

DataSentry :

Building Missing Data Management System for In-the-Wild Mobile Sensor Data Collection through Multi-Year Iterative Design Approach

Yugyeong Jung ^[1], Hei Yiu Law ^[1], Hadong Lee ^[2],
Junmo Lee ^[1], Bongshin Lee ^[3], and Uichin Lee ^[1]



[1]



[2]



[3]



Mobile Data Collection: The Foundation of Mobile Sensing Studies



Diagnosing health conditions [1]



Predicting productivity [2]

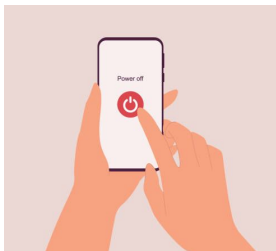


Analyzing social interactions [3]

[1] The Perceived Utility of Smartphone and Wearable Sensor Data in Digital Self-tracking Technologies for Mental Health, CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems
[2] Understanding Personal Productivity: How Knowledge Workers Define, Evaluate, and Reflect on Their Productivity, CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems
[3] Social Sensing: Assessing Social Functioning of Patients Living with Schizophrenia using Mobile Phone Sensing, CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems

Missing Data Issues in Mobile Data Collection

Participant behaviors [1]



Powering-off

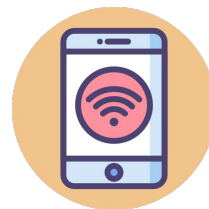


No self-reports



Dropout

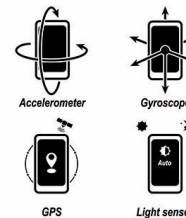
System-related issues [2]



Network



Server



Sensors



Sensor
Logging



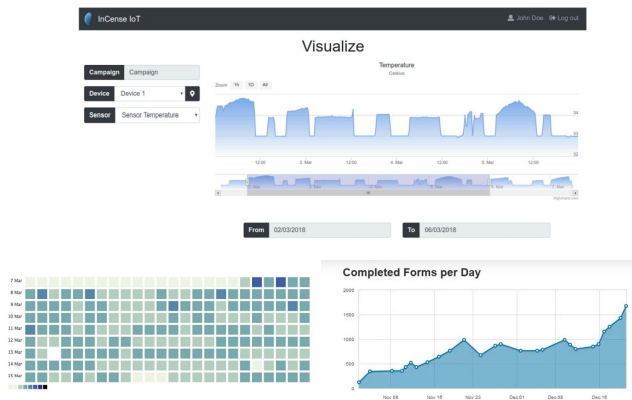
Missing data

[1] Xuhai Xu, Jennifer Mankoff, and Anind K Dey. 2021. Understanding practices and needs of researchers in human state modeling by passive mobile sensing. CCF Transactions on Pervasive Computing and Interaction 3 (2021), 344–366.

[2] Jennifer Healey, Lama Nachman, Sushmita Subramanian, Junaith Shahabdeen, and Margaret Morris. 2010. Out of the lab and into the fray: Towards modeling emotion in everyday life. In Pervasive Computing: 8th International Conference, Pervasive 2010, Helsinki, Finland, May 17-20, 2010. Proceedings 8. Springer, 156–173

Challenges in Missing Data Management

Holistic understanding of missing data across many people and sensors



Inspecting **individual** sensor streams or aggregated metrics of **a specific sensor**

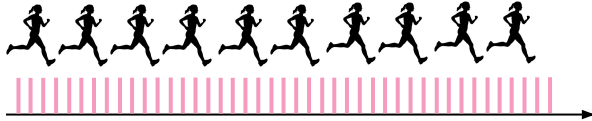


Lack of holistic understanding of missing data **across many people and sensors**

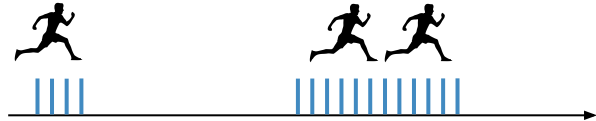
Challenges in Missing Data Management

Needs for considering between- and within-participant variability

Between-participant variability

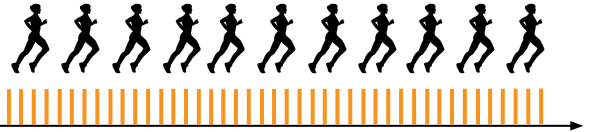


Physical activity data of **participant A**



Physical activity data of **participant B**

Within-participant variability



Physical activity data **during a day**



Physical activity data **during night**

Challenges in Missing Data Management

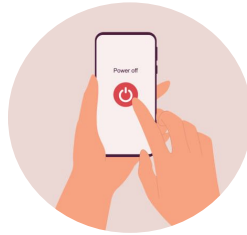
Difficulty in diagnosing root causes of missing data



Difficult to diagnose...



Participant's common
behavior patterns?



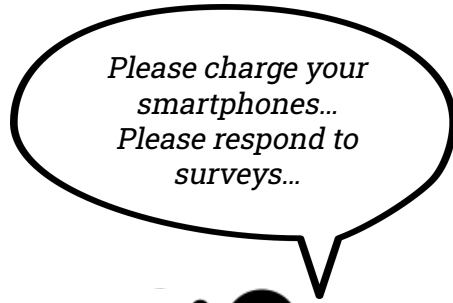
Powering off
smartphones?



Network or data
sync issues?

