COMEX: Identifying Mislabeled Human Behavioral Context Data Using Visual Analytics

Hamid Mansoor[1], Walter Gerych[2], Luke Buquicchio[2], Kavin Chandrasekaran[2], Emmanuel Agu[1,2], and Elke Rundensteiner[1,2]

[1]Department of Computer Science
[2]Data Science Program
Worcester Polytechnic Institute
Worcester, USA
{*hmansoor, wgerych, ljbuquicchio, kchandrasekaran, emmanuel, rundenst*}*@wpi.edu*

***Abstract*—Context-Aware (CA) systems that adapt their behavior based on their users' current context have broad applications in areas from healthcare to smart environments. Context Recognition (CR) is currently solved using machine and deep learning approaches that require realistic datasets with ground truth labels collected "in-the-wild" as users live their lives. In such studies, users periodically label their current context (current activity, body position, and location) while a mobile app continuously gathers sensor data. Unfortunately, users sometimes assign wrong or incomplete context labels, reducing the quality of the labeled dataset; and causing lower classification accuracy. We present COMEX, an interactive visual analytics tool that assists analysts in identifying instances of mislabeled context data to improve the quality of CA datasets. For this, we first provide a conceptual categorization of mislabeling error types. Thereafter we develop linked visualizations, augmented by anomaly scores indicating suspected labeling issues, which provide richer insights into the diverse characteristics of the target dataset. We validate our approach on an open source dataset that contains context information for 60 participants gathered over several days using smartphones. With the help of COMEX, participants of our case study identified numerous mislabelled instances in the dataset. We re-ran the classification task after excluding mislabelled data and saw improvements in classification accuracy.**

***Index Terms*—Human Context, Context Recognition, Interactive Visualizations, Mislabelled data, Machine Learning**

I. Introduction

Context-Aware (CA) computing systems adapt their behavior to the users' current context and facilitate exciting new applications. CA systems must first recognize the user's context before they can adapt to it. State-of-the-art CA systems often use machine or deep learning to perform Context Recognition (CR) by classifying data gathered from sensors on the user's smartphone or wearables [1].

Creating accurate CR classifiers requires rich datasets gathered from users while they live their lives (*"in-the-wild"*). Subjects are recruited into context data gathering studies in which they label contexts (e.g. place type, activity) while a mobile application gathers data from their smartphones or smartwatches continuously [2] [3]. Such datasets are used to train CR machine learning classifiers [4]. Most smartphones are able to capture sensor data such as accelerometer, gyroscope, temperature, humidity and light which provide important clues to determine context.

However, the subjects in such CR studies may assign wrong context labels for several reasons. For example, when labeling retroactively, subjects may forget the time and place of their contexts. Some users tend to be careless while labeling the data which may lead them to mislabel their contexts. Such labeling errors reduce the accuracy achievable by CR classifiers.

Human contexts are hard to describe because they involve different factors and so a visual tool that can identify mislabelled data in other domains such as image labelling does not work here. To mitigate this problem, we designed a visual paradigm to relate objective sensor data with human provided labels along with other relevant features. We realized that paradigm by implementing COntext Mislabel EXplorer (COMEX), a web based tool with multiple visually linked panes, to support the exploration and identification of mislabelled instances of context.

Excluding such erroneous labels can improve the accuracy and performance of machine learning classifiers. We validate our approach using data from the ExtraSensory (ES) project, which gathered smartphone and smartwatch sensor data from 60 subjects at the University of California San Diego who provided context labels. Specifically, our work makes the following contributions:

- We categorized various causes of context mislabelling and then identified types of mislabels that they lead to during the collection of the CR data set.
- We designed a visual paradigm to relate context labels (treated as ground truth) with objective measures of their "outlierness", We then integrate this visual paradigm into COMEX.
- We accounted for individual differences in users by calculating "outlierness" of each session against all other sessions of the same label in everyone's data and in that individual user's data.
- We provide results from a case study that illustrates our error identification process and shows improvements in classification accuracy for some labels in the ES dataset.

II. Related Work

A. Human Context Monitoring And Recognition

Using machine learning classifiers, smartphone sensor data can predict human context [5], [6]. Context data can be used to research important implications of human behavior including well being. Wang *et al* [3] used smartphones to gather data to infer the mental health and academic situation of college students for the StudentLife project. He and Agu [4] used the StudentLife dataset to predict sedentary behavior of smartphone users for up to one hour in the future. Such information can be leveraged to design CA systems that aim to mitigate harmful behaviors [1]. Ghods *et al* [7] designed a clinician in the loop monitoring system to track patient behaviors for health care professionals using smart home data.

B. Data Visualizations to Analyze Behavior

Data visualization is a powerful technique to understand human behavior. The time-series nature of human data makes it challenging to analyze and find patterns and trends in behavior, due to the presence of unimportant co-occurring information. Visualizations can help filter important data for analysis. Nguyen *et al* [8] designed a tool called U4 to detect anomalies in user interaction sequences. They proposed the notion of "multi-semantic linking" where they would highlight information in sequences that were semantically similar, thus allowing users to gain information about only relevant event sequences. Polack *et al* [9] created Chronodes, a tool to mine human activity data to compare groups of human behaviors, based on event sequences. Data visualizations can also be used to design systems that aim to intervene and mitigate harmful behaviors. Sharmin *et al* [10] proposed data visualizations as tools to design intervention systems for stress management.

III. Background

A. The ExtraSensory Study

COMEX was designed and evaluated using context data from the ExtraSensory (ES) project [2]. This project recruited 60 subjects to provide "in-the-wild" context data as they lived their lives. Sensor data was continuously gathered from the subjects' smartphones and smartwatches, while they periodically assigned context labels (activity, location/place, phone placement) to the data.

Subjects installed the ES smartphone and smartwatch apps for labeling their contexts. The app had an intuitive interface to label contexts including activities, location/place, phone placement. The app gathered data such as the accelerometer, gyroscope and magnetometer along with audio features from smartphone and smartwatch sensors. The app also collected the phone's battery level, state, light intensity around the phone, humidity and temperature.

During a data gathering "session", the app would continuously run in the background and gather sensor and discrete data for twenty seconds between one minute intervals. The participants had the option of stopping the app from collecting data. The project also gave participants optional smartwatches with the ES app for more sensor data. The watch was paired with the phone application to enable collection for the same twenty second window as the phone.

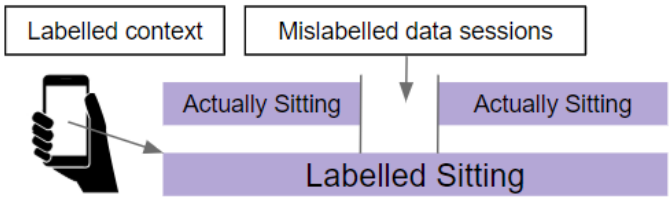

Fig. 1. Example of mislabelled context using a smartphone.

The participants could label their current, previous and future contexts (activity, location/place, phone placement and phone state) using the app's interface and also by responding to notifications. The ES app recorded context as a large number of binary labels that could be true or false at any point in time. Multiple context labels could be labelled to be true at the same time. For example "Driving" while "Sitting" and "Talking". These activities are considered as *co-occurring*.

The app let the users set the frequency at which they received notifications, prompting them to record current or past context labels. The app had a "History Page" that had a list of labels ordered by time, which were provided by the participants and predictions of labels for unlabelled time periods. This view allowed participants to retroactively label their contexts, with the predictions serving as helpful reminders.

This dataset contained mislabeled contexts since it was gathered by humans "in-the-wild". As illustrated in Fig. 1, a participant might label a time period as continuous "Sitting" while in reality there were gaps. Mislabelled data sessions negatively impact machine learning classifiers as these labels are considered to be ground truth. Our goal is to make such classifiers more robust by using a visual analytics approach to identify and remove mislabeled data that "taint" the real data.

B. The ExtraSensory Dataset

The ExtraSensory (ES) project gathered data for 60 participants for an average of 7.6 days each. The raw smartphone sensors included the gyroscope, accelerometer and magnetometer along with discrete measurements such as phone state (battery level, charging vs. unplugged), user location, and phone placement. The app also collected audio for sessions lasting twenty seconds. The app computed 13 Mel Frequency Cepstral Coeffcients (MFCCs). The data was stored on the phone temporarily before being sent to a server. The data was used to calculate 255 features for every twenty-second long data gathering session. The features extracted from the sensor data included measurements such as mean and standard deviation of each time segment for the high frequency sensors such as the accelerometer and gyroscope.

IV. Visualizing Human Context Data

A. Challenges

- *Visualizing the large feature set:* Every twenty second data session was converted into 255 features which makes

it difficult to visualize anomalies.

- *Varied user interpretation:* The participants were told not to modify their lives in any way and label contexts naturally. This meant that there could be discrepancies in the ways people interpret the labels. For example, some users co-labelled being "At Home" and "At School" which might seem odd, but it may have made sense to the users because many of them were students who resided on campus. Different user interpretations could make it hard to establish ground truth.
- *Missing and inconsistent data:* The smartwatch data was unavailable for a large number of sessions which makes it difficult to compare sessions.

B. Goals and Tasks

We analyzed the potential causes of mislabels to design our visual paradigm. We installed the Android version of the ES app [11] on a phone and evaluated the interface to identify use cases that could lead a participant to wrongly label their context.

We identified three broad categories for causes of mislabelling:

- **Recall Bias (C1)**: Human memory is prone to biases and retroactive labelling ("History Page") may lead to some participants mislabelling.
- **Inopportune Interruption (C2)**: Participants may receive notifications at inopportune moments and in their haste to deal with the notification, they may mislabel.
- **Careless Reporting (C3)**: Some participants may not be motivated to do a good job and may not make an effort to accurately recall contexts or provide all relevant labels.

Keeping these causes in mind, we now characterize the types of mislabels that may arise:

- **Wrong Duration (E1)**: Participants may incorrectly recall **(C1)** the start and end times of context; thus under or over estimate the duration of some activity or enter a future activity into the app and then not perform the activity.
- **Incomplete Labels (E2)**: Participants may not report all the relevant context labels **(C1, C3)**. For instance, a user may be "Eating" while "Sitting" but only labels for "Sitting".
- **Wrong Labels (E3)**: Some participants may provide labels that do not make sense and are unlikely to co-occur if interrupted at a bad time **(C2, C3)**. For instance "Driving" while "Standing".

The dataset contained multifaceted context data that needs to be viewed holistically to judge its veracity. We designed a visual paradigm that related contextual information such as time of day, labelling mechanism used and co-occurring labels with objective measures of irregularity of the underlying data. To realize this, we implemented COMEX with several interconnected visualizations that would visualize data at different levels of granularity that allow for cohort-level as well as individual-level views. It is important to show both levels as the habits of some individuals may not be captured at the cohort level.

We designed four tasks that we wanted to accomplish using COMEX:

- **Get an Overview of the Data (T1):** Allowing users to get an overview of all data including the number of data sessions for each label and their commonly co-occurring labels. This cohort level analysis requires minimal interaction.
- **Evaluate Irregular Data (T2):** Showing how unusual some sessions were compared to other sessions. For instance, the "outlierness" of a walking session could be compared to other walking sessions. This measure needs to be displayed in conjunction with other relevant information such as participant ID, time of the activity, co-occurring labels and labelling mechanisms.
- **Filter and Mine Data (T3):** This task allows analysts to filter out irrelevant information and focus on the contexts that they are interested in. The analyst would need a quick way to filter data and display important features since the dataset is massive.
- **Mark Sessions for Exclusion (T4):** The analyst should be able to keep track of all mislabelled data sessions that they discover. These sessions can then be marked and excluded from classification.

C. Calculating Anomaly Measures

To realize our design paradigm, it is important to view contexts with an objective measure of how "normal" the underlying sensor data is for a session compared to other similar instances. We used two anomaly detection methods: a weighted z score (the number of standard deviations from the mean) and the Isolation Forest [12] algorithm. The z score assumes a normal distribution while Isolation Forest assumes that outliers are sparse in the feature space. All the sensor, audio and discrete data were used to compute 255 features for each individual session. Features have different importance values based on their significance for classification; for instance audio features might have a higher importance in detecting a person talking while the magnetometer might be relatively unimportant. For this reason, we weight the anomaly score for each feature by the importance of that feature. We calculated the feature importance values for the "featurized" data by training a Random Forest [13] classifier on the data.

We calculated the z-scores for all individual features and aggregated the product of the z scores with their respective importance value to get a weighted z score for each session. The newly calculated scores are normalized to values between 0 and 1.

$$weighted\_z\_score = \sum_{i=1}^{N} feature\_Z\_score * feature\_imp,$$

where N is the number of features.

The second method used the Isolation Forest algorithm. Isolation forest assumes that outliers occur in a low-density

Fig. 2. COntext Mislabel EXplorer (COMEX).

space compared to the inliers. The algorithm is built on an ensemble of isolation trees, where each isolation tree randomly splits the data until every data point is uniquely identified. The average number of splits required to isolate a point is used to calculate the anomaly score. We set a ”contamination parameter” that represents an approximation of the expected percentage of data believed to be anomalous to 0.2. The calculated values are normalized to values between 0 and 1.

Humans have different styles of performing Activities of Daily Living (ADL) such as walking, speaking and phone placement habits. This variation presents an issue when calculating anomaly measures for all the data across a diverse group of people, which is why we used “Global” and “Individual” approaches for the scores. For the “Global” approach, we ran the two algorithms to generate scores for all the labels collected across all participants. For the “Individual” approach, we calculated the anomaly scores for the labels for each individual participant’s data and not across the entire dataset. Every data session has four anomaly scores associated with it because of the two algorithms and the two approaches. The analyst is able to select the desired score.

V. COMEX: The Context Mislabel Explorer

We researched and developed COMEX, a web-based tool using a JavaScript library called D3.js [14], which realizes the above requirements. The pane on the left (Fig. 2A) presents an overview of all gathered labels. The user selects labels in this pane for further analysis in the two panes on the right in Fig. 2 (split horizontally). Next, we describe key features of COMEX designed to achieve the tasks described in Section IV B.

A. Overview of Labels Collected

The pane in Fig. 2 A provides an overview of all the collected context labels. The labels are ordered from left to right and from top to bottom by the number of data sessions collected. For example, ”Indoors” and ”Sitting” labels had the highest and third highest numbers respectively. Hovering over a label’s circle causes co-occurring labels (other labels that had happened at the same time) to also get highlighted. The amount of fill in the circle represents the proportion of times that the highlighted labels co-occurred with the hovered over label. In Fig. 2 A, the user is hovering over ”Walking” and the co-occurring labels for ”Walking” are shown.

Labels with no instance of co-occurrence are greyed out. This pane performs task **T1** by providing a quick overview of the data without much interaction as the users can visually discern co-occurrence frequencies by the level of fill, quickly find unlikely co-occurrences (**E3**) and also see trends that make intuitive sense, such as “Indoors and “Sleeping” co-occurring often (**E2**). Using circles and ordering them makes it easier to compact the fifty one labels into limited space. The user can select a label by clicking on it which changes the pane in Fig. 2 A to show the label’s co-occurring labels, ordered by frequency from left to right and top to bottom. The amount of fill in the circles represents the frequency of co-occurrence.

Clicking on a circle also populates the pane in Fig. 2 C with bars representing continuous ”chunks” of the selected label. These represent periods of time in the dataset, when successive data sessions for a participant had the same label as the one that the user clicked in Fig. 2 A (details in section V B). The user can filter (enabling **T2**) the chunks shown in that pane (Fig. 2 C) by clicking on a co-occurring label to only show the continuous activity chunks with at least one instance of the co-occurring label being present.

B. Continuous Label Chunks

The pane in Fig. 2C displays continuous ”chunks” for the selected label. These chunks are ordered from left to right and from top to bottom by their lengths. These chunks represent periods where the time difference between successive or back-to-back collected data sessions was less than or equal to sixty seconds and had the same label assigned by the same

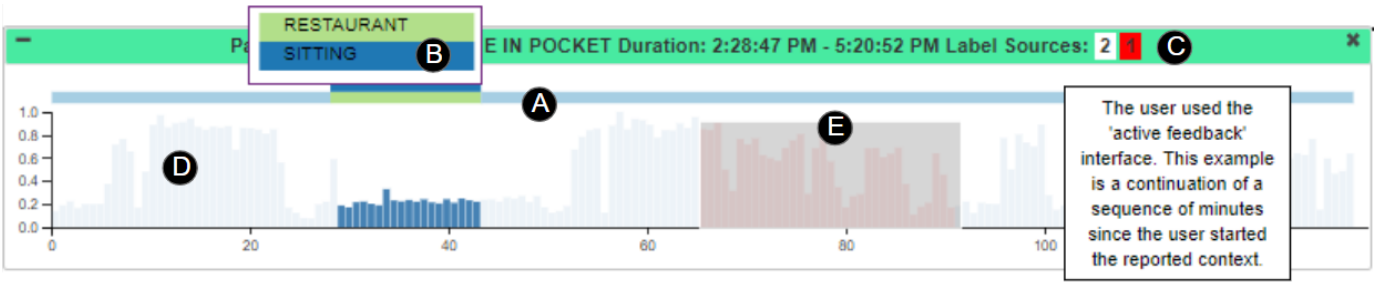

Fig. 3. Label Chunk View.

participant. Since the app sent data after approximately every minute, it is reasonable to consider successive sessions with the same label as continuation of that context. The length of the chunks represents the duration of the continuous label instance or chunk. Encoding duration by length and then ordering chunks makes it easier to compare relative durations. The opacity of the chunks represents the average anomaly score (between 0 - 1) for sessions comprising the chunk. This supports task **T2** and **T3** as the user can focus on the darker ones which contain anomalies. The user can update the chunk opacity by selecting the anomaly score algorithm and calculation mode in Fig. 2 B.

Clicking a chunk causes the Label Chunk View (Section V C) to show on the pane (Fig. 2 D). This then shows the detailed information about the underlying sessions comprising the chunk.

C. Label Chunk View

The Label Chunk View (LCV) (Fig. 3) drills into continuous chunks to see the underlying data. The user can see co-occurring labels represented as horizontal lines above the bars (Fig. 3 A). The user can hover over the lines to see labels and the colors representing them (Fig. 3 B). This helps us to find labels that are unlikely to co-occur (**E3**). The anomaly scores for the sessions in the chunk are plotted as bars with values between 0 and 1 (Fig. 3 D) (**T2**) (**E1, E3**). The score calculation method and algorithm can be changed using Fig. 2 B. The continuity of the colored bars in Fig. 3 A shows the continuous co-occurrence of the respective label. If the user wants to view that continuous label chunk with the same pane layout, they can click on the horizontal line for the co-occurring label and a new pane with the information for the co-occurring label chunk will be appended to the pane in Fig. 2 D.

The header contains the duration of the continuous label, the user's ID (shortened to fit), time of day for the session and the labelling mechanisms. The numbers in Fig. 3 C represent the different label reporting sources used to label the data sessions. There were eight categories for the reporting mechanism such as using the "History Page", replying to notifications etc. The dataset used numbers between -1 to 6 to encode categories. Hovering over a number shows a tooltip (Fig. 3 C) with information about the mechanism. Clicking on a number highlights the bars for the sessions which were labelled using the mechanism encoded by the number. In Fig. 3, the user selected "1" which highlights the bars in blue, with this reporting mechanism, and reduces the opacity of others. The analyst can find instances of **E1** and **E2** by showing information about how the participant interacted with the phone when they labelled. For example, Fig. 3 shows the information for a continuous chunk of "Phone in pocket". We can highlight the sessions that were labelled using "active feedback" (encoded as "1") which is when the participant provides labels on the app and sets a duration. The co-occurring labels change during this chunk: "Walking" to "Restaurant" and"Sitting", and then back to "Walking". The anomaly scores for the "Phone in pocket" are consistently low for the highlighted sessions and a few sessions beyond. The anomaly scores are variable for other sessions which were marked using the "History page" (encoded by "2") on the app. Such detailed, fine grained exploration is meant to aid in finding errors **E1**, **E2**, which may be the result of recall bias and careless reporting.

The user may conclude that they have found mislabeled data which they want to exclude (**T4**). The user can drag the mouse over the data sessions that they want to mark and then confirm a pop up. The marked sessions are colored red to indicate their selection as shown in Fig. 3 E. The user can view the selected data sessions by clicking on "Marked Sessions" in the pane in Fig. 2 A.

VI. Evaluation

A. Case Study

In a small case study, we utilized COMEX to help us identify and then remove mislabels for 13 labels. We chose these labels because they have some unusual co-occurring labels that did not make sense; for example, "Driving" while being at "Home" and also because we could select enough sessions to make a difference in classification accuracy. The dataset is unbalanced and there were some labels for which we would not have been able to select enough sessions to noticeably change classification results.

We marked the data sessions keeping in mind the errors and tasks discussed previously. We looked at unlikely co-occurring labels, the calculated anomaly scores, label reporting mechanisms used and time of day. After the data sessions were selected, they were marked in the original dataset and

Label	Number of Marked Sessions	Total Sessions	Original AUC-ROC	Recalculated AUC-ROC with Marked Sessions Removed	Change
Exercise	2111	8081	85.82	91.74	5.92
Toilet	911	2655	75.08	78.10	3.02
Passenger	436	2526	85.31	86.78	1.47
Walking	1756	22136	87.83	88.84	1.01
Phone in hand	2121	14573	75.77	77.01	2.24
Driving	2955	7975	93.19	95.71	2.52
Bicycling	1010	5020	93.59	95.54	1.95
Outside	1369	12114	91.98	92.79	0.81
On bus	305	1794	86.81	89.07	2.26
Running	201	1090	78.99	76.31	-2.68
Shower	661	2087	73.75	75.75	-1.00
Cleaning	1101	3806	63.60	63.21	-0.39
Standing	730	6224	74.90	74.75	-0.15

TABLE I
RECLASSIFICATION RESULTS

excluded from any subsequent machine learning training and classification tasks.

