veljko Pejovic, and Mirco Musolesi. Ask, but don't interrupt: the case for interruptibility-aware mobile experience sampling. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pages 723–732, 2015.

[134] Giovanna Miritello, Rubén Lara, Manuel Cebrian, and Esteban Moro. Limited communication capacity unveils strategies for human interaction. *Scientific reports*, 3(1):1–7, 2013.

[135] Varun Mishra, Byron Lowens, Sarah Lord, Kelly Caine, and David Kotz. Investigating contextual cues as indicators for ema delivery. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 935–940, 2017.

[136] Marzia Mortati, Stefano Magistretti, Cabirio Cautela, and Claudio Dell'Era. Data in design: How big data and thick data inform design thinking projects. *Technovation*, 122:102688, 2023.

[137] Francisco Munguia-Galeano, Ah-Hwee Tan, and Ze Ji. Deep reinforcement learning with explicit context representation. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[138] Alexandre Nanchen, Lakmal Meegahapola, William Droz, and Daniel Gatica-Perez. Keep sensors in check: Disentangling country-level generalization issues in mobile sensor-based models with diversity scores. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 217–228, 2023.

[139] United States. Executive Office of the President and John Podesta. *Big data: Seizing opportunities, preserving values*. White House, Executive Office of the President, 2014.

[140] Hyungik Oh, Laleh Jalali, and Ramesh Jain. An intelligent notification system using context from real-time personal activity monitoring. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2015.

[141] Jukka-Pekka Onnela, Caleb Dixon, Keary Griffin, Tucker Jaenicke, Leila Minowada, Sean Esterkin, Alvin Siu, Josh Zagorsky, and Eli Jones. Beiwe: A data collection platform for high-throughput digital phenotyping. *Journal of Open Source Software*, 6(68):3417, 2021.

[142] Antti Oulasvirta, Sakari Tamminen, Virpi Roto, and Jaana Kuorelahti. Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile hci. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 919–928, 2005.

[143] Heinz Paetzold. Review essay: Respect and toleration reconsidered (under consideration: Rainer forst's toleranz im konflikt: Geschichte, gehalt, und gegenwart eines umstrittenen begriffs (frankfurt am main: Suhrkamp, 2003)(english translation forthcoming, cambridge university press)). *Philosophy & social criticism*, 34(8):941–954, 2008.

[144] Vikram Patel, Alan J Flisher, Sarah Hetrick, and Patrick McGorry. Mental health of young people: a global public-health challenge. *The lancet*, 369(9569):1302–1313, 2007.

[145] Ella Peltonen, Parsa Sharmila, Kennedy Opoku Asare, Aku Visuri, Eemil Lagerspetz, and Denzil Ferreira. When phones get personal: Predicting big five personality traits from application usage. *Pervasive and Mobile Computing*, 69:101269, 2020.

[146] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. Beyond interruptibility: Predicting opportune moments to engage mobile phone users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–25, 2017.

[147] Youcheng Qian and Xueyan Yin. Semantic-enhanced graph neural networks with global context representation. *Machine Learning*, pages 1–21, 2024.

[148] Mashfiqui Rabbi, Min Hane Aung, Mi Zhang, and Tanzeem Choudhury. Mybehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 707–718, 2015.

[149] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–27, 2018.

[150] Mika Raento, Antti Oulasvirta, and Nathan Eagle. Smartphones: An emerging tool for social scientists. *Sociological methods & research*, 37(3):426–454, 2009.

[151] Elangovan Ramanujam, Thinagaran Perumal, and S Padmavathi. Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review. *IEEE Sensors Journal*, 21(12):13029–13040, 2021.

[152] Yatharth Ranjan, Zulqarnain Rashid, Callum Stewart, Pauline Conde, Mark Begale, Denny Verbeeck, Sebastian Boettcher, Richard Dobson,

Amos Folarin, RADAR-CNS Consortium, et al. Radar-base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. *JMIR mHealth and uHealth*, 7(8):e11734, 2019.