Challenges in Missing Data Management

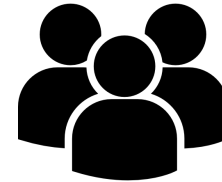
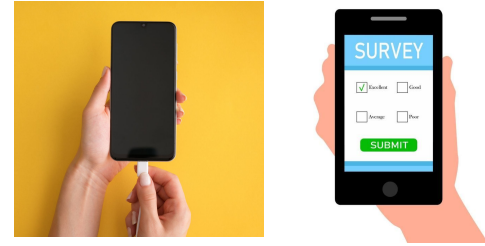
Communication burden of researchers in addressing the issues



Data collection
researcher



**Managing participants
for quality data collection**



Data collection
participants

An illustration depicting a data management system design process. At the top, a person sits on a large blue monitor, working on a laptop. The monitor displays a bar chart and a magnifying glass icon. To the left, a person stands next to a large yellow clipboard labeled 'TESTING' with a checklist of four items, each with a red checkmark or an 'X'. To the right, another person stands holding a large grey cylinder. The background features large green and yellow gears, a red gear, a paper airplane, a hashtag, and a thumbs-up icon.

“Designing a missing data management system to **detect missing data, **diagnose** their root causes, and **address** them during mobile sensor data collection campaign”**

Formative Study

Interviewing **seven**
mobile sensing researchers

Design Requirement 1

Overviewing missing data across
many people and sensors



*"...it would be helpful to display whether
data from each sensor and each person
was collected."*

Design Requirement 2

Identifying long missing data in
event-based sensing



*"If there's only a small number of rows, it
seem like an issue with the sensor.
However, it could be because the user
didn't move..."*

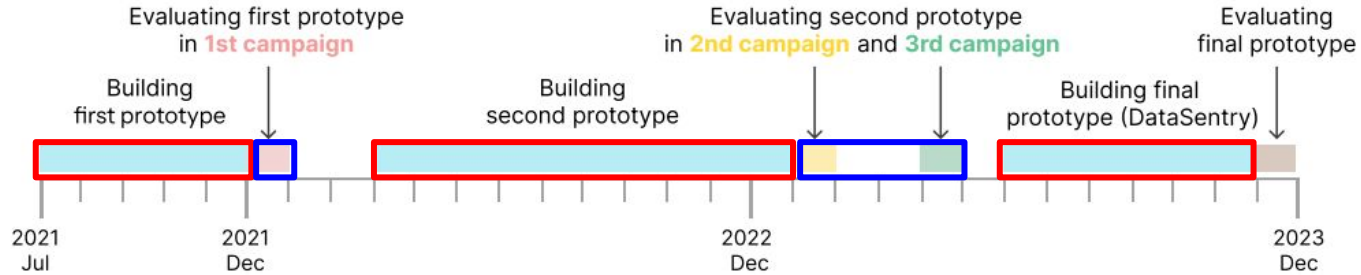
Design Requirement 3

Diagnosing missing data
causes



*"If I can check multiple data items at
a glance, then I can determine why
the data was not collected..."*

Iterative Design Process



First campaign

Data Collection

17 kinds of mobile sensor data
5 kinds of self-report ESM

Participants

116 participants
(43 women; age: $M = 23.5$, $STD = 3.5$)

Researchers

2 researchers
(1 women; age: $M = 28.5$, $STD = 2.2$)

Second campaign

Data Collection

19 kinds of mobile sensor data
4 kinds of self-report ESM

Participants

20 participants
(7 women; age: $M = 24.8$, $STD = 2.7$)

Researchers

2 researchers
(2 women; age: $M = 26.0$, $STD = 1.4$)

Third campaign

Data Collection

5 kinds of mobile sensor data
2 kinds of self-report ESM

Participants

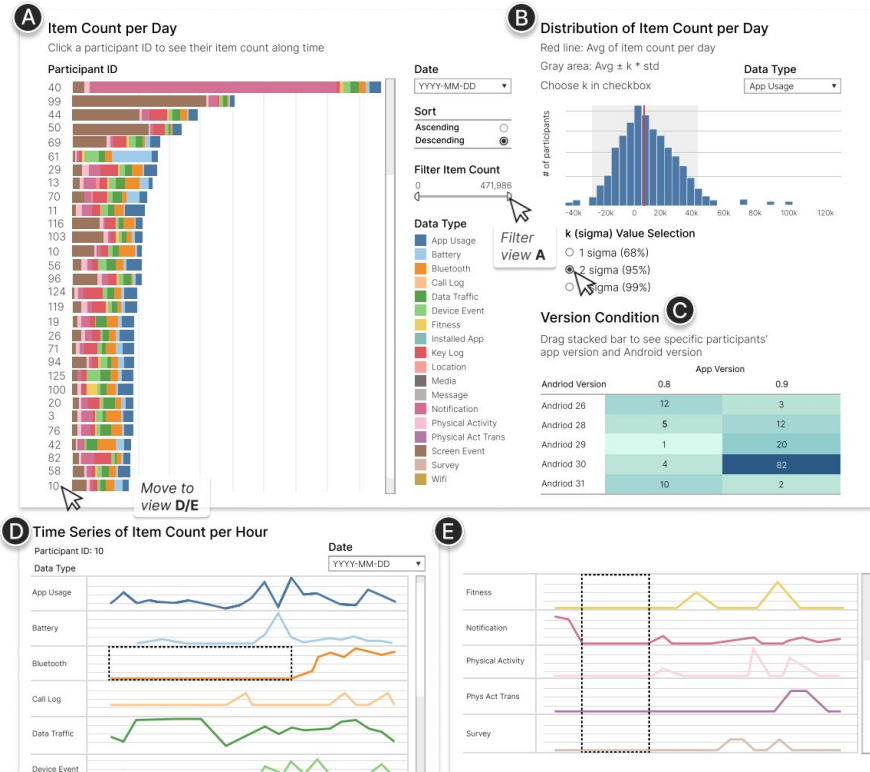
24 participants
(9 women; age: $M = 21.3$, $STD = 2.1$)

