Optimizing for Happiness and Productivity: Modeling Opportune Moments for Transitions and Breaks at Work

Harmanpreet Kaur[1], **Alex C. Williams**[2], **Daniel McDuff**[3], **Mary Czerwinski**[3], **Jaime Teevan**[3], **Shamsi Iqbal**[3]

[1]University of Michigan, [2]University of Waterloo, [3]Microsoft Research

harmank@umich.edu, alex.williams@uwaterloo.ca, {damcduff,marycz,teevan,shamsi}@microsoft.com

ABSTRACT

Information workers perform jobs that demand constant multitasking, which leads to context switches, productivity loss, stress, and unhappiness. Systems that can mediate task transitions and breaks have the potential to keep people both productive and happy. We explore a crucial initial step for this goal: finding opportune moments to recommend transitions and breaks without interrupting people during focused states. Using affect, workplace activity, and task data from a three-week field study ($N = 25$), we build models to predict whether a person should continue their task, transition, or take a break. The R^2 values of our models are as high as 0.7, with only 15% error cases. We ask users to evaluate the timing of recommendations provided by a recommender that relies on these models. Our study shows that users found our break and transition recommendations to be well-timed, rating them as 86% and 77% accurate, respectively. We conclude with a discussion of the implications for intelligent systems that seek to guide task transitions and manage interruptions at work.

Author Keywords

Affect; productivity; workplace

INTRODUCTION

Information workers operate in an environment where multitasking is common [16, 37] and task priorities constantly shift [59]. In practice, multitasking often leads to context switching as people try to manage different tasks and communication channels at once [17, 24]. As a result, information workers often switch context at inopportune moments—when they have maximum context about their current task and are in a state of flow [14]—resulting in high task-resumption costs and loss of productivity [37]. Switching out of unproductive states, though, is important, since these can lead to stress and unhappiness at work [26], which also leads to loss of productivity [56]. This vicious cycle is hard to break if we consider productivity and affect in isolation: a person's affective state is crucial to their workplace effectiveness [28, 51]. Indeed, the "happy-productive worker" hypothesis [73] claims that information workers cannot be their most productive selves, or do their best work, without first being happy.

One way to keep information workers both happy and productive is to recommend state changing actions (such as, "transition to a different task" or "take a break") at times when we believe people to be in unproductive or unhappy states. Such recommendations must be well-timed, as prior work suggests that intelligent systems can do more harm than good if people are interrupted at the wrong times [5, 23, 40, 44]. It is challenging to identify the ideal moment for such recommendations without a fine-grained understanding of the person's affective state and work context. Most prior work has relied on intrusive methods, such as wearable sensors, to gain some understanding of people's workplace affect and context (e.g., [74]), but several of these sensors are challenging to wear continuously, and are subject to technological failure [9].

Our goal is to identify opportune moments for guiding people towards effective states at work in a minimally invasive way. We rely on a tool that logs workplace activity, daily task information, and affect derived from facial expressions in a privacy-preserving way; and conduct a four-week field study with 25 participants at a large technology company. Our work has two phases: (1) we use three weeks of data collected via our tool to build predictive models that jointly optimize people's workplace happiness (represented by positive affect) and productivity, and (2) we deploy these models to make real-time recommendations of transitions and breaks for our participants, and obtain their feedback on the timing of our recommendations.

Our results indicate that it is possible to jointly model positive affect and productivity with reasonable goodness-of-fit (R^2 0.2-0.7) and low error (<15%). These models rely heavily on workstation activity (e.g., mouse and keyboard activity, tab switches), task information from people's to-do lists, and emotion, but the importance of features varies by individual. When applied in practice, these models can be used to identify opportune moments for transitions and breaks in real-time to obtain 85.7% and 77% accuracy, respectively. Our participants appreciate timely reminders for taking these actions, use these times to replenish their work energy, and are more reflective about their work as a result. These findings have implications for building intelligent systems for workplace well-being.

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04…$15.00

DOI: https://doi.org/10.1145/3313831.XXXXXXX

RELATED WORK

Our work relies on principles from HCI, ubiquitous computing, organizational behavior, and psychology, as described below.

The Happy-Productive Worker

Organizational behavior studies show that people who have a happy disposition at work tend to have higher rated (i.e., more productive) performance measures [28, 73]. Coined the "happy-productive worker hypothesis", this has been studied in several organizational settings with different operationalizations of happiness (e.g., job satisfaction, lack of emotional exhaustion) and performance (e.g., meeting monthly targets, ratings from manager). The hypothesis has been supported by prior work in specific domains [13, 65, 66, 72].

Observing 42 software developers—an important class of information workers—Graziotin et al. [27] find that happy developers are indeed better at analytical problem solving and critical thinking. Similarly, [18, 48, 58] mined publicly available issue-tracking data from software repositories (e.g., Apache Jira) to find that positive emotions are correlated with shorter issue fixing time. In a more recent paper, Graziotin et al. [26] suggest that it is more cost-effective to study unhappiness and reduce it – this automatically reduces stress and improves productivity. It is evident from this work that productivity and happiness are intertwined; thus, we consider both in our study.

Multitasking and Interruption Management

Humans are prone to multitasking because they have the cognitive capacity to do so [63, 70], and technology supports this practice [8, 16, 24, 57]. However, multitasking often results in switching tasks at inopportune moments, due to both internal [2, 43] and external [35, 15] interruptions. Information workers switch windows every 40 seconds [52] and working spheres every 3 minutes [24]. Once interrupted, they can take around 15 minutes to resume their task [37]. An interruption at the wrong time, e.g., when they are in a state of flow [14], can result in lower task performance [56] and increased frustration, anxiety and annoyance [4, 23, 38, 40, 5, 49].

Past research has focused on opportune moments for task switching where the goal is to reduce disruption and improve productivity. To support the notion of the Happy-Productive worker, task switches need to take into account the user's current level of engagement as well as their affective state. McFarlane proposes four methods of interrupting a user for switching tasks: immediate, negotiated, mediated, and scheduled [55]. Our work relies on a mediated strategy, where the system uses contextual information to decide when to recommend a task switch or a break to the user, thereby reducing the burden on the user to pick an optimal moment.

Sensor-based Affect and Productivity Monitoring

Prior work has employed sensor-based monitoring to identify opportune moments for a task switch: [41, 45] used pupillary response to measure cognitive load; [32, 33] studied heart rate variability (HRV) as a proxy for focus; and several other studies, including [10, 29, 30, 62, 60, 68] used electromyogram, accelerometry data, electrocardiogram data, skin conductance, sleep and circadian rhythms, mobile phone context, and other signals to measure stress and cognitive load.

A comprehensive overview of sensor-based psychological, physiological, behavioral, and contextual measurements of stress can be found in [3]. Most related to our work is Züger et al.'s prediction of interruptible moments in people's workdays based on a combination of sensor-based data outlined above [74]. They collect ground truth self-reports of interruptibility from people, and train personalized models that use data from several sensors to predict if an individual is interruptible at a given time. We build models to predict joint optimization of productivity and happiness as a complementary aspect to [74]'s interruptibility prediction, and build on their work by using data from an emotion, context, and task logging tools, but without any wearable sensors. Our work builds on the wealth of sensor-based monitoring studies, specifically, those that demonstrate that even simple sensors are valuable for modeling interruptability [21].

Tool-Based Productivity Mediation

Extending sensor-based monitoring into real-world applications, researchers have leveraged these data sources to develop systems that help people better manage their attention spans, to dos, and overall productivity. For example, Busybody [34] applies Bayesian models built using log data and user labels to predict the cost of interruption, and Lilsys [7] predicts availability based on user actions and ambient sensors. Oasis [39] utilizes the perceptual structure of tasks via statistical models that detect breakpoints in real time, and schedules notification delivery accordingly, in order to reduce interruption costs. Several tools now support easy task switching: Groupbar [64] allows windows belonging to the same task to be grouped together for task switching with a single mouse click, and Active Progress bar [36] allows users to switch to temporary tasks during the wait time while computer tasks are in progress. At the day-level, Switchbot [71] helps users disengage from work and reengage the next day, which has a positive impact on productivity and well-being.

Most related to our overarching goal, prior studies of work-related breaks [20] have led to several break-recommendation tools in the CHI literature. Cambo et al. [9] introduced BreakSense, a multi-device application that employs location-based challenges to promote mobility in the workplace. Similarly, Luo et al. [47] designed "Time for Break", a break-prompting system aimed at combating prolonged sedentary behavior, and found that pre-existing habits play an important role in the receptivity of the system. Most recently, Tseng et al. [67] developed and studied UpTime, a conversational system built into Slack that seeks to improve the transition between breaks and work time by blocking distractions (e.g., social media sites) for a fixed period of time.