[153] Brendan Read. Respondent burden in a mobile app: Evidence from a shopping receipt scanning study. In *Survey Research Methods*, volume 13, pages 45–71. European Survey Research Association, 2019.

[154] Ericka Janet Rechy-Ramirez, Antonio Marin-Hernandez, and Homero Vladimir Rios-Figueroa. Impact of commercial sensors in human computer interaction: a review. *Journal of Ambient Intelligence and Humanized Computing*, 9:1479–1496, 2018.

[155] Karen Renaud, Judith Ramsay, and Mario Hair. " you've got e-mail!"... shall i deal with it now? electronic mail from the recipient's perspective. *International Journal of Human-Computer Interaction*, 21(3):313–332, 2006.

[156] Daniele Riboni and Claudio Bettini. Owl 2 modeling and reasoning with complex human activities. *Pervasive and Mobile Computing*, 7(3):379–395, 2011.

[157] Yannick Roos, Michael D Krämer, David Richter, Ramona Schoedel, and Cornelia Wrzus. Does your smartphone "know" your social life? a methodological comparison of day reconstruction, experience sampling, and mobile sensing. *Advances in Methods and Practices in Psychological Science*, 6(3):25152459231178738, 2023.

[158] Patrick Royston. Explained variation for survival models. *The Stata Journal*, 6(1):83–96, 2006.

[159] Jason D Runyan, Timothy A Steenbergh, Charles Bainbridge, Douglas A Daugherty, Lorne Oke, and Brian N Fry. A smartphone ecological momentary assessment/intervention "app" for collecting real-time data and promoting self-awareness. *PloS one*, 8(8):e71325, 2013.

[160] Cheryl L Rusting. Personality, mood, and cognitive processing of emotional information: three conceptual frameworks. *Psychological bulletin*, 124(2):165, 1998.

[161] Gilbert Ryle and Julia Tanney. *Collected Essays 1929-1968: Collected Papers Volume 2*. Routledge, 2009.

[162] Hassan Sarmadi, Alireza Entezami, Ka-Veng Yuen, and Bahareh Behkamal. Review on smartphone sensing technology for structural health monitoring. *Measurement*, 223:113716, 2023.

[163] Laura Schelenz, Avi Segal, Oduma Adelio, and Kobi Gal. Transparency-check: An instrument for the study and design of transparency in ai-based personalization systems. *ACM Journal on Responsible Computing*, 1(1):1–18, 2024.

[164] Bill Schilit, Norman Adams, and Roy Want. Context-aware computing applications. In *1994 first workshop on mobile computing systems and applications*, pages 85–90. IEEE, 1994.

[165] Bill N Schilit and Marvin M Theimer. Disseminating active map information to mobile hosts. *IEEE network*, 8(5):22–32, 1994.

[166] Christina Schmidt, Fabienne Collette, Christian Cajochen, and Philippe Peigneux. A time to think: circadian rhythms in human cognition. *Cognitive neuropsychology*, 24(7):755–789, 2007.

[167] Shalom H Schwartz. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):2307–0919, 2012.

[168] Indira Sen, Fabian Floeck, Katrin Weller, Bernd Weiss, and Claudia Wagner. A total error framework for digital traces of humans. *arXiv preprint arXiv:1907.08228*, 2019.

[169] Mahsa Sheikh, Meha Qassem, and Panicos A Kyriacou. Wearable, environmental, and smartphone-based passive sensing for mental health monitoring. *Frontiers in Digital Health*, 3:662811, 2021.

[170] Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32, 2008.

[171] Muhammad Shoaib, Stephan Bosch, Hans Scholten, Paul JM Havinga, and Ozlem Durmaz Incel. Towards detection of bad habits by fusing smartphone and smartwatch sensors. In *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pages 591–596. IEEE, 2015.

[172] Judith D Singer, John B Willett, John B Willett, et al. *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press, 2003.

[173] Annelien Smets and Bram Lievens. Human sensemaking in the smart city: A research approach merging big and thick data. In *Ethnographic Praxis in Industry Conference Proceedings*, volume 2018, pages 179–194. Wiley Online Library, 2018.