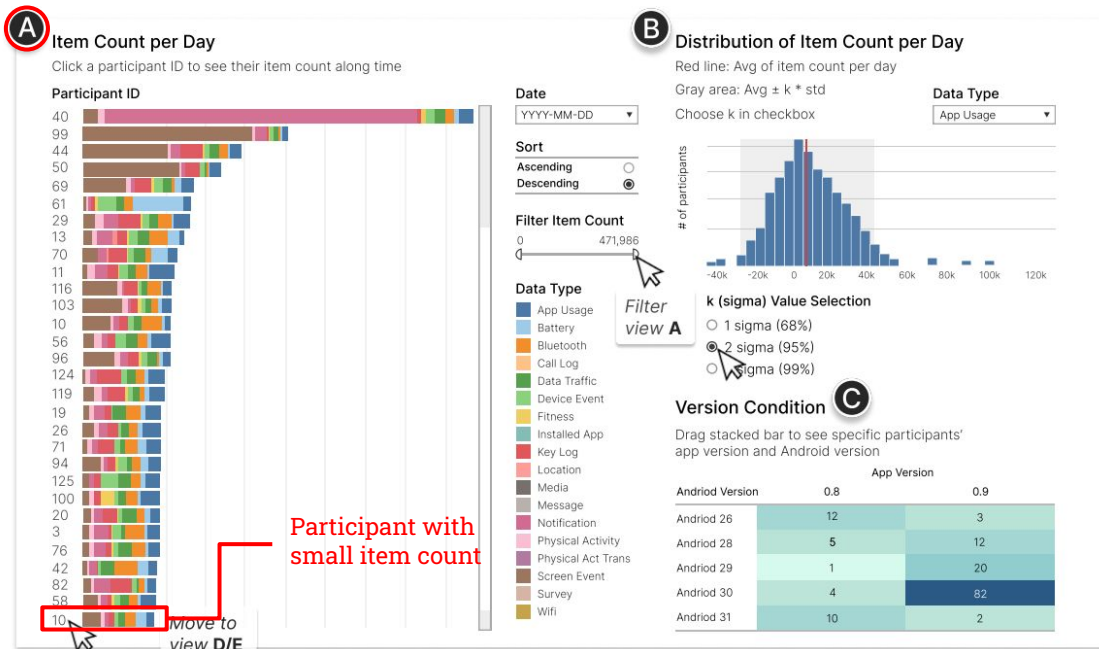
Researchers

1 researcher
(1 man; age = 33)

First Design Iteration

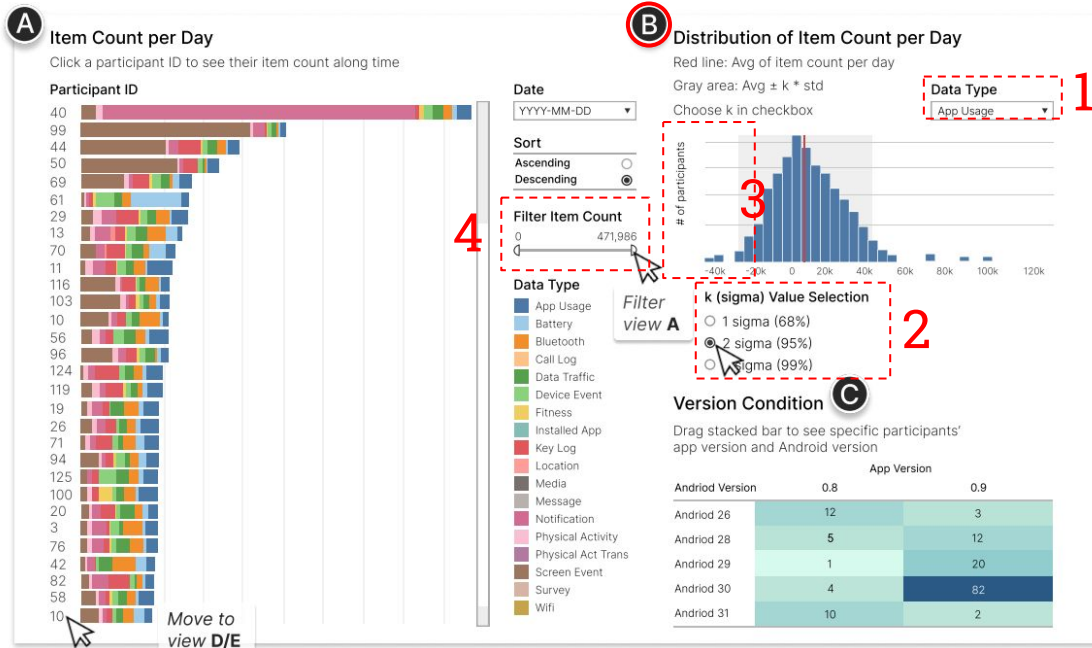
Reflecting three design requirements in formative study





Overview of missing data across people and sensors (Design Requirement 1)

- Calculating **daily item count** for each sensor and participant
- Visualizing the metric as **stacked bars**



Data-driven guidelines using statistical quality control

(Design Requirement 2)

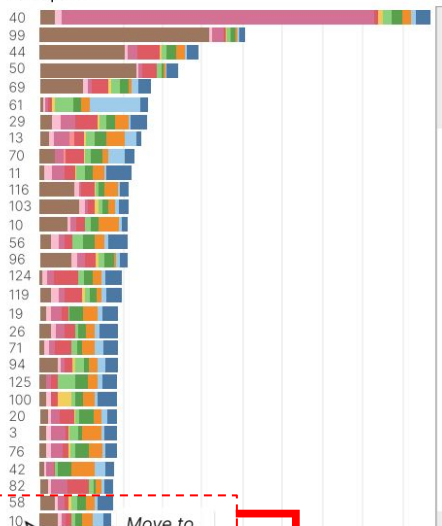
- Providing guidelines to determine which item count might indicate missing data in **event-based sensing**
- The concept of **control charts** → Outlier metric as values outside $[\mu - k\sigma, \mu + k\sigma]$