For classifying the labels, we used the Gradient Boosted Machines Classifier [13]. We re-ran a 5-fold cross validation for the selected labels and computed the Area Under the Curve of the Receiver Operating Characteristic (AUC-ROC) measure. The AUC-ROC measures how effectively our classifier can distinguish between classes, where a random classifier has an AUC-ROC of 0.5. We chose the AUC-ROC over standard accuracy as the AUC-ROC is more robust to class imbalance.

We observed improvements in the classifications results for 9 out of the 13 selected labels (see Table 1), which indicates that overall, our approach worked.

VII. CONCLUSIONS

We introduced a novel visual analytics solution, COMEX, for detecting human labeled behavior data traces that contain mislabeled context data. Our evaluation study demonstrates the effectiveness of our solution by supporting analysts in exploring human context data and identifying mislabeled data sessions. In particular, we demonstrated that removing mislabelled data can improve labeled datasets that are amenable for training higher quality models for more accurately classifying human context data. Future work includes employing additional anomaly detection algorithms to identify outlier data and using different data visualization strategies. Our approach of using visual analytics to explore human labeled data may also be applied to other application domains.

ACKNOWLEDGMENT

The work presented in this paper was funded by DARPA WASH program HR001117S0032

REFERENCES

[1] He and Agu, "A context-aware smartphone application to mitigate sedentary lifestyle," 2014.

[2] Y. Vaizman, K. Ellis, and G. Lanckriet, "Recognizing detailed human context in the wild from smartphones and smartwatches," *IEEE Pervasive Computing*, vol. 16, no. 4, pp. 62–74, 2017.

[3] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proc Ubicomp*. ACM, 2014, pp. 3–14.

[4] Q. He and E. O. Agu, "Smartphone usage contexts and sensable patterns as predictors of future sedentary behaviors," in *Healthcare Innovation Point-Of-Care Technologies Conference (HI-POCT), 2016 IEEE*. IEEE, 2016, pp. 54–57.

[5] D. Micucci, M. Mobilio, and P. Napoletano, "Unimib shar: A dataset for human activity recognition using acceleration data from smartphones," *Applied Sciences*, vol. 7, no. 10, p. 1101, 2017.

[6] S. Ranakoti, S. Arora, S. Chaudhary, S. Beetan, A. S. Sandhu, P. Khandnor, and P. Saini, "Human fall detection system over imu sensors using triaxial accelerometer," in *Computational Intelligence: Theories, Applications and Future Directions-Volume I*. Springer, 2019, pp. 495–507.

[7] A. Ghods, K. Caffrey, B. Lin, K. Fraga, R. Fritz, M. Schmitter-Edgecombe, C. Hundhausen, and D. J. Cook, "Iterative design of visual analytics for a clinician-in-the-loop smart home," *IEEE journal of biomedical and health informatics*, 2018.

[8] P. H. Nguyen, C. Turkay, G. Andrienko, N. Andrienko, O. Thonnard, and J. Zouaoui, "Understanding user behaviour through action sequences: from the usual to the unusual," *IEEE transactions on visualization and computer graphics*, 2018.

[9] P. J. Polack Jr, S.-T. Chen, M. Kahng, K. D. Barbaro, R. Basole, M. Sharmin, and D. H. Chau, "Chronodes: Interactive multifocus exploration of event sequences," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 8, no. 1, p. 2, 2018.

[10] M. Sharmin, A. Raij, D. Epstien, I. Nahum-Shani, J. G. Beck, S. Vhaduri, K. Preston, and S. Kumar, "Visualization of time-series sensor data to inform the design of just-in-time adaptive stress interventions," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 505–516.

[11] Y. Vaizman, K. Ellis, G. Lanckriet, and N. Weibel, "Extrasensory app: Data collection in-the-wild with rich user interface to self-report behavior," in *Proc CHI 2018*. ACM, 2018, p. 554.

[12] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008, pp. 413–422.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[14] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 data-driven documents," *IEEE Transactions on Visualization & Computer Graphics*, no. 12, pp. 2301–2309, 2011.