[174] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[175] Thanos G Stavropoulos, Georgios Meditskos, and Ioannis Kompatsiaris. Demaware2: Integrating sensors, multimedia and semantic analysis for

the ambient care of dementia. *Pervasive and Mobile Computing*, 34:126–145, 2017.

[176] Thanos G Stavropoulos, Georgios Meditskos, Ioulietta Lazarou, Lampros Mpaltadoros, Sotirios Papagiannopoulos, Magda Tsolaki, and Ioannis Kompatsiaris. Detection of health-related events and behaviours from wearable sensor lifestyle data using symbolic intelligence: A proof-of-concept application in the care of multiple sclerosis. *Sensors*, 21(18):6230, 2021.

[177] Piers Steel. The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, 133(1):65–94, 2007.

[178] Piers Steel. Arousal, avoidant and decisional procrastinators: Do they exist? *Personality and Individual Differences*, 48(8):926–934, 2010.

[179] Fiona Steele. Multilevel models for longitudinal data. *Journal of the Royal Statistical Society: series A (statistics in society)*, 171(1):5–19, 2008.

[180] Arthur Stone, Saul Shiffman, Audie Atienza, and Linda Nebeling. *The science of real-time data capture: Self-reports in health research*. Oxford University Press, 2007.

[181] Bella Struminskaya, Peter Lugtig, Florian Keusch, and Jan Karem Höhne. Augmenting surveys with data from sensors and apps: Opportunities and challenges. *Social Science Computer Review*, page 0894439320979951, 2020.

[182] Lucille Alice Suchman. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press, 1987.

[183] Seymour Sudman and Norman M Bradburn. Effects of time and memory factors on response in surveys. *Journal of the American Statistical Association*, 68(344):805–815, 1973.

[184] Jessie Sun, Mijke Rhemtulla, and Simine Vazire. Eavesdropping on missing data: What are university students doing when they miss experience sampling reports? *Personality and Social Psychology Bulletin*, 47(11):1535–1549, 2021.

[185] Hui Yie Teh, Andreas W Kempa-Liehr, and Kevin I-Kai Wang. Sensor data quality: A systematic review. *Journal of Big Data*, 7(1):11, 2020.

[186] Fetene B Tekle and Jeroen K Vermunt. Event history analysis. 2012.

[187] Roger Tourangeau et al. Cognitive sciences and survey methods. *Cognitive aspects of survey methodology: Building a bridge between disciplines*, 15:73–100, 1984.

[188] Roger Tourangeau, Lance J Rips, and Kenneth Rasinski. The psychology of survey response. 2000.

[189] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. Recognizing detailed human context in the wild from smartphones and smartwatches. *IEEE pervasive computing*, 16(4):62–74, 2017.

[190] Yonatan Vaizman, Nadir Weibel, and Gert Lanckriet. Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–22, 2018.

[191] Niels van Berkel, Jorge Goncalves, Simo Hosio, Zhanna Sarsenbayeva, Eduardo Velloso, and Vassilis Kostakos. Overcoming compliance bias in self-report studies: A cross-study analysis. *International Journal of Human-Computer Studies*, 134:1–12, 2020.

[192] Niels van Berkel, Jorge Goncalves, Peter Koval, Simo Hosio, Tilman Dingler, Denzil Ferreira, and Vassilis Kostakos. Context-informed scheduling and analysis: improving accuracy of mobile self-reports. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[193] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, pages 3–14, 2014.

[194] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T Campbell. Smartgpa: how smartphones can assess and predict academic performance of college students. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 295–306, 2015.

[195] Rui Wang, Weichen Wang, Alex DaSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–26, 2018.

[196] Weichen Wang, Gabriella M Harari, Rui Wang, Sandrine R Müller, Shayan Mirjafari, Kizito Masaba, and Andrew T Campbell. Sensing behavioral change over time: Using within-person variability features from mobile sensing to predict personality traits. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–21, 2018.

[197] Weichen Wang, Shayan Mirjafari, Gabriella Harari, Dror Ben-Zeev, Rachel Brian, Tanzeem Choudhury, Marta Hauser, John Kane, Kizito Masaba, Subigya Nepal, et al. Social sensing: assessing social functioning of patients living with schizophrenia using mobile phone sensing. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–15, 2020.

[198] Yanda Wang, Weitong Chen, Dechang Pi, Lin Yue, Sen Wang, and Miao Xu. Self-supervised adversarial distribution regularization for medication recommendation. In *IJCAI*, pages 3134–3140, 2021.

[199] Herbert F Weisberg. *The total survey error approach: A guide to the new science of survey research*. University of Chicago Press, 2009.

[200] Alexander Wenz, Annette Jackle, and Mick P Couper. Willingness to use mobile technologies for data collection in a probability household panel. In *Survey Research Methods*, volume 13, pages 1–22. European Survey Research Association, 2019.

[201] Robert West, Kelly J Murphy, Maria L Armilio, Fergus IM Craik, and Donald T Stuss. Effects of time of day on age differences in working memory. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 57(1):P3–P10, 2002.

[202] Peter Wilhelm, Meinrad Perrez, and Kurt Pawlik. Conducting research in daily life: A historical review. 2012.

[203] Li Xiaoyue, Rodas Britez Marcelo, Busso Matteo, and Giunchiglia Fausto. Representing habits as streams of situational contexts. In *Advanced Information Systems Engineering Workshops: CAiSE 2022 International Workshops*, 2022.

[204] Jameel Shehu Yalli, Mohd Hilmi Hasan, Nazleeni Samiha Haron, Mujeeb Ur Rehman Shaikh, Nafeesa Yousuf Murad, and Abdullahi Lawal Bako. Quality of data (qod) in internet of things (iot): An overview, state-of-the-art, taxonomy and future directions. *International Journal of Advanced Computer Science & Applications*, 14(12), 2023.

[205] Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. In-context learning with representations: Contextual generalization of trained transformers. *arXiv preprint arXiv:2408.10147*, 2024.

[206] Fengpeng Yuan, Xianyi Gao, and Janne Lindqvist. How busy are you? predicting the interruptibility intensity of mobile users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.

[207] Lin Yue, Hao Shen, Sen Wang, Robert Boots, Guodong Long, Weitong Chen, and Xiaowei Zhao. Exploring bci control in smart environments: intention recognition via eeg representation enhancement learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–20, 2021.

[208] Lin Yue, Dongyuan Tian, Weitong Chen, Xuming Han, and Minghao Yin. Deep learning for heterogeneous medical data analysis. *World Wide Web*, 23(5):2715–2737, 2020.

[209] Lin Yue, Dongyuan Tian, Jing Jiang, Lina Yao, Weitong Chen, and Xiaowei Zhao. Intention recognition from spatio-temporal representation of eeg signals. In *Australasian Database Conference*, pages 1–12. Springer, 2021.

[210] Lin Yue, Haonan Zhao, Yiqin Yang, Dongyuan Tian, Xiaowei Zhao, and Minghao Yin. A mimic learning method for disease risk prediction with

incomplete initial data. In *International Conference on Database Systems for Advanced Applications*, pages 392–396. Springer, 2019.

[211] Özgür Yürür, Chi Harold Liu, Zhengguo Sheng, Victor CM Leung, Wilfrido Moreno, and Kin K Leung. Context-awareness for mobile sensing: A survey and future directions. *IEEE Communications Surveys & Tutorials*, 18(1):68–93, 2014.

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

[213] Lina Zhang, Dongwon Jeong, and Sukhoon Lee. Data quality management in the internet of things. *Sensors*, 21(17):5834, 2021.

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

[215] Zhi-Hua Zhou and Ji Feng. Deep forest: towards an alternative to deep neural networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3553–3559, 2017.

[216] Zhi-Hua Zhou and Ji Feng. Deep forest. *National Science Review*, 6(1):74–86, 2019.

[217] Xingquan Zhu and Ian Davidson. *Knowledge Discovery and Data Mining: Challenges and Realities: Challenges and Realities*. Igi Global.