A

Item Count per Day

Click a participant ID to see their item count along time

Participant ID



Move to view D/E

B

Distribution of Item Count per Day

Red line: Avg of item count per day

Gray area: Avg \pm k * std

Choose k in checkbox

Data Type

App Usage

Date

YYYY-MM-DD

Sort

Ascending

Descending

Filter Item Count

0 471,985

Filter view A

Data Type

App Usage

Battery

Bluetooth

Call Log

Data Traffic

Device Event

Fitness

Installed App

Key Log

Location

Media

Message

Notification

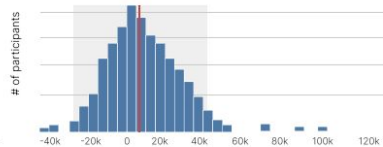
Physical Activity

Physical Act Trans

Screen Event

Survey

Wifi



k (sigma) Value Selection

☐ 1 sigma (68%)☒ 2 sigma (95%)☐ 3 sigma (99%)

Version Condition

Drag stacked bar to see specific participants' app version and Android version

Android Version	App Version	
	0.8	0.9
Android 26	12	3
Android 28	5	12
Android 29	1	20
Android 30	4	82
Android 31	10	2

D

Time Series of Item Count per Hour

Participant ID: 10

Date

YYYY-MM-DD

Data Type

App Usage

Battery

Bluetooth

Call Log

Data Traffic

Device Event

Fitness

Notification

Physical Activity

Physical Act Trans

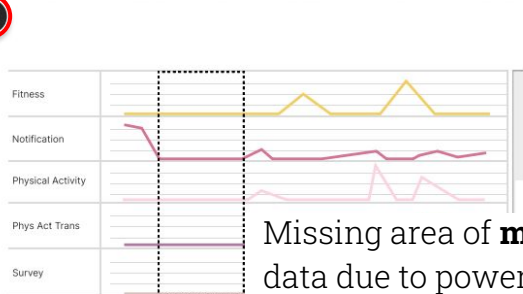
Screen Event

Survey

Wifi

Missing areas of a **single** data due to sensor malfunction

E



Missing area of **multiple** data due to powering off or out of battery

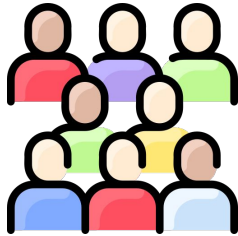
Visual exploration for missing data diagnosis

(Design Requirement 3)

- Inspecting **temporal trend** of item counts in hourly level

Field deployment

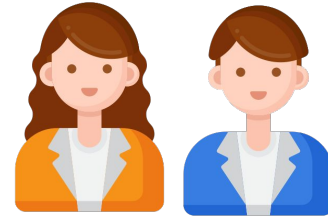
First data collection campaign



116 data collection
participants



17 mobile sensor data
5 self-report ESM

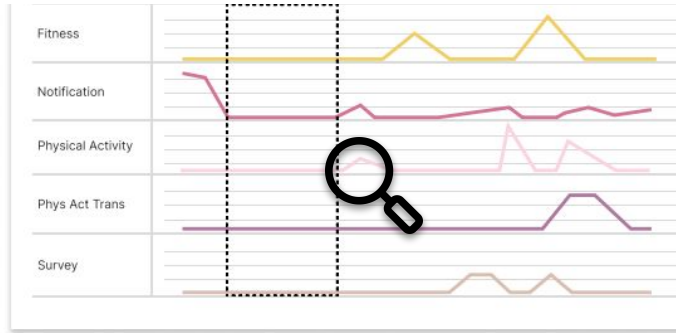


2 managing
researchers

Design Insight from Field Deployment

Design Insight 1

Needs for reviewing
raw sensor data



*Difficult to pinpoint **exactly when**
the data was missing...*

Design Insight 2

Needs for diagnosing the causes of unexpected
missing data by observing multiple sensor streams

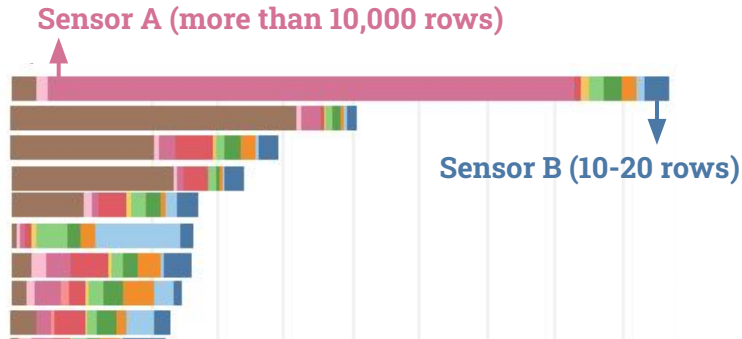


*Typical sensing pattern?
Sensor issues?
How about other participants' data?*

Usability Issues from Field Deployment

Usability Issue 1

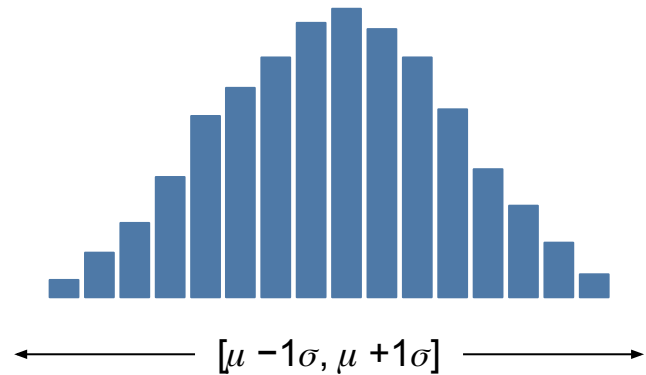
Difficulty of inspecting stacked bars



Wide variation in the scale of count metrics across different sensors

Usability Issue 2

Ineffectiveness of a statistical quality control method



Distribution was usually within $[\mu - 1\sigma, \mu + 1\sigma]$
→ Difficult to find outlying participants

