Survey of Experience Sampling Method in Big Data - AI Integration Perspective

TAECKYUNG LEE, KAIST, Republic of Korea

DISCLAIMER: This is a course project report of KAIST EE616 (Prof. Steven Euijong Whang). This report is NOT peer-reviewed.

The experience sampling method is a widely used data labeling technique for studying ongoing experiences. The experience sampling method relies on participants' self-reports to collect data, which can be prone to human error and missing data. In this paper, we identify noise sources in the experience sampling method and propose solutions to improve the quality of the collected data. We also discuss how recent advances in machine learning, such as semi-supervised learning and robust training, can be applied to data generated by the experience sampling method to improve the accuracy and fairness of the results. Overall, our work provides insights into the limitations of the experience sampling method and suggests ways to enhance its effectiveness for generating high-quality data.

ACM Reference Format:
Taeckyung Lee. 2022. Survey of Experience Sampling Method in Big Data - AI Integration Perspective. 1, 1 (February 2022), 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Artificial intelligence (AI), including deep learning, has recently gained great attention. From traditional domains of computer vision, natural language processing, and speech processing, AI is now applied to human-computer interaction (HCI) domains of psychology [39], physiology [11, 51], emotion sensing [48], and digital health [40].

Since humans are primarily involved in the loop of data acquisition, labeling, and improvement, there is a higher chance of human error and bias in the data. Recent work focused on data quality, creation, management, and analysis in HCI domains to solve the issue. For example, the CHI 2022 workshop discussed how humans create, collect, manage, curate, analyze, interpret, and communicate data [50].

Data labels are often from human annotations[1], therefore understanding human factors in labeling is important. The main limitation of existing literature on human factors in data labeling is that they mostly consider the data annotated/generated from crowdsourcing. For example, existing works cover data annotators [55], data annotation infrastructure [74], annotation tools [83], or end-to-end (data annotation to model generation) [8, 70] - all in the crowdsourcing context.

Although crowdsourcing has been the main line of research, the experience sampling method (ESM) also have been widely used as the human-in-the-loop data labeling in HCI-related domains as in Figure 1. Human factors of ESM are also crucial for data quality. For example, imagine building a smartphone system to predict emotions based on smartphone usage records. You must first collect the data of automatically-recorded smartphone usage logs with

[1]For example, 65% of machine learning papers of Twitter data utilize labels from human annotations [21].

Author's address: Taeckyung Lee, taeckyung@kaist.ac.kr, KAIST, Daejeon, Republic of Korea.

Manuscript submitted to ACM

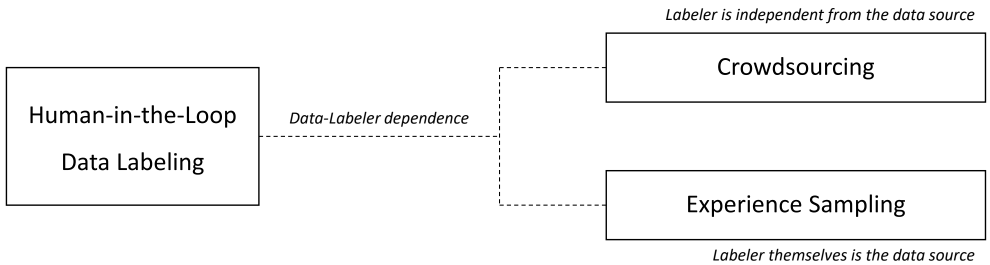

Fig. 1. Overview of human-in-the-loop data labeling dependence on the data source.

corresponding user emotions, which must be self-reported. Such self-reporting data is known to be noisy likewise [32]. However, such human factors in ESMs are underexplored.

Therefore, we focus on human factors affecting data quality in ESM data in this survey. We intensively analyze existing approaches from the cause of noisy experience sampling data to its solutions. We provide guidelines for preventing noisy labels and methods from validating and cleaning the data and improving the models.

2 RELATED WORKS

2.1 Big Data - AI Integration

Big data - AI integration, or data-centric AI, is a research trend focusing on the data itself to improve the AI system, focusing on the data point of view in the machine learning pipeline, including data collection, data cleaning, validation, and integration, robust model training, and fairness [52, 56, 79].