While the goal of these systems is to reduce interruption and maximize productivity, leveraging user affect in supporting productive practices remains relatively unexplored. As a first step towards building such tools, in this work, we leverage emotion, context, and task information to predict opportune moments for task transitions and breaks for people, with the goal of helping people become happy-productive workers.

RESEARCH GOALS

Our broad research goal is to help people achieve their work-related goals while also optimizing positive affect in the workplace. To approximate this, we use predicted emotion labels for people's facial expressions, their workplace activity, and their daily task list, to recommend actions for productivity and positive affect at any given time – specifically, switching to a different task or taking a break. Our research questions are:

RQ1. Can we predict productivity and affect from emotion, activity, and task data, without using wearable sensors?

RQ2. Can we identify opportune times for transitioning tasks and taking breaks for people during their workday?

RQ3. How do people respond to recommendations of transitions and breaks at opportune times based on optimizations of productivity and positive affect?

We study this in two phases: in Phase 1, we develop models to predict opportune moments for intervention using a jointly optimized value for positive affect and productivity; in Phase 2 we evaluate these moments via real-time recommendations.

PHASE 1: MODEL DEVELOPMENT

To guide people towards positive affect and productivity at work, we perform optimization over data collected about people's emotions, workstation activity, and tasks. Here, we describe how we collect this data, followed by the specifics of our features, the models used for prediction, and finally the metrics we use for evaluating our predictions.

Tracking Task, Emotion, and Context

We collect 8 categories of data to make predictions about productivity and happiness at work: (1) emotion data, (2) heart rate, (3) physical movement, (4) interaction data, (5) task information, (6) time and day information, (7) digital actions being performed, and (8) productivity and affect reports. We use existing emotion and context logging software to collect categories 1-4 and 6-7 [54], and build an interface, FLOWZONE (Figure 1), on top of this software to collect 5 and 8. More details on the specific features obtained is provided after the tool description below.

Emotion and Context Logging Software

We obtain emotion expressions and context by processing data collected via a standard webcam (participant privacy is preserved by never storing raw data). The software [54] analyzes people's facial expressions while at their desk. It consists of a visual and a context pipeline.

Visual Pipeline. The tool processes video data from a webcam. First, it detects faces in the video and extracts landmark positions of key facial features. The distance of the user's face from the camera is extracted using the inter-ocular distance calculated from the facial landmarks. Next, the facial regions of interest are analyzed using an emotion detection algorithm, returning eight probabilities for each of the following basic emotional expressions—anger, disgust, fear, joy, sadness, surprise, contempt, and neutral [19]—with an accuracy of $\sim$87%. We use Microsoft's publicly-available EmotionAPI to detect emotion expression (for more information on its classification of facial expressions, see [6]). Using image frames, the software also extracts heart rate via the photoplethysmographic signal [61, 53].

Context Pipeline. The software [54] logs information about the open applications and interactions with computer peripherals. Each time applications are opened, closed, in focus (the front application), minimized, or maximized, it records these activities with the corresponding timestamp. The software only logs the title of the window, indicating the page or application the user was on, and these values are hashed before storing. It also logs mouse movements and clicks and keyboard inputs.

FLOWZONE: An Interface to Collect Self Reports on Productivity and Affect

We developed FLOWZONE , a user interface on top of the aforementioned Emotion and Context Logging Software [54] to collect additional information on people's daily tasks, and self-reports of task progress, productivity, and affect. FLOWZONE comprises of two components: the Task Tracker, and the Productivity and Affect Self-report interface. The data collected through FLOWZONE is temporally aligned with the data collected by the Emotion and Context Logging Software.

Task Tracker. The Task Tracker is a simple to-do list interface which asks people about the type of activities involved in doing a task (e.g., reading, writing, coding, etc; a list of eight categories borrowed from [22]), its urgency and difficulty, and an estimate for the anticipated completion time for it (Figure 1). Prior work shows that the emotion and context-based markers can change based on the task being performed [22], making this task information critical.

Productivity and Affect Self-Reports. Based on [22], we know that people's reported affect data along with task information can provide a deeper understanding of their observed facial expressions. Our interface collects self-reports of affect and task progress. People report affect via 6 variables derived from the Positive and Negative Affect Scale (PANAS) [69]. Of the 6, 3 are positive items (inspired, enthusiastic, determined) and the other 3 are negative items (irritable, nervous, upset) from the original 20 on the scale. The values for these are selected via sliders ranging from 0-10. We use a smaller subset of items here to minimize time spent filling out the report (reducing interruption costs) – a practice that has been seen in prior work with similar goals of reducing self-report costs [50, 71]. People also report how productive and busy they feel (range: 0-10), and their progress per task (range: 0-100).

Data Collection for Model Building

We recruited 30 participants from a large technology company, and asked them to install our data collection tool on their desktop computers for four consecutive weeks. We used data from the first 3 weeks of the study to build models, and the last week for validation (see Phase 2). 5 participants provided incomplete data due to incorrect setup, insufficient self-reports, or taking time off. Our data set thus comprised of data from 25 people (F=6, M=19) with job roles: Software Engineer (8), Senior Software Engineer (5), Designer (3), Data Scientist (2),

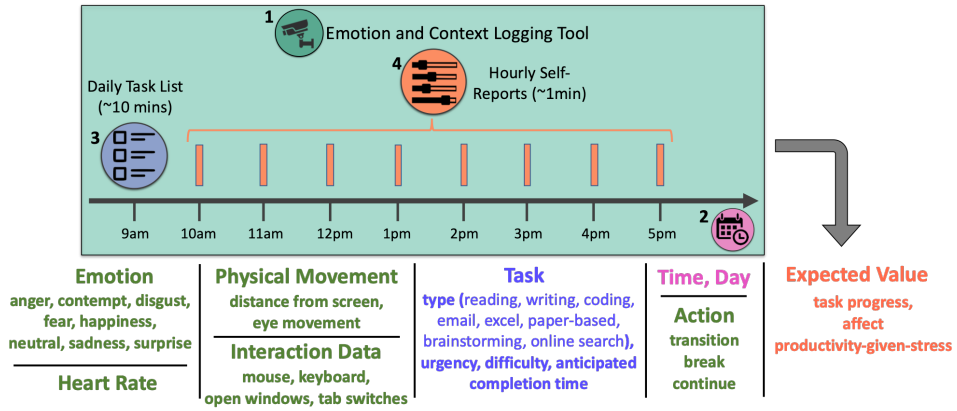

Figure 1. Task, Emotion and Context Tracking Setup. The different components are: (1) a logging software that tracks rich emotion and context data via a webcam [54]; (2) time and day tracking; (3) a daily task list interface where people enter information about the task type, urgency, difficulty, and anticipated completion time; and (4) hourly self-reports of task progress, overall affect, and feeling of productivity-given-stress

Finance Manager (2), Senior Program Manager (1), Senior Content Developer (1), Principal Development Manager (1), Applied ML Engineer (1), and Project Manager (1). Participants were compensated with $150 post-study. Participants were asked to engage in their regular activities on their computers. The only change to their routine was filling out the Task Tracker list at the start of their day, and the hourly Productivity and Affect self-reports (Figure 1). The emotion and context logging software collected data (with informed consent) in the background as participants used their computers.

Features Categories

We collected the following data using the emotion and context logger and FLOWZONE. We generated a list of 35 features in eight categories from the collected data (full list in Figure 1).

1-Emotion. Classified into eight emotion categories—anger, contempt, disgust, fear, happiness, neutral, sadness, surprise—by the logging software, with probabilities that represent the magnitude of each emotion at a given time, adding up to 1.

2-Heart rate (HR). Prior work shows that a low heart rate and low heart rate variability is reflective of focus [32, 33]. Heart rate variability cannot yet be calculated without using wearable sensors; we use heart rate captured by the logging software to observe if the magnitude accounts for any importance.

3-Physical Movement. This includes eye movement and distance from screen, captured via the logging software.

4-Interaction Data. Also captured by the logging software, this includes mouse and keyboard activity, number of tab switches, and number of open windows.