Second Design Iteration

Reflecting design insights and usability issues from first design iteration



Missing data diagnosis via one participant's multiple sensor streams

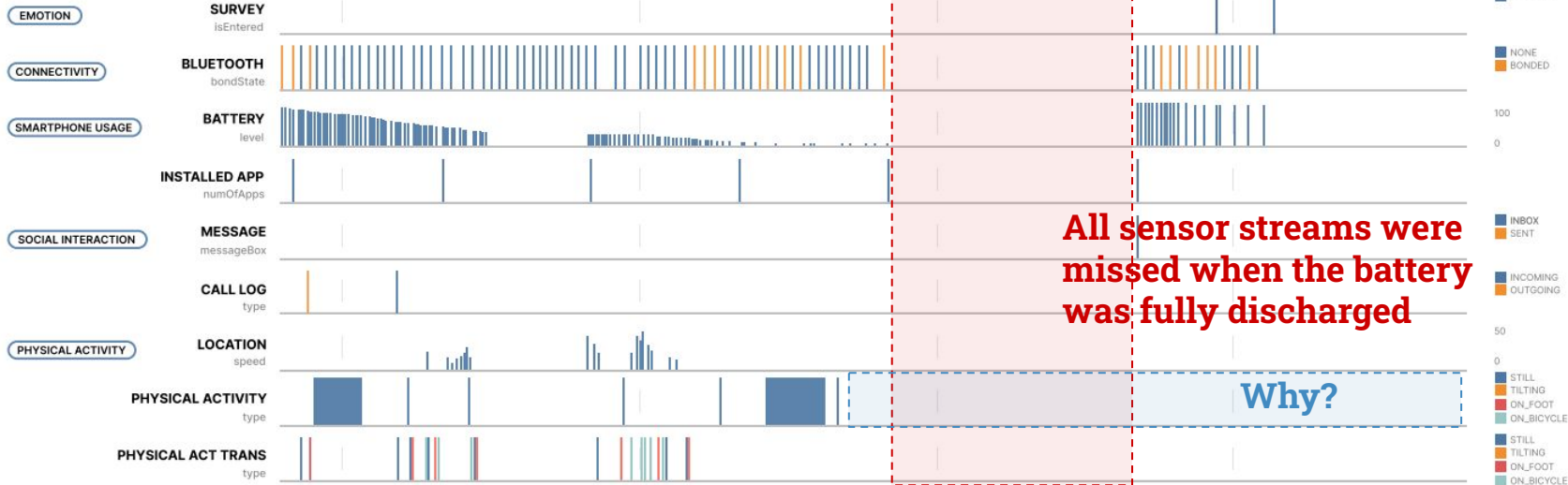
B

Detail of One Participant

Email: participant10@email.com

Period

Selected period: 2023-01-01 00:00 - 2023-01-01 23:59



Missing data diagnosis using within-participant comparison

PHYSICAL ACTIVITY

type

Within-Person Time Series Comparison

Between-Person Time Series Comparison



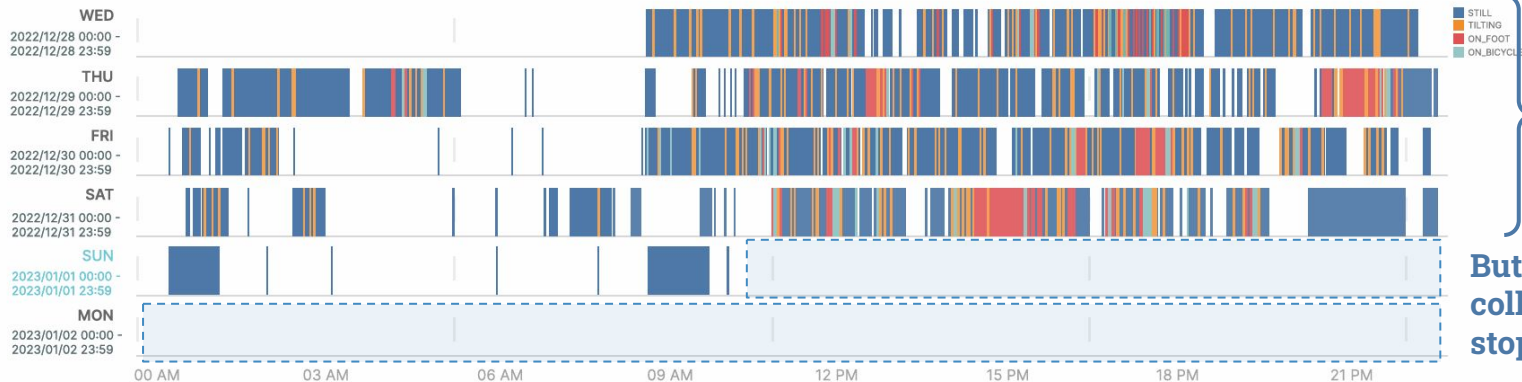
Within-Participant Time-Series Comparison