Data collection in big data - AI integration is categorized as three approaches: data acquisition, data labeling, and improving existing data and models [56, 79]. If there is insufficient data, data acquisition must be the first option, including data discovery, augmentation, and generation. Data labeling includes utilizing existing labels (semi-supervised learning), manual labeling (crowdsourcing and active learning), and automatic labeling (weak supervision). Finally, improving existing data aims to improve poor-quality datasets. For example, popular benchmark datasets in ML applications still include label errors [12, 46], where high-quality small-size data could perform better than a low-quality large-size dataset [46].

The importance of data quality, including label reliability, is explored with various ML efforts [22, 35, 49, 80]. However, without partnerships with data-specific domain experts, it can be difficult for ML experts to validate the accuracy and efficacy of existing labeling operations, which can affect the performance and validation of models [70].

2.2 Crowdsourcing

The crowdsourcing approach distributes data labeling tasks to various digital workers and is widely used in multiple AI fields [21, 24, 30]. For example, Amazon Mechanical Turk [2] is one of the most popular platforms for connecting human workers and task requesters. As the role of human factors in crowdsourcing is critical, there is heavy literature (including surveys) in the view of big data - AI [1, 13, 15, 41, 58] and human-computer interaction [8, 21, 43, 44, 55, 70, 75].

2.2.1 Collaborative Workflow. One area of research has focused on the collaborative workflows of data scientists, ML developers, and data annotators [8, 21, 43, 44, 70]. For example, validating, cleaning, and re-labeling could improve the performance of classification algorithms, but this requires the data scientist to have specific domain knowledge [43]. ML developers also reported challenges in understanding the (1) subjectivity and biases of those who labeled the data and (2) whether labeling was performed under the awareness of the data context [70]. Therefore, it would be essential to prevent any possible biases and errors and preprocess the crowdsourced labels with corresponding domain knowledge during the data creation step.

2.2.2 Data Quality. Controlling the data quality of crowdsourced data could be addressed by simple approaches such as repetitive labeling and majority voting. However, previous methods of repetitive labeling or majority voting require more budget resources. David et al. [34] provide a new crowdsourcing system to achieve reliability with minimizing the budget by an expectation-maximization (EM) algorithm. Also, identifying and filtering the low-quality labelers could achieve better labeling quality (e.g., Vox Populi [17]).

2.2.3 User Interaction. The interface for crowdsourcing is critical for accurate and fast labeling; designing and programming the interface has been identified as time-consuming, costly, and requiring expertise in software engineering. OneLabeler [83] is a flexible system to build data labeling tools through visual programming. Also, CrowdER [76] is the solution for generating labels of comparison of entities using pair-based and cluster-based interfaces.

Also, providing practical instructions is essential to prevent confusion and provide labeling consistency. One approach is providing guidelines before the labeling and letting labelers follow the guidelines. However, such guidelines could be incomplete and do not cover all possible scenarios [56]. Revolt [9] is a collaborative crowdsourcing platform to solve the problem, where workers vote, explain, and categorize the labeling to make post-hoc decision boundaries.

2.2.4 Organazation and Workers. Finally, the human-computer interaction field focuses on organizations and workers in a crowdsourcing context. For example, the organization and infrastructure of crowdsourcing, including its potential to be seen as organized employment, has been investigated [75]. The work practices and career goals of annotators have also been studied, and challenges faced by crowd workers on platforms such as Amazon Mechanical Turk, such as a lack of career advice and limited time and financial resources, have been highlighted [55].

3 NOISE SOURCES IN EXPERIEMNT SAMPLING

The experience sampling method (ESM) is a widely-used data generation and labeling method for collecting individuals' ongoing experiences in human-computer interaction fields [31]. It uses two primary ways: probe-caught and self-caught [32, 63, 64]. In the probe-caught method, participants are periodically interrupted and required to respond to their ongoing experience. These interruptions can be alerts, pop-up screens, or face-to-face interactions. In the self-caught method, participants voluntarily provide reports. Both ways can be combined to improve the diversity of the data collected.

In this section, we summarize the significant noise sources in the experience sampling method (ESM), as shown in Figure 2. We begin by discussing noise sources in the experience sampling method and then examine noise sources specific to probe-caught and self-caught processes.

3.1 Experience Sampling Method

The primary noise sources in the experience sampling method (ESM) are self-reported data [14]. This can lead to (1) incorrect responses due to human error, (2) missing labels due to intentional or unintentional causes, and (3) response shifts.

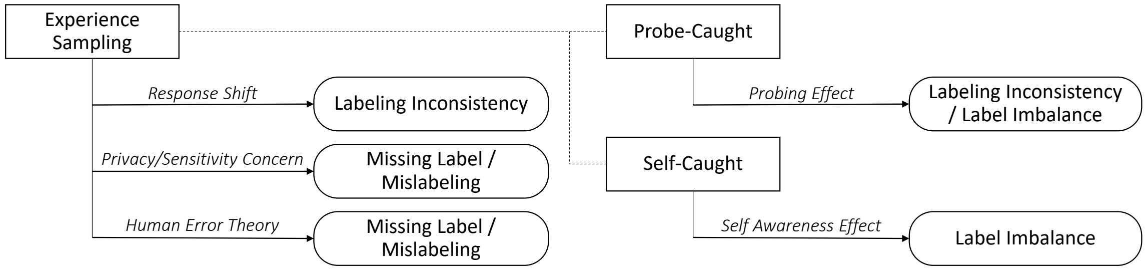

Fig. 2. Overview of labeling noise sources in experience sampling method.

Human error can result in inaccurate data. According to the human error theory [45, 47], momentary slip or lapse of attention could lead to unexpected behaviors. The theory suggests that errors are not simply the result of individual incompetence or carelessness but rather are influenced by a complex interplay of cognitive, social, and environmental factors [45]. One example could be poor instructions and guidelines, or the task is beyond the physical or mental ability of the person [37].

Response shift, which refers to changes in individuals' internal standards, values, priorities, or definitions, can occur in experience sampling studies [6]. This phenomenon is widespread in longitudinal experience sampling studies [54, 60, 68]. Response shifts can affect the validity of experience sampling studies by changing how participants report their experiences. For example, a participant may initially report feeling anxious about a particular situation. Still, after some time, they may no longer feel worried about it and instead report feeling calm. This change in self-report does not necessarily reflect a difference in the participant's actual emotional experience but rather a change in their perception of that experience.

Missing labels can occur if participants intentionally or unintentionally do not respond to the probes [14]. First, participants could suffer from frequent actions required. For example, study dropout rates are relatively high in ESM-related experiments [71]. For example, one participant from Kang et al. [32] ignored the ESM probes if the probe interrupted the primary task. Also, participants could avoid or even falsely report the response due to privacy/sensitivity concerns. For an example of daily activity data collection, one might refuse to report personal activities or even falsely report the response to hide the activity.

Participants could also be affected by the Hawthorne effect, where participants alter behaviors as they realize they are being observed. For example, participants could alter the voice conversations when they are reminded that their conversation is being recorded on smartphones [53].

The label could be biased toward time. For the daily diary case of ESM, participants could fill the response days after an event instead of immediately reporting [62]. Also, with insufficient guidelines, participants could be confused about the time window of the response, which would lead to any unknown bias towards time.

3.2 Probe-Caught Method

The use of probe-caught methods in experience sampling method (ESM) has been prevalent in the field of human-computer interaction (HCI) to collect a sufficient amount of data [7, 28, 32, 33, 39, 42, 57, 77]. Providing the probe could be heuristic-based (e.g., periodic probing, random probing), event-based (e.g., smartphone app launch [57]), or rule-based (e.g., active learning [33]).

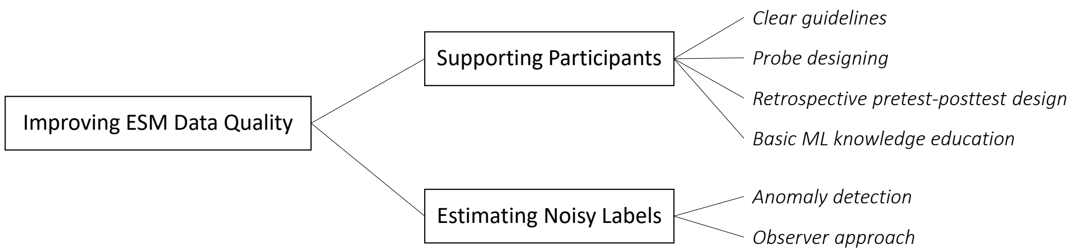

Fig. 3. Overview of methods to improve data quality.

However, using arbitrarily-invoked alerts in these methods can interrupt participants' ongoing tasks and affect their mental states. Unexpected interrupts are known to interfere with people's ongoing tasks, decrease user efficiency, and even affect users' mental states [3, 4, 29].

As a result, participants' psychophysiological responses may be affected by using probe-caught methods in ESM. For example, Kang et al. [32] reported that ESM positively or negatively affected mobile user emotions in at least 38% of emotions by analyzing 2,227 samples of mobile ESM data from 78 participants. Also, a few participants from online lecture viewing experiment [39] reported that periodic probing requests either disturbed or made to focus more.

3.3 Self-Caught Method

In the self-caught method of experience sampling [27, 69], data quality may be compromised due to the reliance on participants' self-reports. This can result in data imbalance and a lack of reports on specific experiences. To improve the quality of data collected using the self-caught method, researchers may need to develop strategies for encouraging more balanced reporting from participants. This could include providing clear guidelines and incentives for participants to report all relevant experiences and implementing techniques such as data under- or oversampling to balance the dataset.

4 IMPROVING DATA QUALITY

Data quality is an essential factor in machine learning systems. Poor data quality can hinder the accuracy of the model, as noted in [61]. Therefore, the experimenter must consider data quality throughout the process, from experiment planning to machine learning. We first define existing data quality measurement approaches in Section 4.1. Then, we provide methods to improve data quality in experience sampling data in Section 4.2 and Section 4.3 as in Figure 3.

4.1 Addressing Data Quality

Few works of literature consider the data quality of the experience sampling method (ESM) response. Yue et al. [82] performed an ESM experiment to ask participants to submit text-based ESM with optional photos. The authors evaluated the data quality with (1) the rate of incomplete/invalid ESM responses and (2) the measured length of the text response. As a result, photo-sharers (who submitted the optional photo to describe the ESM text response) resulted in higher data quality than non-photo-sharers.

Also, Hicks et al. [26] measures the 'participant quality' as the number of interactions and discover that power users (ESM participants whose responding device is their primary device) result in higher data quality compared to survey-only users.

4.2 Supporting Participants

Data quality is primarily dependent on the participants. To reduce human errors in the experience sampling method (ESM), it is essential to provide clear guidelines for data labeling and educate annotators on their work's importance. Previous research has shown that giving detailed examples is more effective than simply explaining guidelines [8]. This can help to motivate and improve the productivity of annotators.

Designing the probing process is crucial. The probing method and interface should minimize the burden on participants by adjusting the number of questions, daily alerts, and question types [36]. In addition, the experience sampling method itself can be adjusted [19], therefore not to fatigue the participants for efficiency and consistency of annotations [8]. Customizable data collection tools [83] or gamification [72] could also improve labeling quality. In addition, making ESM tools monitorable can help managers track outliers, monitor participants, and enforce annotation guidelines [8].

To reduce response shifts (see Section 3.1), a commonly used technique is the 'retrospective pretest-posttest design' [59, 67], where participants periodically answer the questionnaire to reflect on how they were doing at the start of the study. This allows the experimenter to re-calibrate in light of changes in perspective over the study period.

Cha, Oh, and Park et al. [8] stated that data quality could vary based on human labelers' diversity and background knowledge. Therefore, to improve the data quality in ESM, we can educate basic ML knowledge to participants to enhance labeling quality as in labeling [5, 84].

4.3 Estimating Noisy Labelers

Unreliable participants can be detected using anomaly detection and statistical methods to improve data quality. For example, existing human-computer interaction (HCI) works used low response rates (missing data rates) to remove unreliable participants from the training/testing dataset [32, 39]. A short response time can also indicate skipping behavior [71], where the characteristics of the experiment determine the threshold.

Another approach is to use an 'observer,' who observes and labels the participant's state independently of the participant. For example, the K-EmoCon dataset [48] provides emotion labeling from experimenters and independent observers, where the observer infers the experimenter's emotion from their facial expressions. With observers, we can apply existing approaches for dealing with noisy labelers in the big data community to improve the data [16, 18, 23, 61, 65, 66, 80]. For example, cross-replication reliability (xRR) [80] could be utilized to measure the quality of crowdsourced datasets with high cultural and training variances of annotators.

5 IMPROVING MODEL

Although we utilize various techniques to improve data quality, the experience sampling method (ESM) cannot eliminate underlying human errors compared to crowdsourcing. Therefore, it is essential to build robust and fair models in the presence of underlying noisy data.

5.1 Semi-Supervised Learning

Semi-supervised learning is a deep learning technique that enables models to learn from a few labeled samples and apply that knowledge to unlabeled samples. For instance, Wampfler et al. [73] used semi-supervised learning to predict affective states based on smartphone keyboard usage. Since affective states can only be inferred through periodic probes, they used semi-supervised learning to fill the gap between a few labeled and many unlabeled samples.

5.2 Robust Training

There has been extensive research on training models with noisy labels [20]. Most existing methods rely on a small amount of clean data. However, it is difficult to assume that a small amount of clean data is available in the experience sampling method (ESM) because ESM samples each individual's label. Existing model training approaches could be applied if the 'observer' approach (see Section 4.3) were utilized. For instance, Xiao et al. [81] proposed a general framework for training a neural network with few clean and noisy labels by modeling the relationships between data, labels, and noise.

Addressing the issue of data imbalance, particularly in the self-caught method, could improve training performance. The survey by He and Garcia [25] provides traditional approaches for learning from imbalanced data. A common practice is to under or oversamples the data by replicating or removing data. SMOTE [10] is an alternative that oversamples the minority classes by copying and generating synthetic samples using minority examples.

5.3 Fairness

Algorithmic fairness, also known as fair machine learning, has attracted significant attention due to the potential social impact of AI. The goal of fairness is to produce unbiased results concerning specific protected variables. However, suppose the experience sampling method (ESM) questionnaire contains potentially protected variable candidates. In that case, participants may provide inaccurate responses due to human error or concerns about sensitivity and privacy (see Section 3). Recent work on robust fairness in binary classification has shown the potential benefits of using noisy data from sensitive experiences [38, 78]. However, robust fairness in multi-class classification and regression remains an under-researched area.

6 CONCLUSION

The increasing trend toward integrating big data and AI has highlighted the importance of data quality. However, while crowdsourcing has been extensively studied, the experience sampling method has received relatively little attention regarding its potential for generating noisy data. In this paper, we identify critical noise sources in the experience sampling method and propose methods for improving the quality of the data generated. We also discuss how recent advances in machine learning, such as semi-supervised learning and robust training, can be applied to data generated by the experience sampling method to address these noise sources and improve the results' accuracy. Overall, our work provides insights into the limitations of the experience sampling method and suggests ways to enhance its effectiveness for generating high-quality data in human-computer interaction.

REFERENCES

[1] Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. 2013. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing* 17, 2 (2013), 76–81.

[2] Amazon. 2022. Amazon Mechanical Turk. https://www.mturk.com/. Accessed: 2022-12-20.

[3] Brian P Bailey and Shamsi T Iqbal. 2008. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction (TOCHI)* 14, 4 (2008), 1–28.

[4] Brian P Bailey, Joseph A Konstan, and John V Carlis. 2001. The Effects of Interruptions on Task Performance, Annoyance, and Anxiety in the User Interface.. In *Interact*, Vol. 1. 593–601.

[5] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)* 41, 3 (2009), 1–52.

[6] Niels van Berkel and Vassilis Kostakos. 2021. Recommendations for Conducting Longitudinal Experience Sampling Studies. In *Advances in Longitudinal HCI Research*. Springer, 59–78.

[7] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Pentland. 2014. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the 22nd ACM international conference on Multimedia*. 477–486.

[8] Inha Cha, Juhyun Oh, Cheul Young Park, Jiyoon Han, and Hwalsuk Lee. 2022. The Grind for Good Data: Understanding ML Practitioners' Struggles and Aspirations in Making Good Data. *arXiv preprint arXiv:2211.14981* (2022).

[9] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2334–2346.

[10] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[11] Weixuan Chen and Daniel McDuff. 2018. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*. 349–365.

[12] Kate Crawford and Trevor Paglen. 2021. Excavating AI: The politics of images in machine learning training sets. *Ai & Society* (2021), 1–12.

[13] Valter Crescenzi, Paolo Merialdo, and Disheng Qiu. 2015. Crowdsourcing large scale wrapper inference. *Distributed and Parallel Databases* 33, 1 (2015), 95–122.

[14] Mihaly Csikszentmihalyi and Reed Larson. 2014. *Validity and Reliability of the Experience-Sampling Method.* Springer Netherlands, Dordrecht, 35–54. https://doi.org/10.1007/978-94-017-9088-8_3

[15] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)* 51, 1 (2018), 1–40.

[16] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28, 1 (1979), 20–28.

[17] Ofer Dekel and Ohad Shamir. 2009. Vox Populi: Collecting High-Quality Labels from a Crowd.. In *22nd Annual Conference on Learning Theory (COLT)*.

[18] Mohamad Dolatshah. 2018. *Cleaning crowdsourced labels using oracles for statistical classification.* Ph. D. Dissertation. Applied Sciences: School of Computing Science.

[19] L Feldman-Barrett and DJ Barrett. 2001. Computerized experience-sampling: How technology facilitates the study of conscious experience. *Social Science Computer Review* 19 (2001), 175–185.

[20] Benoît Frénay and Michel Verleysen. 2013. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems* 25, 5 (2013), 845–869.

[21] R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 325–336.

[22] Venkat Gudivada, Amy Apon, and Junhua Ding. 2017. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software* 10, 1 (2017), 1–20.

[23] Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology* 8, 1 (2012), 23.

[24] Kotaro Hara, Vicki Le, and Jon Froehlich. 2013. Combining crowdsourcing and google street view to identify street-level accessibility problems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 631–640.

[25] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.

[26] John Hicks, Nithya Ramanathan, Donnie Kim, Mohamad Monibi, Joshua Selsky, Mark Hansen, and Deborah Estrin. 2010. AndWellness: An Open Mobile System for Activity and Experience Sampling. In *Wireless Health 2010* (San Diego, California) *(WH '10)*. Association for Computing Machinery, New York, NY, USA, 34–43. https://doi.org/10.1145/1921081.1921087

[27] Victoria Hollis, Artie Konrad, Aaron Springer, Matthew Antoun, Christopher Antoun, Rob Martin, and Steve Whittaker. 2017. What does all this data mean for my future mood? Actionable analytics and targeted reflection for emotional well-being. *Human–Computer Interaction* 32, 5-6 (2017), 208–267.

[28] Karen Hovsepian, Mustafa Al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 493–504.

[29] Shamsi T Iqbal and Brian P Bailey. 2005. Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. In *CHI'05 extended abstracts on Human factors in computing systems*. 1489–1492.

[30] Humayun Irshad, Eun-Yeong Oh, Daniel Schmolze, Liza M Quintana, Laura Collins, Rulla M Tamimi, and Andrew H Beck. 2017. Crowdsourcing scoring of immunohistochemistry images: Evaluating performance of the crowd and an automated computational method. *Scientific reports* 7, 1 (2017), 1–10.

[31] Daniel Kahneman, Alan B Krueger, David A Schkade, Norbert Schwarz, and Arthur A Stone. 2004. A survey method for characterizing daily life experience: The day reconstruction method. *Science* 306, 5702 (2004), 1776–1780.

[32] Soowon Kang, Cheul Young Park, Auk Kim, Narae Cha, and Uichin Lee. 2022. Understanding Emotion Changes in Mobile Experience Sampling. *CHI Conference on Human Factors in Computing Systems* 14, 1–14. https://doi.org/10.1145/3491102.3501944

[33] Ashish Kapoor and Eric Horvitz. 2008. Experience sampling for building predictive user models: a comparative study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 657–666.

[34] David Karger, Sewoong Oh, and Devavrat Shah. 2011. Iterative learning for reliable crowdsourcing systems. *Advances in neural information processing systems* 24 (2011).

[35] Ramesha Karunasena, Mohammad Sarparajul Ambiya, Arunesh Sinha, Ruchit Nagar, Saachi Dalal, Hamid Abdullah, Divy Thakkar, Dhyanesh Narayanan, and Milind Tambe. 2021. Measuring Data Collection Diligence for Community Healthcare. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–12.

[36] Predrag Klasnja, Beverly L Harrison, Louis LeGrand, Anthony LaMarca, Jon Froehlich, and Scott E Hudson. 2008. Using wearable sensors and real time inference to understand human recall of routine activities. In *Proceedings of the 10th international conference on Ubiquitous computing*. 154–163.

[37] Trevor A Kletz. 1993. *Lessons from disaster: how organizations have no memory and accidents recur.* IChemE.

[38] Alex Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. 2019. Noise-tolerant fair classification. *Advances in Neural Information Processing Systems* 32 (2019).

[39] Taeckyung Lee, Dain Kim, Sooyoung Park, Dongwhi Kim, and Sung-Ju Lee. 2022. Predicting Mind-Wandering with Facial Videos in Online Lectures. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2103–2112. https://doi.org/10.1109/CVPRW56347.2022.00228

[40] Uichin Lee, Gyuwon Jung, Eun-Yeol Ma, Jin San Kim, Heepyung Kim, Jumabek Alikhanov, Youngtae Noh, and Heeyoung Kim. 2022. Toward Data-Driven Digital Therapeutics Analytics: Literature Review and Research Directions. *IEEE/CAA JOURNAL OF AUTOMATICA SINICA* (5 2022). https://doi.org/10.48550/arxiv.2205.01851

[41] Guoliang Li, Jiannan Wang, Yudian Zheng, and Michael J Franklin. 2016. Crowdsourced data management: A survey. *IEEE Transactions on Knowledge and Data Engineering* 28, 9 (2016), 2296–2319.

[42] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. 2013. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. 389–402.

[43] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[44] Michael Muller, Christine T Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, et al. 2021. Designing ground truth and the social life of labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[45] Donald A Norman and Tim Shallice. 1986. Attention to action. In *Consciousness and self-regulation.* Springer, 1–18.

[46] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749* (2021).

[47] Raja Parasuraman, R Parasuraman, and David Roy Davies. 1984. *Varieties of attention.* Academic Press.

[48] Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. 2020. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data* 7 (12 2020), 293. Issue 1. https://doi.org/10.1038/s41597-020-00630-y

[49] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336.

[50] Kathleen Pine, Claus Bossen, Naja Holten Møller, Milagros Miceli, Alex Jiahong Lu, Yunan Chen, Leah Horgan, Zhaoyuan Su, Gina Neff, and Melissa Mazmanian. 2022. Investigating Data Work Across Domains: New Perspectives on the Work of Creating Data. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–6. https://doi.org/10.1145/3491101.3503724

[51] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. 2010. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express* 18 (5 2010), 10762. Issue 10. https://doi.org/10.1364/OE.18.010762 Publisher: Optical Society of America.

[52] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2018. Data lifecycle challenges in production machine learning: a survey. *ACM SIGMOD Record* 47, 2 (2018), 17–28.

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

[54] Lena Ring, Stefan Höfer, Frank Heuston, David Harris, and Ciaran A O'Boyle. 2005. Response shift masks the treatment impact on patient reported outcomes (PROs): the example of individual quality of life in edentulous patients. *Health and quality of life outcomes* 3, 1 (2005), 1–8.

[55] Veronica A Rivera and David T Lee. 2021. I Want to, but First I Need to: Understanding Crowdworkers' Career Goals, Challenges, and Tensions. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–22.

[56] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2019. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering* 33, 4 (2019), 1328–1347.

[57] Mintra Ruensuk, Taewan Kim, Hwajung Hong, and Ian Oakley. 2022. Sad or Just Jealous? Using Experience Sampling to Understand and Detect Negative Affective Experiences on Instagram *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 147, 18 pages. https://doi.org/10.1145/3491102.3517561

[58] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–19.

[59] Carolyn E Schwartz and Mirjam AG Sprangers. 2010. Guidelines for improving the stringency of response shift research using the thentest. *Quality of Life Research* 19, 4 (2010), 455–464.

[60] Carolyn E Schwartz, Mirjam AG Sprangers, Amy Carey, and George Reed. 2004. Exploring response shift in longitudinal data. *Psychology & Health* 19, 1 (2004), 51–69.

[61] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 614–622.

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

[63] Jonathan Smallwood and Jonathan W. Schooler. 2006. The restless mind. *Psychological Bulletin* 132 (11 2006), 946–958. Issue 6. https://doi.org/10.1037/0033-2909.132.6.946

[64] Jonathan Smallwood and Jonathan W Schooler. 2015. The Science of Mind Wandering: Empirically Navigating the Stream of Consciousness. *Annual Review of Psychology* 66 (2015), 487–518. Issue 1. https://doi.org/10.1146/annurev-psych-010814-015331

[65] Padhraic Smyth. 1996. Bounds on the mean classification error rate of multiple experts. *Pattern Recognition Letters* 17, 12 (1996), 1253–1257.

[66] Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. 1994. Inferring ground truth from subjective labelling of venus images. *Advances in neural information processing systems* 7 (1994).

[67] Mirjam AG Sprangers, Frits SAM Van Dam, Jenny Broersen, Litanja Lodder, Lidwina Wever, Mechteld RM Visser, Paul Oosterveld, and Ellen MA Smets. 1999. Revealing response shift in longitudinal research on fatigue: the use of the thentest approach. *Acta Oncologica* 38, 6 (1999), 709–718.

[68] Mirjam AG Sprangers and Carolyn E Schwartz. 1999. Integrating response shift into health-related quality of life research: a theoretical model. *Social science & medicine* 48, 11 (1999), 1507–1515.

[69] Yoshihiko Suhara, Yinzhan Xu, and Alex'Sandy' Pentland. 2017. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *Proceedings of the 26th International Conference on World Wide Web*. 715–724.

[70] Divy Thakkar, Azra Ismail, Pratyush Kumar, Alex Hanna, Nithya Sambasivan, and Neha Kumar. 2022. When is Machine Learning Data Good?: Valuing in Public Health Datafication. *Conference on Human Factors in Computing Systems - Proceedings* (4 2022). https://doi.org/10.1145/3491102.3501868

[71] Niels Van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR)* 50, 6 (2017), 1–40.

[72] Niels Van Berkel, Jorge Goncalves, Simo Hosio, and Vassilis Kostakos. 2017. Gamification of mobile experience sampling improves data quality and quantity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–21.

[73] Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R Schinazi, and Markus Gross. 2020. Affective state prediction based on semi-supervised learning from smartphone touch data. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[74] Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. 2022. Whose AI Dream? In search of the aspiration in data annotation. *CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3491102.3502121

[75] Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. 2022. Whose AI Dream? In search of the aspiration in data annotation.. In *CHI Conference on Human Factors in Computing Systems*. 1–16.

[76] Jiannan Wang, Tim Kraska, Michael J Franklin, and Jianhua Feng. 2012. Crowder: Crowdsourcing entity resolution. *arXiv preprint arXiv:1208.1927* (2012).

[77] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. 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*. 3–14.

[78] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. 2020. Robust optimization for fairness with noisy protected groups. *Advances in Neural Information Processing Systems* 33 (2020), 5190–5203.

[79] Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. 2021. Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective. *arXiv preprint arXiv:2112.06409* (2021).

[80] Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication Reliability–An Empirical Approach to Interpreting Inter-rater Reliability. *arXiv preprint arXiv:2106.07393* (2021).

[81] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2691–2699.

[82] Zhen Yue, Eden Litt, Carrie J. Cai, Jeff Stern, Kathy K. Baxter, Zhiwei Guan, Nikhil Sharma, and Guangqiang (George) Zhang. 2014. Photographing Information Needs: The Role of Photos in Experience Sampling Method-Style Research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1545–1554. https://doi.org/10.1145/2556288.2557192

[83] Yu Zhang, Yun Wang, Haidong Zhang, Bin Zhu, Siming Chen, and Dongmei Zhang. 2022. OneLabeler: A Flexible System for Building Data Labeling Tools. In *CHI Conference on Human Factors in Computing Systems*. 1–22.

[84] Haiyi Zhu, Steven P Dow, Robert E Kraut, and Aniket Kittur. 2014. Reviewing versus doing: Learning and performance in crowd assessment. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1445–1455.