5-Task Information. Includes eight features: task *urgency*; task *difficulty*; anticipated task *completion time*; and task *type* as binary values for reading, coding, content creation, digital communication, brainstorming, paper-based reading or writing, creating spreadsheets, and searching for information online. Each feature vector includes this information for tasks that show progress between self-reports at different time intervals. If multiple tasks show progress, task type information is a union of the values, difficulty and anticipated time are added, and urgency is an argmax over the individual values.

6-Time, Day. We encode time (hours spent at work) and day of week (categorical variable using 7 binary features, one for each day) as two feature proxies for circadian rhythms.

7-Potential Actions. At any given time, a person can take one of three actions to change their work environment: (1) take a *break*, (2) *transition* to a different task, or (3) *continue* their current task (i.e., take no action). Breaks may be digital (e.g., visit social media) or physical (e.g., walk away from their computer). Without wearable sensors, we do not have data for what people do on their physical breaks, so we encode physical breaks given absence of data. Any time sample is considered a digital break if people visit one of the following websites during that time: Facebook, Twitter, LinkedIn, Instagram, Reddit, YouTube, Twitch. It is considered a transition if the foreground windows and tabs being used changes completely in that timeframe. All other time samples fall under the "continue" task category.

8-Output Variable: Joint Productivity-Happiness Expected Value. We define the output variable of this model as the Joint Productivity-Happiness Expected Value (Expected Value going forward), which considers both productivity and positive affect, the two variables we intend to jointly optimize. Expected Value, in our case, is computed from the hourly self-report data from FLOWZONE, thus including both self-reported task progress, and self-reported affect using the modified PANAS sliders, normalized and scaled to be in the range of 0-100. However, productivity and affect also have an interaction effect [51]. For example, people can be happy doing rote work, which may or may not be productive, or people can be focused but also stressed because of a challenging task. We thus add a combined feature using the "productivity" and "busy" self-reports. We scale each value to be between -5 to 5, and multiply them to get an overall *productivity-given-stress* value [51]. We normalize and scale this value to range from 0-100 as well. For our study, Expected Value is calculated with equal weights for all three components, but other weightings could be used based on which values users want to prioritize (we comment on this in our Discussion). Thus,

$$ExpectedValue = (\tfrac{1}{3} \times TaskProgress) + (\tfrac{1}{3} \times Affect) + (\tfrac{1}{3} \times Productivity-given-stress)$$

Datasets

Original Dataset. Our original dataset is comprised of data collected over three weeks ($N = 25$). Emotion and context were logged at a microsecond-level granularity whenever people were at their workstation; self-reported data was collected at hourly intervals. Since we did not force people to answer self-reports to minimize disruptions, there were some hours with missing datapoints. On average, there were 7 hourly self-reports collected per participant, per day (min=4, max=15). Expected Value is dependent on self-reports; thus our original *complete* dataset comprised of time instances when this self-report value was available (hourly at best). This gave us 62.44 datapoints per participant, on average (min=33, max=122).

Original + Simulated Dataset. Our two data sources—log data and self-reported data—operate at different time intervals (microsecond and hourly, respectively). To better align these and get a complete picture of a user's day, we up-sampled the self-report hourly data using growth and decay functions, thus getting Expected Value at more granular time intervals. Given a value at hour h1 and another at h2 (where h1 $<$ h2), we applied a growth function to value at h1 and a decay function to that at h2, and took the max value for every time interval t between h1-h2. We experimented with several growth and decay functions: exponent with $\gamma = \frac{1}{25}, \frac{1}{5}$, 1, 5, 25; natural log (*Ln*); and Log_{10}. We also tested different time samples for up-sampling: $1, 2, 3, 4, 5, 7, 8, 10, 15, 20, 25, 30$ *mins*. The microsecond level data obtained from the logging software was similarly down-sampled to the same time samples by applying aggregation functions. This dataset spanned every 1-30 mins per day, with aggregated down-sampled emotion and context logging data, and up-sampled self-report data.

Models

We built models that took as input all of our data sources and predicted an Expected Value. Our data effectively represents a timeseries per person, and our output variable's continuous nature called for regression models. We thus modeled our setup as a classic timeseries forecasting problem using Auto-Regressive Integrated Moving Average (ARIMA) models. ARIMA models have 3 main components: (1) *the Auto-Regressive part*, the number of prior (*lagged*) values of the dependent variable to be used for each new training and prediction datapoint; (2) *the Integrated part*, the degree of differencing required to convert a non-stationary variable into a stationary time series; and (3) *the Moving Average part*, the number of random errors of the past to be used to account for current datapoint's errors. ARIMA models traditionally use only one timeseries: the main variable being regressed (here, Expected Value). We build ARIMAX models—ARIMA models with eXogenous variables—to account for input features (e.g., emotion labels, task information) which are potential explanatory variables – exogenous in ARIMA terminology.

While ARIMAX models are the best representation of our timeseries data, they are complex and thus expensive to compute. With our Phase 2 goal being a deployment, we also sought to model our data using other regression models. We tried several different ones (e.g., Support Vector Regression and Multiple Linear Regression), and finally picked Random Forest Regression (RFR) models for our real-time recommendation task because these had the best performance (metrics for gauging performance explained below) and assigned feature importances similar to ARIMAX. We relied on this similarity of feature importances between the two types of models as a form of validation for using the less computationally demanding RFR models in the deployment phase.

Cross-validation. When using the original dataset, we applied leave-one-out cross-validation (LOOCV), training on n-1 datapoints and testing on 1, averaging results of all possible model combinations done this way. For the original + simulated dataset, we use holdout cross-validation, using 60% of data for training, and 20% each for validation and testing.

For both these methods of cross-validation, we follow day forward chaining to ensure that future values are never used to predict past values. That is, for each day, we treat each future datapoint as a new test case, and use all prior ones as our training set. Similarly for the train-validation-test dataset split, we use ordered splitting such that no future data points are in the training or validation sets.

Metrics. For the ARIMAX models, we use Akaike Information Criterion (AIC) values to find the best-fitting model. AIC values are better suited for timeseries model results because they represent goodness of fit for past and future of the time-series data; lower AIC values indicate better fit. We use R^2 and Adjusted-R^2 to evaluate the goodness-of-fit of our RFR models. These metrics are used to report how well the selected independent features explain the variability of our dependent variable (Expected Value). For example, an R^2 value of 0.X is read as "the model explains X% of variance in the data." R^2 values can be biased to the addition of new features, even when the features do not add any explanatory power. Adj-R^2 handles this bias, and thus is a better measure for model comparison. We report both for our RFR models, but pick the best models using the Adj-R^2 values. We also compute Root Mean Square Error (RMSE) values for both types of models to indicate the difference between actual and predicted Expected Values.

PHASE 1: MODEL EVALUATION

We use emotion, interaction, task, and action data to model our output variable – the Expected Value of people's workday (which is designed to jointly capture their productivity and affect). We test several regression models on our original and original + simulated datasets. We ultimately rely on the original + simulated dataset for all our model-building after validating this dataset against the original dataset: there is no significant difference ($p > 0.1$) in model performance or feature importances between the two datasets.

ARIMAX Model Performance

ARIMAX models are commonly applied to timeseries data like ours. Since each timeseries is unique to the context it was collected in, we treat all participants' data separately, and build personalized ARIMAX models for all of them. The core AR, I, MA features of an ARIMAX model rely on this unique context per timeseries (see AR, I, MA values in Table 1). ARIMAX models output results in the form of estimates for each independent variable along with p-values for significance.

	AR,I,MA	Feature (Estimate)	AIC
P1	(2,0,4)	Time (1.3)***, Tab Switches(-0.7)***, Continue(-0.4)**	2650
P2	(5,1,2)	Complete Time (-0.7)***, Mon (-0.4)**, Transition (0.6)**	2218
P3	(1,0,4)	Screen Distance (1.5)**, Surprise (-0.3)**, Continue(-0.1)*	4902
P4	(3,1,5)	Keyboard (1.07)***, Coding (0.26)**, Continue (0.72)***	2291
P5	(2,1,5)	Keyboard (0.89)***, Surprise (-1.3)***, Break (1.22)**	1880
P6	(5,2,3)	Screen Distance (1.1)**, Continue (0.35)**	3105
P7	(3,1,4)	Tab Switches (-1.33)**, HR (-0.13)*, Continue (-1.07)**	1284
P8	(4,0,4)	HR (-1.78)**, Transition (-1.2)**	4171
P9	(1,1,5)	Mouse (0.94)***, Anger (-0.42)**, Break (1.05)**	1255
P10	(2,0,3)	Happiness (0.8)***, Urgency (-1.23)***, Continue (-0.28)*	2260
P11	(5,0,4)	Mouse (0.63)**, Continue (0.35)**	1315
P12	(5,1,4)	Eye Movement (0.23)*, Wed (1.53)**, Surprise (0.18)*	3147
P13	(1,0,5)	Sadness (0.77)**, Surprise (0.14)*, Transition (-0.9)*	3357
P14	(3,2,4)	Time (-1.6)***, Mon (-0.66)**, Continue (-0.83)**	2469
P15	(1,1,5)	Keyboard (-0.33)*, Difficulty (-0.92)**	2195
P16	(2,1,5)	Screen Distance (0.72)**, Break (0.59)**	1109
P17	(1,0,5)	Brainstorming (0.31)*, Continue (0.41)**	3692
P18	(4,0,3)	Tab Switches (-0.39)***, Screen Distance (-0.62)**	2774
P19	(2,1,4)	Sadness (0.25)**, Reading (-0.51)*, Break (1.51)**	2053
P20	(2,1,3)	Coding (0.79)***, Continue (1.27)**	1529
P21	(1,0,5)	Urgency (-1.25)**, Monday (-0.68)**, Break (1.04)***	1893
P22	(2,1,5)	Keyboard (0.93)**, Difficulty (0.17)**, Continue (0.65)*	3041
P23	(3,2,4)	Writing (-1.64)***, Break (0.49)**, Continue (0.11)*	2063
P24	(2,1,5)	Tab Switches (0.16)*, Reading (0.79)**, Transition (1.14)**	1148
P25	(4,0,4)	Complete Time (-1.29)***, Tues (-0.09)*, Anger (-0.17)*	1547

Table 1. Results from the ARIMAX models per participant: AR, I, MA denote values for the auto-regressive, integrated, and moving average components of the model; Features presented are those with significant estimate values; and AIC value represents goodness-of-fit. Significant levels: *=p<0.05 **=p<0.01 ***=p<0.001

We find that ARIMAX models output 2-3 significant features per participant. To better understand the broader categories of features that are important, we bin our 35 features into 8 categories representing different data sources: Emotion, Heart Rate (HR), Physical Movement, Interaction Data, Task Information, Time of Day, Day of Week, Action. Observing the categories with at least one significant feature per participant, Action is the most popular category (21 out of 25 participants show at least one of break, transition, or continue as having a significant estimate), followed by Task Information (12 out of 25), Interaction Data (10 out of 25), Emotion (8 out of 25), Physical Movement (5 out of 25), Day of Week (5 out of 25), Time of Day (2 out of 25), and Heart Rate (2 out of 25).

ARIMAX models consistently return significant estimates for a feature in the *Action* class: whether someone has recently taken a break, transitioned tasks, or has been continuing the same task is important for predicting future actions. Time-series models are well-known for capturing such historical nuance. We find that samples aggregated at 7- and 10-minutes (time sample variable t used in Original + Simulated dataset) provide the best results for these models, with average AR and MA values being 3 and 4, respectively. This means that the ARIMAX models consider the past 21-30 minutes (3x7 and 3x10) of data in forecasting the Expected Value for a given time interval, and do this with an average RMSE of 8.6% and AIC value of 2374. The validation split highlights $Exp(\frac{-1}{25}x)$ as the time decay function for the best performing model.

Random Forest Regression Model Performance

ARIMAX models are complex and computationally demanding (processing time of $\sim$15 mins per participant), making it hard to use them in real-time settings. We thus test regression models, settling on Random Forest Regression (RFR) models because they have the best performance. Since ARIMAX models are more naturally suited to our timeseries data setting, we rely on the results of the ARIMAX models to validate the performance of our RFR models.

We build RFR predictive models at three levels: general, per participant, and per cluster, where clustering is done based on job role. A general model with good performance has the potential of being applied at a larger scale, because it would indicate people's data can be used interchangeably. Personalized models per participant with good performance can help us understand which features matter most when modeling different individuals. Models for different job role clusters can highlight whether people's work practices, productivity, and affect are defined by something specific about their job role.

Table 2 presents results for all models using the metrics explained above. It also includes the distribution of data (mean and S.D.) for each participant and cluster, to better contextualize our R^2 and Adj-R^2 results. Further, Table 2 highlights the best values of the constants used for modeling via the holdout validation set. All models with the best validation set performance use $Exp(\frac{-1}{25}x)$ as their time decay function; the best values for Time Sample per model are indicated in Table 2. Below, we share results from each of these models, and then compare the feature importances seen across them.

General Model Performance

Given prior work that suggests that people have unique patterns of activity, affect, and daily to-dos at work, it comes as no surprise that our general model that includes all participants as one data source has mediocre performance. With an R^2 and Adj-R^2 value of 0.2, the general model is able to explain 20% variance in data, making it a moderate fit. The Root Mean Squared Error (RMSE) for this model is 26.5, on a scale of $0-100$; RMSE values share the same scale as the output variable, Expected Value (Table 2, header "All").

Personalized Model Performance

Our personalized models have high R^2 and Adjusted-R^2 (Adj-R^2): R^2 values range from $0.2-0.7$, with an average of 0.52, and Adj-R^2 values range from $0.2-0.7$, with an average of 0.47 (Table 2, header "Participants"). High values for both these metrics indicate that our models are a good fit for people's data, and a large percentage (up to 70% in the best case) of the variance in data is explained by the models. The RMSE values range from 3.5 to 13.2, the average value being 7.1. Overall, these models perform extremely well both in terms of goodness-of-fit and low error values, considering the wide distribution of data per participant.

Cluster Model Performance

Our cluster models have similar performance to the personalized models, with R^2 values ranging from $0.4-0.7$, and RMSE values between $3.9-5.2$. In fact, in some cases, these models perform better than the personalized models for the participants in the cluster. Since the clusters are formed based on job role, this suggests that people doing similar jobs have similar task progress, affect, and productivity-given-stress rates. In a cold-start setting—when we don't have enough data from a participant to build personalized models for them

	Participants																									Clusters					All
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	1	2	3	4	5	
Mean	26.2	39.7	35.9	21	29.4	19.9	33	31.8	33.8	33.8	27.5	26	29.1	29	30.5	30.7	35.4	23.8	25.7	22.6	23.3	32.9	33.6	35.6	17.6	29.6	30	30	29.6	24.2	29.1
S.D.	23.6	22.5	31.4	19.9	25.5	12.3	27.8	25.5	22.2	31.4	24.8	25.5	18.2	22.5	20	19.5	21.9	20.9	19.3	16.5	19	27.5	25.8	24.3	13.8	23.5	24.4	20.7	21.4	17.9	22.4
RMSE	4.7	4.9	5.4	3.6	5.6	10.3	7.6	5.6	5.2	4.6	7.6	9.7	9	13.2	6.1	6.5	3.5	9.6	7	10	11.7	5.4	8.3	6.4	5.6	5.2	5.1	5.1	4.7	3.9	26.5
R^2	0.4	0.7	0.4	0.5	0.2	0.6	0.6	0.6	0.7	0.4	0.3	0.3	0.6	0.5	0.6	0.6	0.7	0.2	0.3	0.6	0.7	0.6	0.6	0.5	0.7	0.7	0.6	0.7	0.4	0.5	0.2
Adj R^2	0.4	0.7	0.3	0.4	0.2	0.6	0.5	0.6	0.6	0.3	0.3	0.2	0.5	0.4	0.6	0.4	0.7	0.2	0.3	0.5	0.7	0.6	0.6	0.5	0.6	0.7	0.5	0.7	0.3	0.5	0.2
Time Sample	7Min	2Min	4Min	7Min	1Min	5Min	4Min	8Min	3Min	7Min	4Min	1Min	8Min	1Min	3Min	1Min	3Min	1Min	1Min	8Min	5Min	2Min	1Min	7Min	4Min	2Min	2Min	1Min	4Min	2Min	2Min

Table 2. Results of Random Forest Regression Models for all Participants and Clusters, and a Generalized Model for "All" Participants. Participants are color-coordinated according to their cluster membership. E.g., P1-8 belong to cluster C1.

immediately—modeling based on data from their job role cluster would be a viable alternative. The clusters we chose here were based on the official job roles of our participants – Software Engineer, Senior Software Engineer, Designer, Data Scientist, Finance Manager, Other (which included Senior Program Manager, Senior Content Developer, Principal Development Manager, Applied ML Engineer, and Project Manager).

Understanding Feature Importance

Similar to ARIMAX results, we bin our 35 features into 8 categories to highlight the importance of each class. The feature importances of all categories sum up to 1; the maximum importance value assigned to any individual category is 0.60. Figure 2 shows the importance for all categories.

We find interaction data to be the most important feature category on average, followed by task information, emotion, physical movement, time, heart rate, day, and potential actions. The average feature importances for different categories across all participants were: interaction data=0.22, task information=0.19, emotion=0.17, physical movement=0.14, time=0.13, heart rate=0.08, day=0.07, and potential action=0.02. Even though interaction data is the most important feature category on average, it is not always the most important feature for each participant. For example, emotion is the most important category for P13, task information for P2 and P20, and combinations of other categories are equally important for other participants. The order of feature importance remains the same if we look at the frequency at which each feature is most important.

Interaction data, while an important feature category per participant, is not the most important feature for any of the clusters or the general model (Figure 2, Clusters start with "C" and general model under "All"). On average, the feature importances per cluster are not aligned with those of the individual participants, especially if a participant has high feature importance for a particular feature category. The general model's feature importance values are more spread out across all categories of features, as expected from an aggregated model.

Overall, we find that there are differences in the features that are important per person, when compared to those important for a job role cluster or for the general model with all participant data. This is interesting given that the R^2, Adj-R^2, and RMSE values are not too different across these, especially when comparing personalized and cluster models. Indeed, it seems that an important consideration when applying these models in a real-world setting is the eventual need for personalized models. While starting with cluster-based models might rid one of the cold-start problem, no general or cluster model represents the participant and what is important for their Expected Value in the same way as their own data.

Comparing ARIMAX and RFR Models

We find that at least one of the features with significant estimates in the ARIMAX models also consistently belongs to the same feature category as the RFR results. For example, task information is the most important feature class for P2, and anticipated completion time (a feature that falls under the task information category) has a significant estimate from the ARIMAX model for P2. The primary difference between the two models is in the *Action* class: ARIMAX models consistently return significant estimates for a feature in the Action class whereas RFR models do not. We hypothesize that this is due to the nature of the action variable: whether someone has recently taken a break, transitioned tasks, or has been continuing the same task becomes a more important consideration over time. Timeseries models capture exactly this nuance, whereas RFR models do not consider these prior values. Once we built these models, our next goal was to validate them via a real-time recommendation setup. We use RFR models for our deployment phase; while we tested ARIMAX models in this setting, the high processing time for ARIMAX ($\sim$15 mins per participant) made it infeasible to use them in a real-time context. More nuanced engineering efforts could reduce processing times to make ARIMAX models also work in real-time settings – we leave these explorations to future work.

PHASE 2: MODEL DEPLOYMENT

In Phase 2, we design and build a system that uses the models built in Phase 1 to recommend breaks and transitions in real time. We deploy this system to understand how people perceive the timing of our recommendations (whether we were able to find these opportune moments for breaks and transitions that we set out to), and observe people's reactions to these recommendations. We employ descriptive methods to evaluate our recommender system, and highlight themes for what people liked and disliked in our setup.

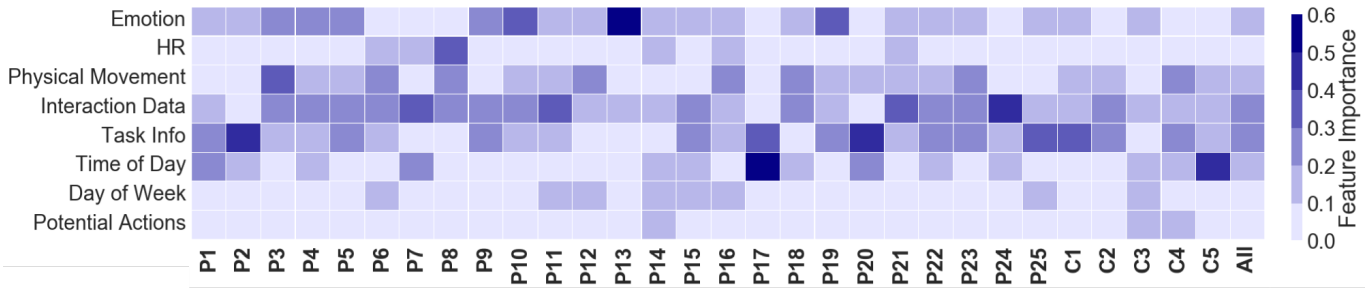

Figure 2. Feature importance output from the random forest regression models for all participants, clusters, and the general model.

FLOWZONE v2: Real-time Recommendations

We add frontend and backend components to FLOWZONE to recommend transitions and breaks in real time.

Frontend Modifications. We introduce two Windows forms that appear for transition and break recommendations; nothing appears for the "continue" recommendation. Each form shows the recommendation along with an explanation (e.g., for break recommendation, it says "Wow, you've been working hard! FlowZone thinks a break right now will replenish your energy and keep you going!"). We do not provide any personalized explanation for the recommendation. The forms ask participants to select one of these options about the recommendation provided: (1) "Yes, going to take a break", (2) "Yes, it's time for a break, but I can't take one right away", (3) "Yes I just took a break", and (4) "No, this is not a good time for a break." This granularity in "Yes" options supports our goal of understanding whether the recommendation is an interruption, or comes at an opportune time.

Model-based Backend. Our backend enables real-time queries to both the logging databases and the models built using the initial three weeks of data in Phase 1. We hosted a webserver that interacts with these components using API calls; the logging databases were hosted on Azure Table Service, and the model files were hosted on our webserver, after being converted to a compressed format. Our backend pipeline was: (1) logging software stores data every microsecond (as before); (2) for each participant's chosen time sample t (i.e., the Time Sample parameter, in minutes, that had the best model performance in Phase 1) FLOWZONE pings the server to look at the last t minutes of data; (3) the backend computes the feature vector by aggregating t minutes of data, and computes an argmax over Expected Value (our model's output variable) based on three potential action values (transition, break, or continue); (4) the potential action with the maximum Expected Value is returned as a recommendation to the frontend, where it is shown to the participant with the corresponding form.

Study Design

We deployed our updated FLOWZONE app for three days during the fourth week of our study. To ensure that our participants' responses about the recommendations were not biased by system novelty, we added a control condition which used the same system setup and outputs, but relied on pseudo-random, heuristics-based rules for recommending breaks and transitions. Our goal was not to compare the two conditions, rather validate that people were rating the timing, and not rating favorably because of the novelty of the system.

For the model condition, participants received recommendations for transitions or breaks using the predictive models built in Phase 1. For the control condition, we did not use models; we assigned heuristics-based probabilities to the three potential actions: transition and break were assigned $\frac{1}{6}$th probability each, and continue was assigned $\frac{2}{3}$rd probability because continuing a task is more common than transitioning or taking a break. At every 30-minute interval, the control condition picked one out of the three options based on the probabilities assigned, and recommended that to the participant. We set recommendation checks at 30-minute intervals for the control condition because this is traditionally the smallest time interval on people's work calendars. Both model and control condition participants were shown the same interface and explanations.

Post-Study Survey. All participants took a post-study survey that asked about their experience with FLOWZONE, and their opinions on guided recommendations. The survey included open-text questions about people's opinion of FLOWZONE; on whether the transition and break recommendations were well-timed or not, appropriately frequent or not, examples of cases of good and bad recommendations (and why), if they felt better after following a recommendation than the state they were in before; and 2 Likert questions on whether FLOWZONE made them feel more productive and happy at work (range: strongly disagree - strongly agree, 1-5).

The survey also included questions about the idea of intelligent systems guiding people at work to jointly optimize their productivity and happiness. We asked an open-text question on what they thought would be good or bad about this idea, and 4 Likert questions on whether they thought this tool would (1) be useful for their work practices, (2) make them feel positive at / about work, (3) make them feel negative at / about work, and (4) be helpful for their productivity at work (all ranged: strongly disagree - strongly agree, 1-5).

Dataset

We set up the study with 15 participants in the *model* condition and 10 in the *control* condition. The 10 participants in our control condition were classified as having relatively low data volume and quality in Phase 1 – there were some gaps in their data from training weeks in Phase 1 due to frequent meetings away from their desk (the tool was recording data only at their desk), remote work days (again, away from their desk), or unexpected vacation time. This led to Adj-R^2 values for their models ranging from $0.2 - 0.4$ (moderate to low variance in data explained by the model). Given our goal of understanding perceptions around opportunely-timed recommendations,

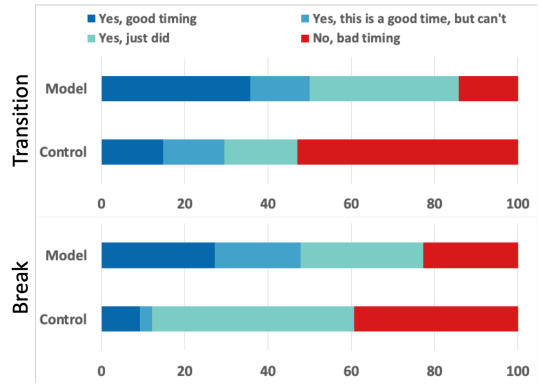

Figure 3. The number of times transition and break recommendations fell under each of the categories provided to participants.

we designed our model condition to represent the best case modeling we could do, and our control condition to counteract any novelty effects from the FLOWZONE that would hurt our understanding of the value of these opportune recommendations. Participants were not aware of any system differences, and their data was only included in the Phase 2 results if they stayed active at their workstations, per our request.

Eight out of 15 people in our model condition, and six in our control condition continued to use FLOWZONE with the updated real-time recommendations. Given that this was the fourth week of a field study, we were not surprised to have some dropouts in both our conditions. Even with dropouts in the system field study, we received some additional responses for the post-study survey (19 responses: 11 model, 8 control).

Deployment Results

Participants in the model condition received on average 5.6 recommendations for breaks and 1.2 for transitions every day (s.d. break=1.6, transition=0.56), and those in the control condition received on average 4 break and 3.5 transition recommendations every day (s.d. break=2.7, transition=0.72). In this way, participants in the model condition received relatively more recommendations for breaks over transitions on average, whereas the number was evenly in the control condition. This was not unexpected given the probability setup for the control condition which assigned equal probabilities to transitions and breaks. The large difference in the number of transitions recommended by the model vs. control conditions shows that our models have a nuanced understanding of the number of tasks people do throughout their day, and when they need to be directed to transition between them.

Model condition recommendations had a high chance of being at opportune times. We calculated accuracy via summation of responses for all "Yes" categories divided by the total number of responses per condition. Using this descriptive metric, we find model-based recommendations of transitions and breaks to be 85.7% and 77% accurate respectively (Figure 3). We verified these numbers against the control condition to check for novelty bias in favor of a transition and break recommender system: control condition transitions and breaks were only 47% and 60.6% accurate, respectively, indicating that people's evaluation in favor of the model condition was due to different, more opportune-timed recommendations.

When asked about whether FLOWZONE made them more productive at work, model condition participants leaned positive (Agree=6, Neutral=2, Disagree=4), whereas control condition participants had mostly neutral responses (A=2, N=4, D=1). When asked the same question in the context of happiness at work, model condition participants were positive (Strongly Agree=1, A=6, N=2, D=3), and those in the control condition remained neutral or negative (N=5, D=2).

When asked to rate if this future intelligent system would be useful at work, most people responded positively (SA=1, A=12, N=4, D=2). They similarly had a positive response for whether this tool could help them feel positive about work (SA=2, A=10, N=6, D=1) and help their productivity (A=12, N=4, D=3). Most people were appreciative of the idea behind task tracking, productivity, and positive affect at work, and were excited about tools like this becoming commonplace in the future: "it can be a digital assistant looking after you and your well-being, what more could you want?!" (P13).

DISCUSSION AND FUTURE WORK

We have shown that it is possible to build models (R^2 0.2-0.7; RMSE $<$15%) that jointly optimize productivity and happiness at work, using emotion, workplace context, and task-based data. When deployed, these models allowed us to recommend transitions and breaks to people at opportune times (people evaluated the accuracy of timing as: transitions=85.7%, breaks=77%). Below, we discuss several design implications and considerations that resulted from our studies.

Frequency and Timing of Recommendations is Crucial

We began our exploration of recommending transitions and breaks with the intuition that finding opportune times for providing these recommendations was important to avoid disrupting people's focused work times. Our results highlight this from people as well, making it an important design consideration for systems in this space. Participants in our model condition felt that the frequency of recommendations for transitions and breaks was "just right" (P3) or "frequent, but good for my health" (P10). With higher frequency recommendations (as in the control condition) participants felt that the frequency "very rarely matched my own assessment" (P11). The timing of recommendations was also a key consideration for people's decisions to follow through on what was recommended. Well-timed recommendations lead to more follow-through:

> "I found the timing to be surprisingly good, actually. Following the recommendations did help me feel happier and more productive because I took more breaks that I realized after the fact that I needed. Hours turn to blurs without something to break them up so taking breaks helped the day seem more full." (P4)

Intervention Design Needs Personalization

FLOWZONE was meant to study and validate the *timing* of the recommendations, but looking ahead, this is simply a starting point for designing interventions that might help people follow through a recommended action. Our phase 2 was simply a real-world deployment test of our recommender, but intervention-style applications that would apply our recommender need to

design personalized strategies. For example, P12 felt that the window popup recommending a break was not enough, "if you can provide a joke instead of asking me to take a break, or sending me analytics about what other employees are doing at this moment, or how many people are suffering at the same problem I have may help me feel better."

Keeping Control with the User

When asked about their thoughts on a hypothetical intelligent system that could guide them throughout the day to keep them both productive and happy, people were generally positive, but wanted control as needed. Knowledge of deadlines, meetings, and other collaborative factors affect people's workday. Unless the tool let them manipulate this meta-level of factors that affect their work, some people felt that an intelligent system could not guide their workflow effectively. Technology can never completely meet the fluid social needs of users – Ackerman calls this the "social-technical gap" and suggests that instead of attempting to build these impossible perfect solutions, we should build first order approximations of them [1]. For our recommender, we added opportunities for user-control in our models via the design of our output variable, which assigns weights to productivity, affect, and an interaction variable between them for productivity-given-stress. We use equal weights, but a user might care more about productivity on one day because of an imminent deadline, or might need more mood-based recommendations on another day when they are feeling particular stressed or unhappy [31].

Understanding Context in the Workplace

We introduced and examined FLOWZONE as a tool for recommending actions for an *individual* worker. However, information workers rarely work in isolation—they are a part of teams within organizations, often collaborating on a daily basis [11]. Several participants mentioned a desire for a team-centric version of FLOWZONE that recommends actions based on the team's context. This requires context about how the team works together: group coherence, communication, and reliance become important. This is not as simple as jointly optimizing happiness and productivity for each individual team member—when people actively rely on each other in a team setting, their productivity and happiness is dependent on that of other members, and interaction effects must be considered. While we wait for technical developments that can enable an understanding of team context, we can apply cluster-based data aggregation as a starting point. When we clustered people by their official job role, we were able to achieve reasonable goodness-of-fit (R^2 values ranged from $0.4 - 0.7$), but the feature importances that were unique to an individual were lost. Our hope is that future work will consider more nuanced clusters by conducting extensive surveys to surface the tacit roles people perform under the umbrella of an official job title, which was out of scope for our data collection and methods.

Ethical Considerations

While emotion, workplace context, and task data logging can help build accurate models for happiness and productivity, there are concerns about worker privacy, both from us and our participants: "feeling like you're being watched all the time would just be bad" (P3). This is an important consideration, as prior work (e.g., [25]) cautions us of the privacy breaches that are impossible to manage once tracking becomes a required or coerced aspect of work. Beyond privacy, building tools for productivity and efficiency is often seen as supporting Taylorism, where employees' effort is optimized for the most output, with no consideration of the individuals [46]. Our efforts oppose this, instead aiming to keep employees happy while completing fulfilling work. We believe in the "happy-productive worker"—being happy at work is what causes people to be more productive [73]—thus our focus is to optimize happiness, while recognizing that getting things done is also necessary.

LIMITATIONS

One limitation of our study is that our modeling setup relies on hourly self-reports of productivity and affect. These hourly self-reports can be a form of interruption of their own. Similar to prior work (e.g., [74], our hope is that our minimal input-based optional setup does not significantly interrupt people.

We validated our approaches via a deployment study lasting three days. Since users can take time to adapt to suggestions and integrate them in their work patterns, longitudinal studies of such recommendations may provide additional insights. Our control condition was designed for the purpose of validating the user feedback of the model condition: we wanted to ensure that people's responses were not an artifact of the system's novelty in their work environment. We thus designed a simple heuristics-based setup rather than comparing to more specific workplace behavior-change approaches (e.g., the Pomodoro approach [12]). In future work, we hope to compare these existing, nuanced approaches in a longitudinal study to better characterize the differences.

Finally, our models use simulated datasets (in combination with real user data) to enable complex modeling techniques such as timeseries forecasting. While simulated datasets are commonplace in other domains (e.g., natural language processing [42]), they are new to domains like workplace recommendations. We validated the integrity of our original + simulated dataset via comparison tests with the original dataset, but hope that future work will consider other ways to acquire and validate these simulated datasets. This can enable complex modeling and system-building at unprecedented scales, for the benefit of users, practitioners, and researchers alike.

CONCLUSION

We explore how user emotion data, workplace context, and task data can be used to develop predictive models for recommending task transitions or breaks with the goal of guiding workers towards more productive, happy work. We find these models to be highly personalized, though we see some commonalities across the same job roles. Validation of our models with real-time recommendations shows 86% accuracy in predicting opportune moments for transitions, and 77% accuracy in predicting breaks. While open research questions remain around how to support users in following through with the recommendations and how to support collaborative settings, our work is a crucial first step towards supporting intelligent systems by providing timely predictions for individuals.

REFERENCES

[1] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203.

[2] Rachel Adler and Raquel Benbunan-Fich. 2013. Self-interruptions in discretionary multitasking. 29 (07 2013), 1441–1449.

[3] Ane Alberdi, Asier Aztiria, and Adrian Basarab. 2016. Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. *Journal of biomedical informatics* 59 (2016), 49–75.

[4] 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), 21.

[5] Brian P Bailey and Joseph A Konstan. 2006. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in human behavior* 22, 4 (2006), 685–708.

[6] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 279–283.

[7] James "Bo" Begole, Nicholas E. Matsakis, and John C. Tang. 2004. Lilsys: Sensing Unavailability. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work (CSCW '04)*. ACM, New York, NY, USA, 511–514.

[8] Raquel Benbunan-Fich and Gregory Truman. 2009. Multitasking with laptops during Meetings. 52 (02 2009), 139–141.

[9] Scott A. Cambo, Daniel Avrahami, and Matthew L. Lee. 2017. BreakSense: Combining Physiological and Location Sensing to Promote Mobility During Work-Breaks. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3595–3607. DOI:http://dx.doi.org/10.1145/3025453.3026021

[10] Daniel Chen, Jamie Hart, and Roel Vertegaal. 2007. Towards a physiological model of user interruptability. In *IFIP Conference on Human-Computer Interaction*. Springer, 439–451.

[11] Jan Chong and Rosanne Siino. 2006. Interruptions on Software Teams: A Comparison of Paired and Solo Programmers. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work (CSCW '06)*. ACM, New York, NY, USA, 29–38. DOI: http://dx.doi.org/10.1145/1180875.1180882

[12] Francesco Cirillo. 2006. The pomodoro technique (the pomodoro). *Agile Processes in Software Engineering and* 54, 2 (2006).

[13] Russell Cropanzano and Thomas A Wright. 2001. When a" happy" worker is really a" productive" worker: A review and further refinement of the happy-productive worker thesis. *Consulting Psychology Journal: Practice and Research* 53, 3 (2001), 182.

[14] Mihaly Csikszentmihalyi. 1997. *Finding flow: The psychology of engagement with everyday life.* Basic Books.

[15] Edward B. Cutrell, Mary Czerwinski, and Eric Horvitz. 2000. Effects of Instant Messaging Interruptions on Computing Tasks. In *CHI '00 Extended Abstracts on Human Factors in Computing Systems (CHI EA '00)*. ACM, New York, NY, USA, 99–100. DOI: http://dx.doi.org/10.1145/633292.633351

[16] Mary Czerwinski, Eric Horvitz, and Susan Wilhite. 2004. A Diary Study of Task Switching and Interruptions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 175–182. DOI: http://dx.doi.org/10.1145/985692.985715

[17] Laura Dabbish, Gloria Mark, and Víctor M González. 2011. Why do i keep interrupting myself?: environment, habit and self-interruption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3127–3130.

[18] Giuseppe Destefanis, Marco Ortu, Steve Counsell, Stephen Swift, Michele Marchesi, and Roberto Tonelli. 2016. Software development: do good manners matter? *PeerJ Computer Science* 2 (2016), e73.

[19] Paul Ekman, E Richard Sorenson, and Wallace V Friesen. 1969. Pan-cultural elements in facial displays of emotion. *Science* 164, 3875 (1969), 86–88.

[20] Daniel A. Epstein, Daniel Avrahami, and Jacob T. Biehl. 2016. Taking 5: Work-Breaks, Productivity, and Opportunities for Personal Informatics for Knowledge Workers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 673–684. DOI: http://dx.doi.org/10.1145/2858036.2858066

[21] James Fogarty, Scott E Hudson, Christopher G Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny C Lee, and Jie Yang. 2005. Predicting human interruptibility with sensors. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12, 1 (2005), 119–146.

[22] Anonymized for review. "I Didn't Know I looked Angry": Characterizing Misalignment of Observed Emotion and Reported Affect. In *Under Review for CHI 2020*. ACM.

[23] Pamela S Galluch, Varun Grover, and Jason Bennett Thatcher. 2015. Interrupting the workplace: Examining stressors in an information technology context. *Journal of the Association for Information Systems* 16, 1 (2015), 1.

[24] Victor M González and Gloria Mark. 2004. Constant, constant, multi-tasking craziness: managing multiple working spheres. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 113–120.

[25] Nanna Gorm and Irina Shklovski. 2016. Sharing steps in the workplace: Changing privacy concerns over time. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4315–4319.

[26] Daniel Graziotin, Fabian Fagerholm, Xiaofeng Wang, and Pekka Abrahamsson. 2017. On the unhappiness of software developers. In *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*. ACM, 324–333.

[27] Daniel Graziotin, Xiaofeng Wang, and Pekka Abrahamsson. 2014. Happy software developers solve problems better: psychological measurements in empirical software engineering. *PeerJ* 2 (2014), e289.

[28] Barry Gruenberg. 1980. The happy worker: An analysis of educational and occupational differences in determinants of job satisfaction. *American journal of sociology* 86, 2 (1980), 247–271.

[29] Fangfang Guo, Yu Li, Mohan S Kankanhalli, and Michael S Brown. 2013. An evaluation of wearable activity monitoring devices. In *Proceedings of the 1st ACM international workshop on Personal data meets distributed multimedia*. ACM, 31–34.

[30] Eija Haapalainen, SeungJun Kim, Jodi F Forlizzi, and Anind K Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 301–310.

[31] F Maxwell Harper, Funing Xu, Harmanpreet Kaur, Kyle Condiff, Shuo Chang, and Loren Terveen. 2015. Putting users in control of their recommendations. In *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, 3–10.

[32] Jennifer Healey, Rosalind W Picard, and others. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems* 6, 2 (2005), 156–166.

[33] Nis Hjortskov, Dag Rissén, Anne Katrine Blangsted, Nils Fallentin, Ulf Lundberg, and Karen Søgaard. 2004. The effect of mental stress on heart rate variability and blood pressure during computer work. *European journal of applied physiology* 92, 1-2 (2004), 84–89.

[34] Eric Horvitz, Paul Koch, and Johnson Apacible. 2004. BusyBody: Creating and Fielding Personalized Models of the Cost of Interruption. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work (CSCW '04)*. ACM, New York, NY, USA, 507–510. DOI:http://dx.doi.org/10.1145/1031607.1031690

[35] James M. Hudson, Jim Christensen, Wendy A. Kellogg, and Thomas Erickson. 2002. "I'D Be Overwhelmed, but It's Just One More Thing to Do": Availability and Interruption in Research Management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '02)*. ACM, New York, NY, USA, 97–104. DOI: http://dx.doi.org/10.1145/503376.503394

[36] Christophe Hurter, Benjamin R. Cowan, Audrey Girouard, and Nathalie Henry Riche. 2012. Active Progress Bar: Aiding the Switch to Temporary Activities. In *Proceedings of the 26th Annual BCS Interaction Specialist Group Conference on People and Computers (BCS-HCI '12)*. British Computer Society, Swinton, UK, UK, 99–108.

[37] Shamsi T. Iqbal. 2007. Disruption and Recovery of Computing Tasks: Field Study, Analysis, and Directions. In *In Proceedings of the Conference on Human Factors in Computing Systems - CHI 2007 (Apr. 28-May 3*. ACM, 677–686.

[38] Shamsi T Iqbal, Piotr D Adamczyk, Xianjun Sam Zheng, and Brian P Bailey. 2005. Towards an index of opportunity: understanding changes in mental workload during task execution. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 311–320.

[39] Shamsi T. Iqbal and Brian P. Bailey. 2010. Oasis: A Framework for Linking Notification Delivery to the Perceptual Structure of Goal-directed Tasks. *ACM Trans. Comput.-Hum. Interact.* 17, 4, Article 15 (Dec. 2010), 28 pages.

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

[41] Shamsi T Iqbal, Xianjun Sam Zheng, and Brian P Bailey. 2004. Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI'04 extended abstracts on Human factors in computing systems*. ACM, 1477–1480.

[42] Youxuan Jiang, Jonathan K. Kummerfeld, and Walter S. Lasecki. 2017. Understanding Task Design Trade-offs in Crowdsourced Paraphrase Collection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 103–109.

[43] Jing Jin and Laura A. Dabbish. 2009. Self-interruption on the Computer: A Typology of Discretionary Task Interleaving. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 1799–1808. DOI: http://dx.doi.org/10.1145/1518701.1518979

[44] Pamela Karr-Wisniewski and Ying Lu. 2010. When more is too much: Operationalizing technology overload and exploring its impact on knowledge worker productivity. *Computers in Human Behavior* 26, 5 (2010), 1061–1072.

[45] Ioanna Katidioti, Jelmer P Borst, Douwe J Bierens de Haan, Tamara Pepping, Marieke K van Vugt, and Niels A Taatgen. 2016. Interrupted by your pupil: An interruption management system based on pupil dilation. *International Journal of Human–Computer Interaction* 32, 10 (2016), 791–801.

[46] Craig R Littler. 1978. Understanding taylorism. *British Journal of Sociology* (1978), 185–202.

[47] Yuhan Luo, Bongshin Lee, Donghee Yvette Wohn, Amanda L. Rebar, David E. Conroy, and Eun Kyoung Choe. 2018. Time for Break: Understanding Information Workers' Sedentary Behavior Through a Break Prompting System. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 127, 14 pages. DOI:http://dx.doi.org/10.1145/3173574.3173701

[48] Mika Mäntylä, Bram Adams, Giuseppe Destefanis, Daniel Graziotin, and Marco Ortu. 2016. Mining valence, arousal, and dominance: possibilities for detecting burnout and productivity?. In *Proceedings of the 13th International Conference on Mining Software Repositories*. ACM, 247–258.

[49] Gloria Mark, Daniela Gudith, and Ulrich Klocke. 2008. The Cost of Interrupted Work: More Speed and Stress. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 107–110. DOI: http://dx.doi.org/10.1145/1357054.1357072

[50] Gloria Mark, Shamsi Iqbal, Mary Czerwinski, and Paul Johns. 2015. Focused, Aroused, but So Distractible: Temporal Perspectives on Multitasking and Communications. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 903–916. DOI: http://dx.doi.org/10.1145/2675133.2675221

[51] Gloria Mark, Shamsi T Iqbal, Mary Czerwinski, and Paul Johns. 2014. Bored mondays and focused afternoons: the rhythm of attention and online activity in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3025–3034.

[52] Gloria Mark, Shamsi T. Iqbal, Mary Czerwinski, Paul Johns, and Akane Sano. 2016. Neurotics Can'T Focus: An in Situ Study of Online Multitasking in the Workplace. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1739–1744. DOI: http://dx.doi.org/10.1145/2858036.2858202

[53] Daniel McDuff, Sarah Gontarek, and Rosalind W Picard. 2014. Improvements in remote cardiopulmonary measurement using a five band digital camera. *IEEE Transactions on Biomedical Engineering* 61, 10 (2014), 2593–2601.

[54] Daniel McDuff, Kael Rowan, Piali Choudhury, Jessica Wolk, ThuVan Pham, and Mary Czerwinski. 2019. A Multimodal Emotion Sensing Platform for Building Emotion-Aware Applications. *arXiv preprint arXiv:1903.12133* (2019).

[55] Daniel McFarlane and Kara Latorella. 2002. The Scope and Importance of Human Interruption in Human-Computer Interaction Design. *Human-computer Interaction* 17 (03 2002), 1–61. DOI: http://dx.doi.org/10.1207/S15327051HCI1701_1

[56] Wesley P McTernan, Maureen F Dollard, and Anthony D LaMontagne. 2013. Depression in the workplace: An economic cost analysis of depression-related productivity loss attributable to job strain and bullying. *Work & Stress* 27, 4 (2013), 321–338.

[57] Brid O'Conaill and David Frohlich. 1995. Timespace in the Workplace: Dealing with Interruptions. In *Conference Companion on Human Factors in Computing Systems (CHI '95)*. ACM, New York, NY, USA, 262–263. DOI: http://dx.doi.org/10.1145/223355.223665

[58] Marco Ortu, Bram Adams, Giuseppe Destefanis, Parastou Tourani, Michele Marchesi, and Roberto Tonelli. 2015. Are bullies more productive?: empirical study of affectiveness vs. issue fixing time. In *Proceedings of the 12th Working Conference on Mining Software Repositories*. IEEE Press, 303–313.

[59] Leslie A Perlow. 1999. The time famine: Toward a sociology of work time. *Administrative science quarterly* 44, 1 (1999).

[60] June J Pilcher, Douglas R Ginter, and Brigitte Sadowsky. 1997. Sleep quality versus sleep quantity: relationships between sleep and measures of health, well-being and sleepiness in college students. *Journal of psychosomatic research* 42, 6 (1997), 583–596.

[61] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. 2011. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering* 58, 1 (2011), 7–11.

[62] Peter Richter, Thomas Wagner, Ralf Heger, and Gunther Weise. 1998. Psychophysiological analysis of mental load during driving on rural roads-a quasi-experimental field study. *Ergonomics* 41, 5 (1998), 593–609.

[63] Joshua S. Rubinstein, David Meyer, and Jeffrey E. Evans. 2001. Executive Control of Cognitive Processes in Task Switching. 27 (09 2001), 763–97.

[64] Greg Smith, Patrick Baudisch, George Robertson, Mary Czerwinski, Brian Meyers, Daniel Robbins, and Donna Andrews. 2003. GroupBar: The TaskBar Evolved. In *PROCEEDINGS OF OZCHI 2003*. 34–43.

[65] Barry M Staw. 1986. Organizational psychology and the pursuit of the happy/productive worker. *California Management Review* 28, 4 (1986), 40–53.

[66] Toon W Taris and Paul JG Schreurs. 2009. Well-being and organizational performance: An organizational-level test of the happy-productive worker hypothesis. *Work & Stress* 23, 2 (2009), 120–136.

[67] Vincent W.-S. Tseng, Matthew L. Lee, Laurent Denoue, and Daniel Avrahami. 2019. Overcoming Distractions During Transitions from Break to Work Using a Conversational Website-Blocking System. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 467, 13 pages. DOI: http://dx.doi.org/10.1145/3290605.3300697

[68] Antoine U Viola, Lynette M James, Luc JM Schlangen, and Derk-Jan Dijk. 2008. Blue-enriched white light in the workplace improves self-reported alertness, performance and sleep quality. *Scandinavian journal of work, environment & health* (2008), 297–306.

[69] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.

[70] Christopher D. Wickens. 2008. Multiple Resources and Mental Workload. *Human Factors* 50, 3 (2008), 449–455. DOI: http://dx.doi.org/10.1518/001872008X288394 PMID: 18689052.

[71] Alex C Williams, Harmanpreet Kaur, Gloria Mark, Anne Loomis Thompson, Shamsi T Iqbal, and Jaime Teevan. 2018. Supporting Workplace Detachment and Reattachment with Conversational Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 88.

[72] Thomas A Wright and Russell Cropanzano. 1997. Well-being, Satisfaction and Job Performance: Another Look at the Happy/Productive Worker Thesis.. In *Academy of Management Proceedings*, Vol. 1997. Academy of Management Briarcliff Manor, NY 10510, 364–368.

[73] Thomas A Wright and Barry M Staw. 1999. Affect and favorable work outcomes: two longitudinal tests of the happy–productive worker thesis. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 20, 1 (1999), 1–23.

[74] Manuela Züger, Sebastian C Müller, André N Meyer, and Thomas Fritz. 2018. Sensing Interruptibility in the Office: A Field Study on the Use of Biometric and Computer Interaction Sensors. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 591.