Email: participant10@email.com

Sensor Type: PHYSICAL ACTIVITY

Column: type

Selected Period: 2023-01-01 00:00 - 2023-01-01 23:59



Data was
previously
collected,

But data
collection
stopped

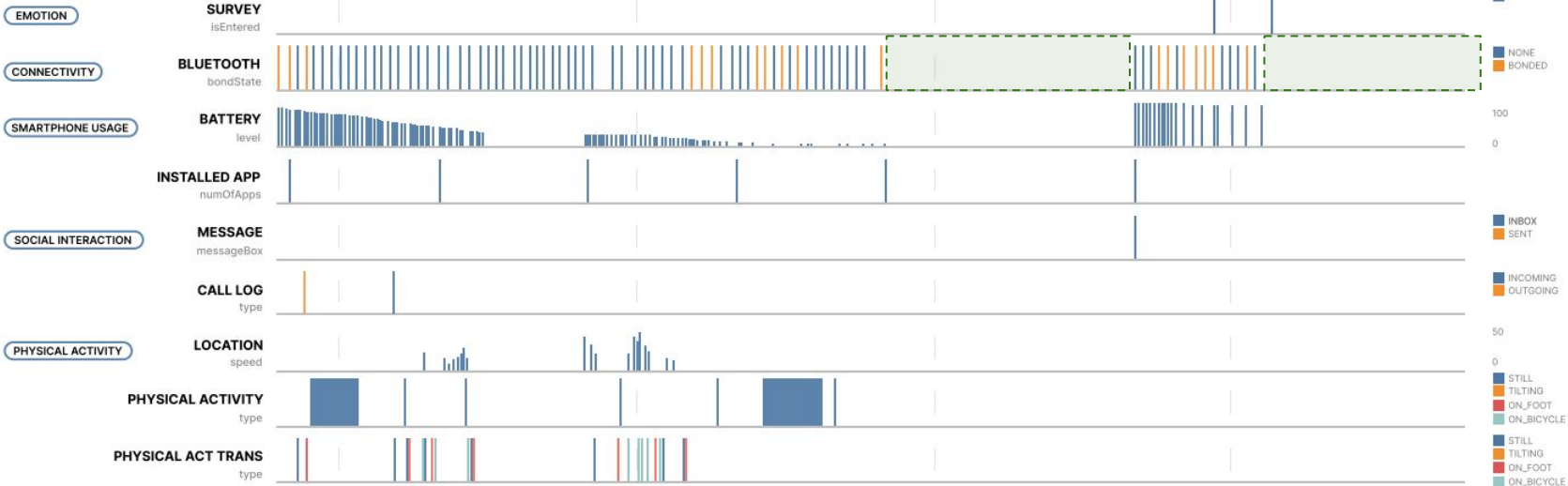
B

Detail of One Participant

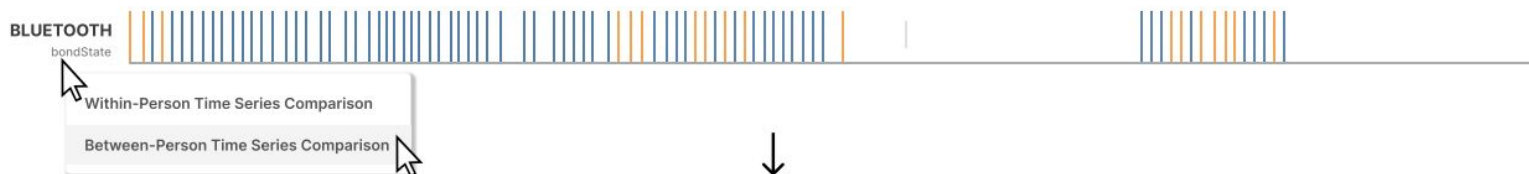
Email: participant10@email.com

Period

Selected period: 2023-01-01 00:00 - 2023-01-01 23:59



Missing data diagnosis using between-participant comparison



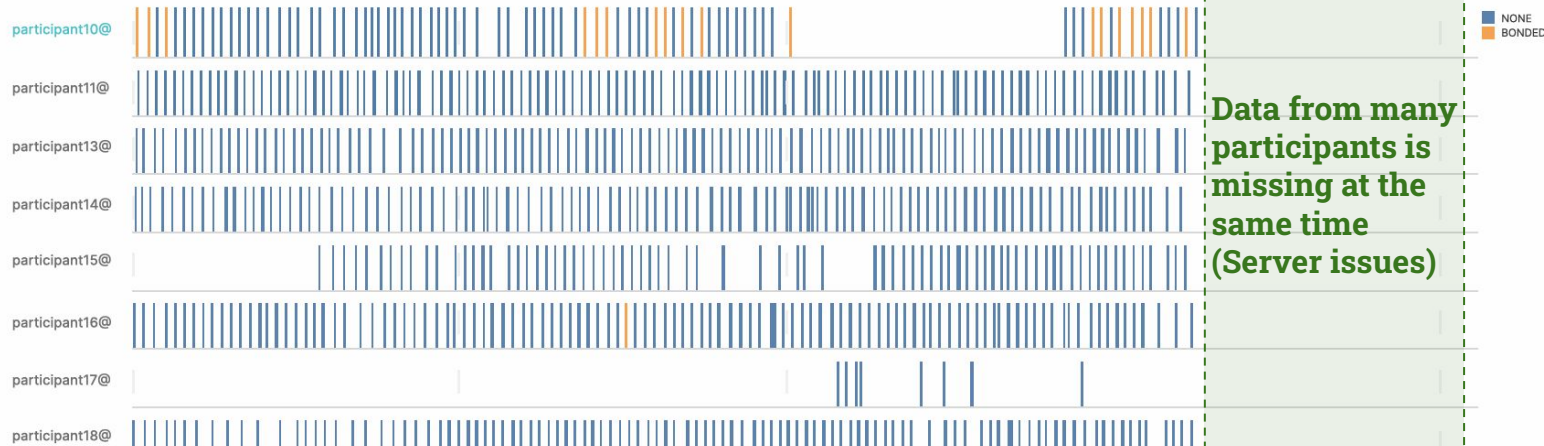
D Between-Participant Time-Series Comparison

Email: participant10@email.com

Sensor Type: BLUETOOTH

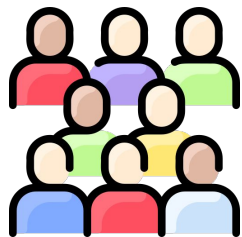
Column: bondState

Selected Period: 2023-01-01 00:00 - 2023-01-01 23:59



Field Deployment

Second and third data collection campaigns



Second campaign

20 data collection participants

19 mobile sensor data
4 self-report ESM

2 managing researchers

1 month

Third campaign

24 data collection participants

5 mobile sensor data
2 self-report ESM

1 managing researcher

1 month

Design Insight from Field Deployment

Design insight 1

Needs for streamlined detection
of long missing periods



Need to diagnose long missing period
by switching several pages repetitively

Design insight 2

Needs for lowering the burden
of communication with participants

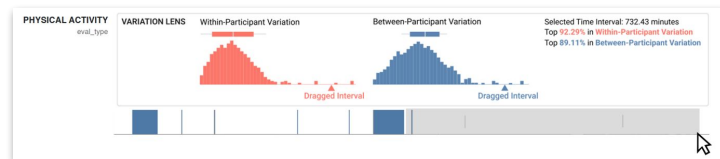


Third Design Iteration

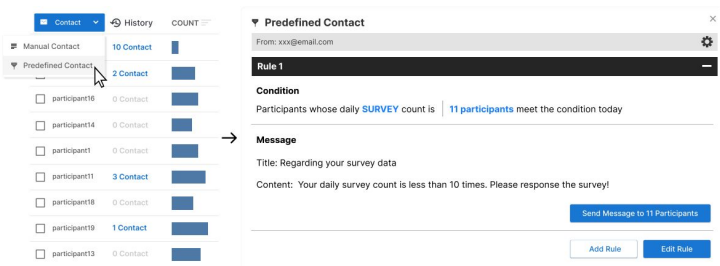
Adding **two main features** reflecting design insights from second design iteration



(D1)



(D2)



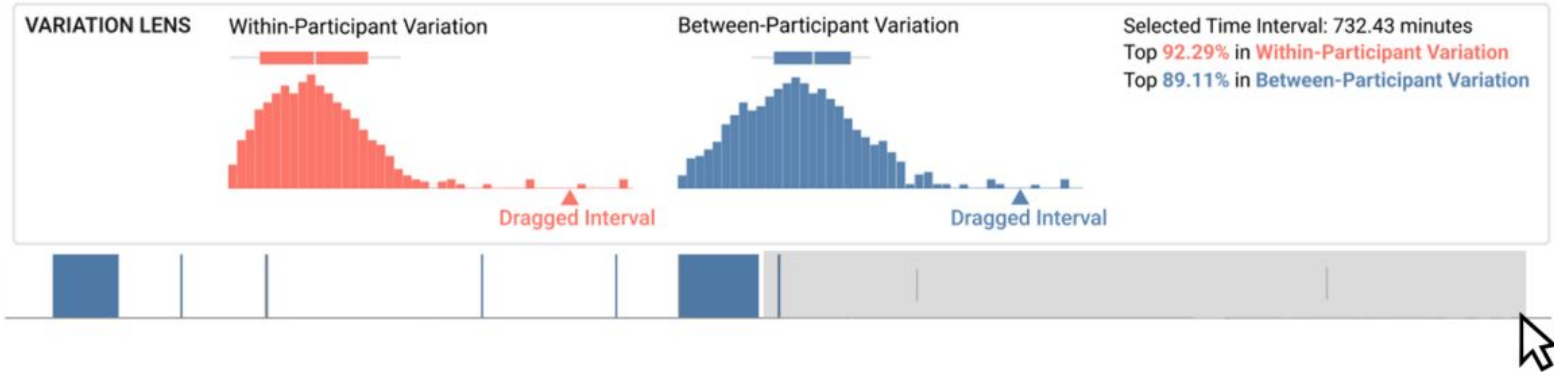
Within- and Between- Participants Variation Lens

Reflecting Design insight 1

How the selected time interval compares in terms of the distribution of time intervals

- 1) **Within** the participant's data
- 2) **Between** participants' data

PHYSICAL ACTIVITY
eval_type



Rule-Based Contact Feature

Reflecting Design insight 2

The interface displays a list of participants and their contact counts, with a dropdown menu for 'Predefined Contact'. An arrow points from the table to a detailed view of 'Rule 1'.

	Contact	COUNT
Manual Contact	10 Contact	
Predefined Contact	2 Contact	
<input type="checkbox"/> participant16	0 Contact	
<input type="checkbox"/> participant14	0 Contact	
<input type="checkbox"/> participant1	0 Contact	
<input type="checkbox"/> participant11	3 Contact	
<input type="checkbox"/> participant18	0 Contact	
<input type="checkbox"/> participant19	1 Contact	
<input type="checkbox"/> participant13	0 Contact	

Predefined Contact

From: xxx@email.com

Rule 1

Condition
Participants whose **SURVEY** count is less than **10 times** | **11 participants** meet the condition

Message
Title: Regarding your survey data
Content: Your daily survey count is less than 10 times. Please response the survey!

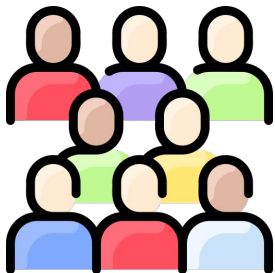
Send Message to 11 Participants

Add Rule **Edit Rule**

Final Evaluation

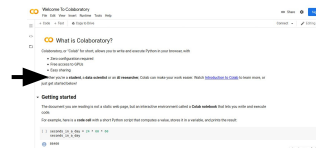
Goal 1. To evaluate DataSentry by researchers from various research groups

Goal 2. To observe how user experiences differ depending on whether within- and between-person comparisons are supported

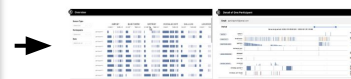


26 researchers from
11 different groups
(Academia and industry)

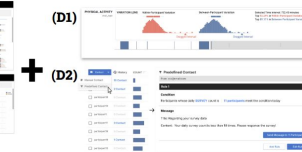
1. Python plotting (Google Colab)



2. DataSentry Basic version (excluding within- and between- participant comparison features)



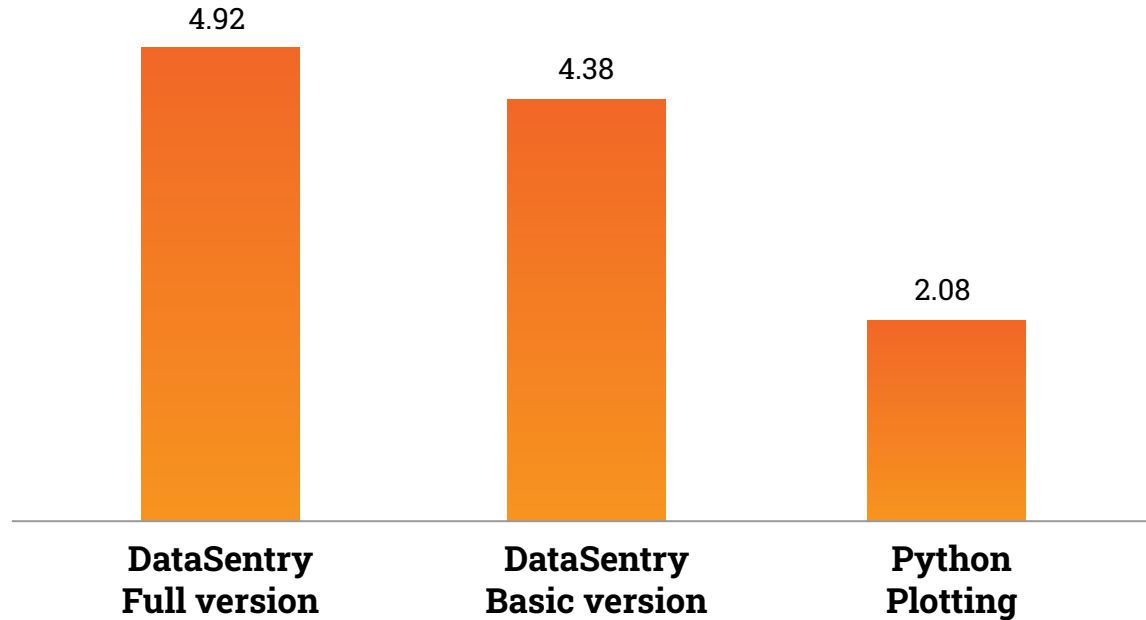
3. DataSentry Full version (including all features)



Within-subject,
in-lab user study

Final Evaluation

PSSUQ (Post-study system usability questionnaire) score (7-point Likert scale)



Final Evaluation

Helpful in managing missing data, specifically...

Overviewing missing data and
diligence of participants



*"...Keeping track of participant **diligence** has always been a key part of our data collection efforts."*

Detection and diagnosis by
within/between-participant
variability



*"By brushing over the empty periods, I could tell if the missing data was an issue, **both within and between participants.**"*

Streamlining **communication**
via rule-based supports



*"I appreciated the ability to **set rules and contact** relevant participants..."*

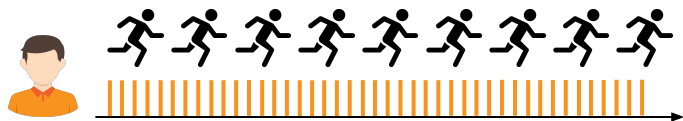
Discussion

Design implication 1

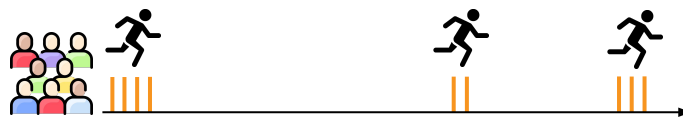
The critical role of understanding **within- and between-participant variability** to detect and diagnose missing data issues



Detecting and diagnosing missing data considering **within- and between-person sensing routines**



Frequent logging along
a participant's commuting path



Sparse logging on weekend mornings
between participants

Discussion

Design implication 2

Researchers wanted to define **diverse rules** related to missing data and **communicate** with participants based on these rules



Enhancing the **expressiveness** of missing data management rules



Semantically
meaningful predicates



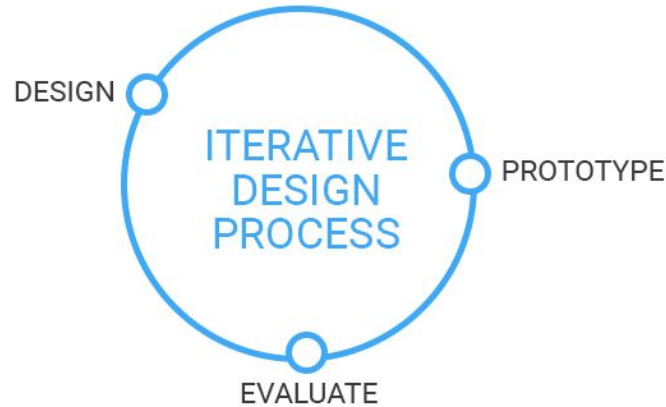
AND/OR conditions



LLM-based rule automation
and communications

Discussion

Lessons learned through multi-year, iterative design process



“The *in-the-wild deployment* and *iterative design process* was instrumental in uncovering and addressing *real-world issues* that might have been overlooked in the lab.”

DataSentry :

Building Missing Data Management System for In-the-Wild Mobile Sensor Data Collection through Multi-Year Iterative Design Approach

Yugyeong Jung, Hei Yiu Law, Hadong Lee,
Junmo Lee, Bongshin Lee, Uichin Lee



← Paper QR code

Contact: yugyeong.jung@kaist.ac.kr

Homepage:

