# Pearl: A Technology Probe for Machine-Assisted Reflection on Personal Data

MATTHEW JÖRKE, Stanford University, USA
YASAMAN S. SEFIDGAR, University of Washington, USA
TALIE MASSACHI, Brown University, USA
JINA SUH, Microsoft Research, USA
GONZALO RAMOS, Microsoft Research, USA

Reflection on one's personal data can be an effective tool for supporting wellbeing. However, current wellbeing reflection support tools tend to offer a one-size-fits-all approach, ignoring the diversity of people's wellbeing goals and their agency in the self-reflection process. In this work, we identify an opportunity to help people work toward their wellbeing goals by empowering them to reflect on their data on their own terms. Through a formative study, we inform the design and implementation of Pearl, a workplace wellbeing reflection support tool that allows users to explore their personal data in relation to their wellbeing goal. Pearl is a calendar-based interactive machine teaching system that allows users to visualize data sources and tag regions of interest on their calendar. In return, the system provides insights about these tags that can be saved to a reflection journal. We used Pearl as a technology probe with 12 participants without data science expertise and found that all participants successfully gained insights into their workplace wellbeing. In our analysis, we discuss how Pearl's capabilities facilitate insights, the role of machine assistance in the self-reflection process, and the data sources that participants found most insightful. We conclude with design dimensions for intelligent reflection support systems as inspiration for future work.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; • **Computing methodologies → Machine learning**.

Additional Key Words and Phrases: datasets, neural networks, gaze detection, text tagging

**ACM Reference Format:**
Matthew Jörke, Yasaman S. Sefidgar, Talie Massachi, Jina Suh, and Gonzalo Ramos. 2023. Pearl: A Technology Probe for Machine-Assisted Reflection on Personal Data. In *28th International Conference on Intelligent User Interfaces (IUI '23), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3581641.3584054

## 1 INTRODUCTION

Self-reflection on personal data is a well-established method to help people gain awareness of their health and wellbeing [9, 43] and motivate positive behavior change [42, 54]. As such, providing people with the ability to self-reflect is an important design goal for systems aiming to support wellbeing [10]. However, building a system that encourages reflection can be challenging, particularly when supporting people without data analysis expertise.

First, many self-reflection tools rely on manual data entry to collect personal data, which is burdensome to the user [15] and can lead to abandonment [25, 57]. A popular alternative, particularly among commercial wellbeing support tools, automatically collects personal data through passive sensing. To communicate insights back to the user, the system analyzes the data and presents automatically generated visualizations, summaries, or suggestions. Although this process reduces the barrier to self-reflection, its "one-size-fits-all" approach ignores the diversity of user wellbeing goals and the user's agency to express how *they* want to improve their wellbeing. Consider a person who frequently answers emails in the late evening—should a wellbeing support system nudge them to avoid this behavior? If this person is a parent and their goal is to have flexible working hours such that they can spend more afternoons with their children, such a nudge might be inappropriate.

Moreover, interpreting data is difficult for non-experts [29, 57], as it requires users to bridge the explanatory gap between data and the wellbeing objectives they care about. Many existing systems implicitly assume that reflection occurs naturally once data is transformed, presented, and visualized [9, 10]. This contradicts leading theories of reflection [59], which underscore that people need to be supported in the self-reflection process [10].

In this work, we set out to explore the potential for reflection support systems to move beyond automated insights and empower people to explore their personal data on their own terms. In particular, we identify an opportunity for human-AI collaboration to enable people to flexibly express their wellbeing goals and support them in making sense of their personal data. We do so by leveraging an interplay between machine teaching and sensemaking [51]: we hypothesize that the act of teaching a machine about a goal and the wellbeing concepts one cares about may simultaneously be a reflective activity. Such an activity could help people make sense of their wellbeing goal and gain insights into how their personal data relates to it, while providing a machine learner with goal-relevant information that can be used to provide automated assistance.

Through a formative study, we generate a set of design principles that guide the development of PEARL[1], a workplace wellbeing reflection support system that allows people without data analysis expertise to make connections between the data collected about them and their own wellbeing goals. PEARL provides people with the agency to explore the data collected about them and define wellbeing-related concepts by tagging regions of interest on their calendar. In turn, an interactive machine learning system extracts insights about these concepts to support people in the process of reflecting on their goals and how to achieve them. These insights, along with any personal notes, can be saved in a digital journal to facilitate self-reflection and document one's learnings for later revisitation or potential sharing.

We use PEARL as a technology probe to explore the potential of our approach to support reflection on wellbeing goals for information workers. Our study had participants with limited data science experience use PEARL to reflect on their computer activity collected over 2-3 weeks. We find that all 12 participants were able to gain insights into their wellbeing by using PEARL. In our analysis, we discuss how each of PEARL's capabilities facilitated or hindered insight generation, the role of machine assistance in the self-reflection process, and the data sources participants found most insightful. We find that machine assistance can encourage self-reflection through different means, both by automating tedious tasks and by encouraging goal decomposition. In addition, we report on several design dimensions for intelligent reflection support tools that emerge from our study and serve as inspiration for future work.

Our work makes the following contributions:

1) **PEARL: an interactive machine teaching system for workplace wellbeing reflection support.** PEARL operationalizes our design principles in the context of workplace wellbeing support and implements an interactive data visualization and machine learning system.
2) **The results of a technology probe evaluation of PEARL with twelve participants.** We evaluate PEARL as a technology probe with 12 participants and document insights and themes from analyzing participants' interactions with the system.

[1] PEARL, short for "**pe**rson**a**l **r**ef**l**ection"

3) **Design insights for supporting wellbeing through machine-assisted self-reflection on personal data.** We contribute design principles for intelligent reflection support systems extracted from our formative user study, design insights surfaced in our technology probe evaluation, and design dimensions for future work in this space.

## 2 RELATED WORK

This section summarizes and situates our approach in prior work on workplace wellbeing, self-reflection on personal data, and human-AI interaction.

### 2.1 Workplace Wellbeing

Our work aims to support the wellbeing of information workers and explores how technology can assist the process of self-reflection on personal data that is passively captured in the workplace. Work is a major source of challenge for achieving wellbeing [1]. Since work-related stress leads to many negative downstream consequences (e.g., mental and physical health disorders, decreased productivity due to absenteeism and burnout) [19], the workplace is an opportune domain for which to develop wellbeing support tools. Information workers [45], who are tasked with processing information rather than physical objects, typically by using computing technology, face many challenges to their workplace wellbeing. Information workers' flexible nature of work [41] and constant connectivity to work [49] leads to blurred work-non-work boundaries that impact their overall wellbeing [48, 63]. Digital connectivity and digitization of information work, however, provides an opportunity to study the role of data generated at work in achieving personal workplace wellbeing goals.

There has been a growing interest in using passive sensing technologies to monitor or improve workplace wellbeing [20, 55]. Although prior research has reported misalignment between passively-sensed measurements and self-reported user perceptions ofwellbeing [21, 38], tracked data, regardless of where they come from, can provide visibility into past behaviors and facilitate reflection that informs future behaviors [24]. For example, Meyer et al. found self-monitoring at work to be helpful in increasing the awareness about work [47]. Other work in HCI has demonstrated the value of technology-facilitated wellbeing reflection at work through conversational agents [39], diaries [14], or experience sampling methods (ESM) [30].

### 2.2 Self-Reflection on Personal Data

Supporting reflection is a common design goal in HCI systems [9, 11]. The field of *personal informatics* aims to design and study the use of systems that "help people collect and reflect on personal information" [42]. Li et al. [42] define the term and propose the stage-based model of personal informatics, which has since been expanded in subsequent work [25, 43]. Among personal informatics systems aiming to support reflection, visualization is a common medium for communicating insights [5, 16, 34, 46, 61]. Choe et al. [18] develop the Visualized Self dashboard to characterize the types of insights people gain from visual data exploration, building on a codebook of visual insights extracted from [16]. Sukumar et al. [60] expand on the insights presented in [16, 18] and present design directions for personal visualization. We incorporate these findings in the design of our system's visualizations. In addition, we draw inspiration from systems that utilize journals (or diaries) as a tool for documenting and reflecting on insights [17, 23, 44, 53].

Rapp et al. [57] investigate how users without self-tracking experience perceive and use self-tracking tools. Through a diary study, they discover that manual self-tracking can be burdensome, different tools were poorly integrated, visualizations were too abstract and unactionable, and that tools ultimately required too much effort to be beneficial for users without self-tracking experience. The results reported in this work inform many of our design decisions.

In a systematic review of reflection support systems, Bentvelzen et al. [11] analyze design techniques utilized to support reflection and report design patterns and resources for future systems, which we draw from in our work. Towards standardizing the evaluation of reflection support tools, Bentvelzen et al. [10] propose the Technology-Supported Reflection Inventory (TSRI), a scale that evaluates how effectively a system supports refection. We utilize the TRSI in our system evaluation.

### 2.3 Human-AI Interaction

Human interaction with intelligent agents has long been a subject of academic interest [32, 33, 52]. In light of widespread deployment of highly performant AI systems, Amershi et al. [4] propose a set of guidelines for human-AI interaction. We incorporate these principles in the design of our system.

Approaches in which humans iteratively interact with a machine learning system are typically called human-in-the-loop or interactive machine learning [3, 26]. Among the many perspectives in this space, we are more closely aligned with *interactive machine teaching* (IMT) [56, 58]. IMT specifies the role of the human-in-the-loop as a teacher who provides knowledge and information to the machine learner. Core to this teaching perspective is the notion of a teaching language, which defines the ways in which a teacher can communicate knowledge to the learner. As Ng et al. [51] describe, a teaching language can include not only labels, but also richer forms of knowledge such as concepts, rules, or relationships. Ng et al. [51] also discuss the role of knowledge decomposition in the machine teaching process, whereby the teacher identifies and expresses "useful knowledge by breaking it down into its constituent parts or relationships" such that it can be articulated to the learner. In designing Pearl, we aimed to support richer teaching languages and to support people in the knowledge decomposition process. Lastly, a machine teacher is not required to be an expert in machine learning. Similarly, our aim is to provide reflection support to people without expertise in data analysis.

## 3 FORMATIVE STUDY

Although prior work provides tangible lessons on self-reflection on personal data [9, 10, 57], personal visualization [18, 60], as well as interactive machine learning and machine teaching interactions [3, 51], it does not directly apply to intelligent reflection support in the context of the workplace and data collected about information work (e.g., application usage logs, keystrokes, email activity, etc.). We therefore conducted a formative user study to ground our design in the context of workplace wellbeing support and to study the potential for machine teaching to support self-reflection. In particular, we wanted to investigate: (1) how people form connections between wellbeing goals and passively sensed data, (2) the types of machine assistance people desire in the context of workplace wellbeing support, and (3) the teaching languages people prefer when teaching an intelligent wellbeing support system about their wellbeing goals.

### 3.1 Participants

We conducted the study with six researchers with expertise in workplace wellbeing or machine learning applied to wellbeing. These experts have the domain knowledge to help us identify potential opportunities for machine assistance within the realm of technical feasibility.

### 3.2 Study Design

Participants interacted with a hypothetical reflection support system that we prototyped in a Wizard-of-Oz fashion using a shared Figma[2] board (Fig. 1). Before the session, we asked participants to state a wellbeing goal and answer a

[2] www.figma.com

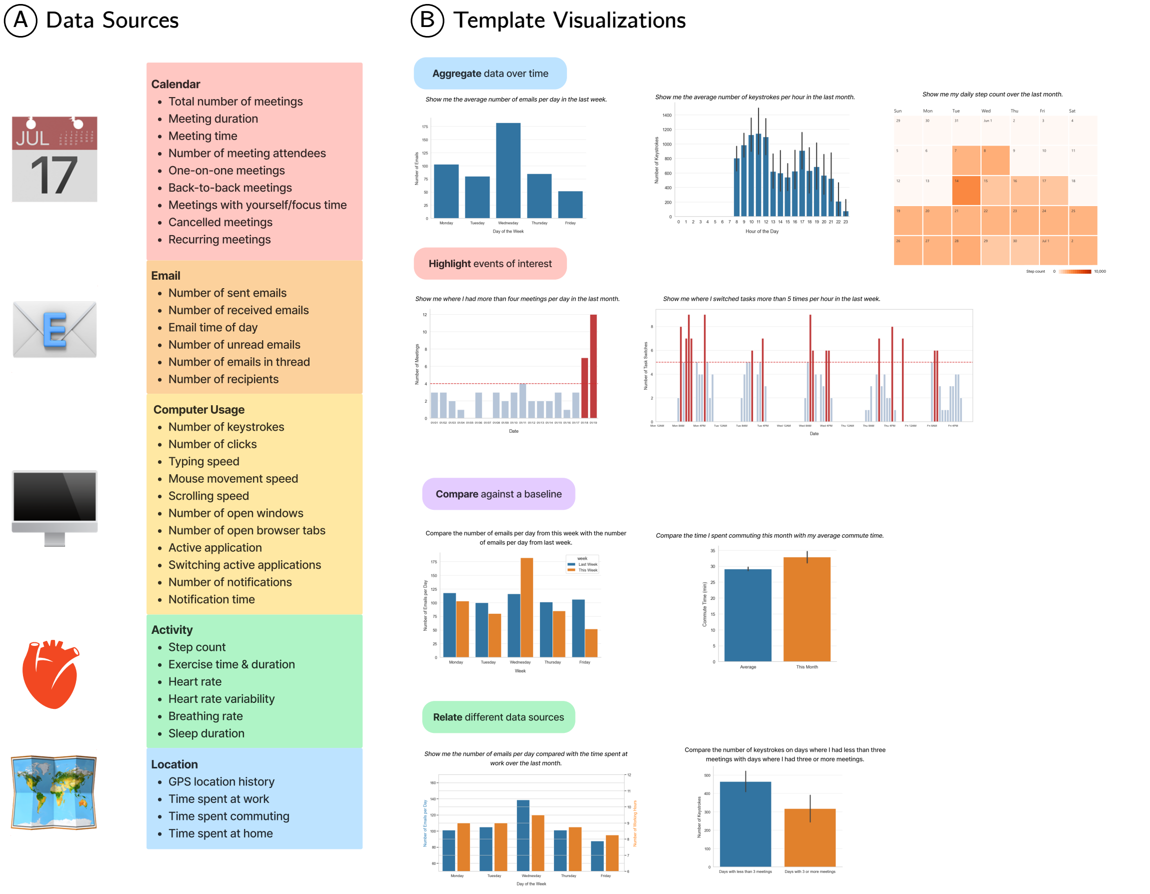


Fig. 1. **Formative Study Materials.** (A) Data sources that a hypothetical wellbeing support system can support, organized by categories. (B) Template visualizations, organized by the kinds of data queries the hypothetical system can answer.

series of questions that were intended to make their goal more specific and actionable.[3] Each study session lasted one hour and consisted of three parts. First, we showed participants a list of data sources organized by categories and asked them to think about which kinds of data might be relevant to their goal (Fig. 1A). After they identified several data sources, we showed participants a series of template visualizations to illustrate the kinds of visual insights the system could provide. We presented charts that aggregated data over time, highlighted events of interest, compared a data source against a baseline, and related two different data sources (Fig. 1B). Drawing from our template visualizations as inspiration, we asked participants to state questions they would like to ask about their personal data that could be answered using a visualization that this system could hypothetically generate. Lastly, we told participants to imagine that they had the ability to tell the system about things that are important in relation to their wellbeing goal and asked them to describe how they could explain this to the system. Additionally, we asked participants to state any other forms of machine assistance that they would like to receive from an intelligent wellbeing support system.

[3] These questions are identical to the goal specification questions in our evaluation study (Section 5.2).

### 3.3 Study Insights

*Connecting Data Sources to Wellbeing Concepts.* We observed a consistent pattern of behavior among all our participants when connecting their wellbeing goals with system data sources. Participants would first scan through the list of data sources until they found one that they believed to be relevant to their wellbeing goal. After identifying a relevant data source, participants would offer an explanation for why they believed that the data source was relevant to their goal. Implicit in this explanation were intermediary concepts that connected their high-level, abstract wellbeing goals with concrete system measurements. For example, a participant with the goal "I would like to improve my work-life balance" connected this goal to the data source "number of keystrokes" by offering the explanation that the number of keystrokes could indicate when they were working after hours. Importantly, the concept of "working after hours" is both measurable with respect to the system's data and relevant with respect to the person's wellbeing goal. This process relates to the notion of *knowledge decomposition* and has previously been studied in the context of machine teaching for document classification [51]. We aim to support knowledge decomposition as a capability of Pearl.

*Machine Assistance.* Many participants expressed hesitation in allowing a wellbeing support system to take control of important work activities, such as rescheduling calendar events or muting notifications. Instead, participants generally preferred system-generated suggestions (e.g., "Try to reduce the number of meetings on Tuesdays") or summarized insights (e.g., "You have 42% more unread emails on days when you have back-to-back meetings.") that they would have the agency to take action on or ignore. In fact, we observed several participants deliberating and navigating the appropriate level of system agency while speaking aloud. For example, a participant who wanted to find more uninterrupted chunks of time for focused work remarked, "*I can start to label, okay this was an uninterrupted chunk, this was an uninterrupted chunk, but I could also give it ways to group meetings to create that space. I don't know exactly how to express that.*" After deliberating how the system might go about automatically scheduling uninterrupted chunks, they concluded, "*I'm not sure if the system would know when is a good time for me to do [focused work], I guess could tell it, like let's do it at 9am, and it could tell me if I'm doing anything at 9am, but... in order for me to tell the system that 9am is a good time, I would have to have discovered it in my data already.*"

*Teaching Languages.* All participants mentioned some form of labeling interaction, such as labeling days when they did or did not meet their goal or labeling calendar events that were indicative of a wellbeing concept (e.g., focus, productivity, energy). While some participants described rule-based teaching interactions (e.g., "working after hours" corresponds to having keyboard activity after 6pm), others mentioned that rule-based teaching would be challenging because the concepts they cared about tracking most tended to be those for which a rule was unknown. For example, it could be challenging to define a rule for "productivity", but it would be easy to identify moments in time that were used productively. Similarly, a participant who wanted to feel more energetic at work commented, "*I would like to give the system some of these examples because it is difficult for me to come up with the rules. I don't always know what is making me energetic, it could be a number of factors.*" They qualified this preference for label-based teaching by stating that if they knew precisely which factors were affecting their energy levels, they would not have needed the system in the first place. Lastly, a few participants expressed interest in labeling individual data sources' relevance (both the directionality and strength) to help the system learn an accurate model more quickly and to increase their trust in the system's inferences.

## 4 PEARL: SYSTEM DESIGN & IMPLEMENTATION

In this section, we present Pearl: a technology probe for machine-assisted reflection on personal data. We discuss our design considerations and present our system's architecture and user interface.

### 4.1 Design Principles

These principles embody our vision for intelligent reflection support systems and incorporate our learnings on how best to support this vision, informed by our literature review and formative user study. We use our principles to guide the design of Pearl and the main capabilities the system provides.

**DP-1: Lower the barrier to gaining wellbeing insights from personal data.**
Our formative study revealed the many ways in which hypothetical data could be used to gain insights into wellbeing. However, it is known that supporting insight generation is challenging in practice [10], particularly for people without prior expertise [57]. We aim to lower the cognitive effort needed to explore, contextualize, and gain insight from personal data.

**DP-2: Support the diverse ways people express, explore, and operationalize their wellbeing goals.**
Standardized reflection support tools ignore the diversity of people's wellbeing goals. Moreover, our formative study uncovered the rich ways in which people decompose wellbeing goals when connecting them with personal data. We recognize the inherent value in intentional goal-setting and aim to support people in connecting personal data with their unique wellbeing goals.

**DP-3: Use machine assistance to augment—not automate—people's self-reflection process.**
While self-reflection is challenging, removing humans from the loop entirely denies their agency in a deeply personal process. Participants in our formative study also emphasized the importance of maintaining an appropriate level of agency. We focus on machine assistance that augments the reflection process and seek interactions that provide a meaningful level of agency.

### 4.2 Technology Probes

We designed and developed Pearl as a *technology probe* [35] to assess the viability of our ideas about intelligent reflection support systems and generate new questions about their design. Probes are a method to engage participants early in the design process with the purpose "not to capture what is so much as to inspire what might be." [12, p. 185]. The use of probes in HCI research originates with Gaver et al.'s [28] cultural probes, in which packets of materials and artifacts are shared with participants in an effort to provoke responses and stimulate design conversations. Boehner et al. [13] discuss how cultural probes have since been adapted and reinterpreted by the HCI community in various ways, often with a stronger emphasis on seeking design information than subverting traditional need-finding methods.

A technology probe [35] is a type of probe in which the artifact presented to the user is a functional piece of technology. Technology probes differ from traditional prototypes in that they have simpler functionality and are introduced early in the design process as a tool for inspiring future designs. Hutchinson et al. [35] outline three goals of a technology probe: a *social science* goal of collecting information about the use and users of the technology in context, an *engineering goal* of field-testing the technology, as well as a *design goal* of inspiring new kinds of technology to support user needs and desires. We aim to support and balance each of the three goals when designing and implementing Pearl.

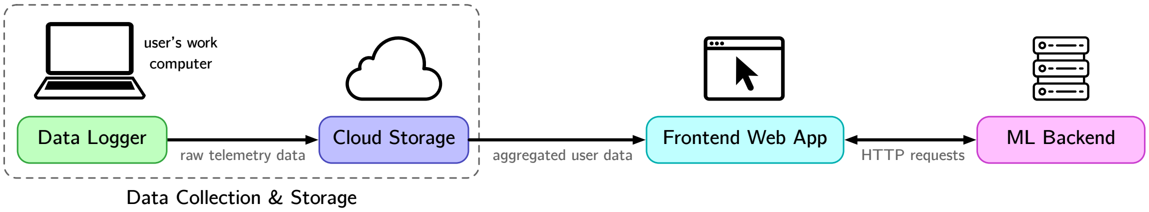


Fig. 2. **System Architecture.** The data logger runs in the background and records raw telemetry data that is sent to cloud storage. The front-end web application fetches data from cloud storage and sends requests to the ML backend for data featurization and model training.

### 4.3 System Architecture

Our implementation of Pearl relies on three interconnected components: data collection and storage infrastructure, an interactive machine learning backend, and a front-end web application.

*Data Collection & Storage Infrastructure.* Pearl's input data is collected using a native Windows application that runs in the background and logs various computer usage metrics, such as keystrokes, mouse events, application usage, and emails. This data is stripped of personally identifiable information and stored in cloud storage. There, the raw interaction signals are processed and aggregated into 30-minute intervals, which determines the minimum temporal granularity of all data sources in our system.

*Machine Learning (ML) Backend.* The back-end server was built using Python and serves API requests from the front-end, performs data processing and featurization, and supports all machine learning functionality. Data processing and featurization are performed using pandas[4], natural language processing of the goal statements and data sources is performed using spacy[5], and the interactive machine learning components are implemented using scikit-learn[6]. We describe our learning algorithms in the relevant sections below.

*Front-End Web Application.* Pearl's front-end is a web application built using React[7]. Upon logging into the system's front-end, the user's data is fetched from our cloud database and sent to the backend for featurization.

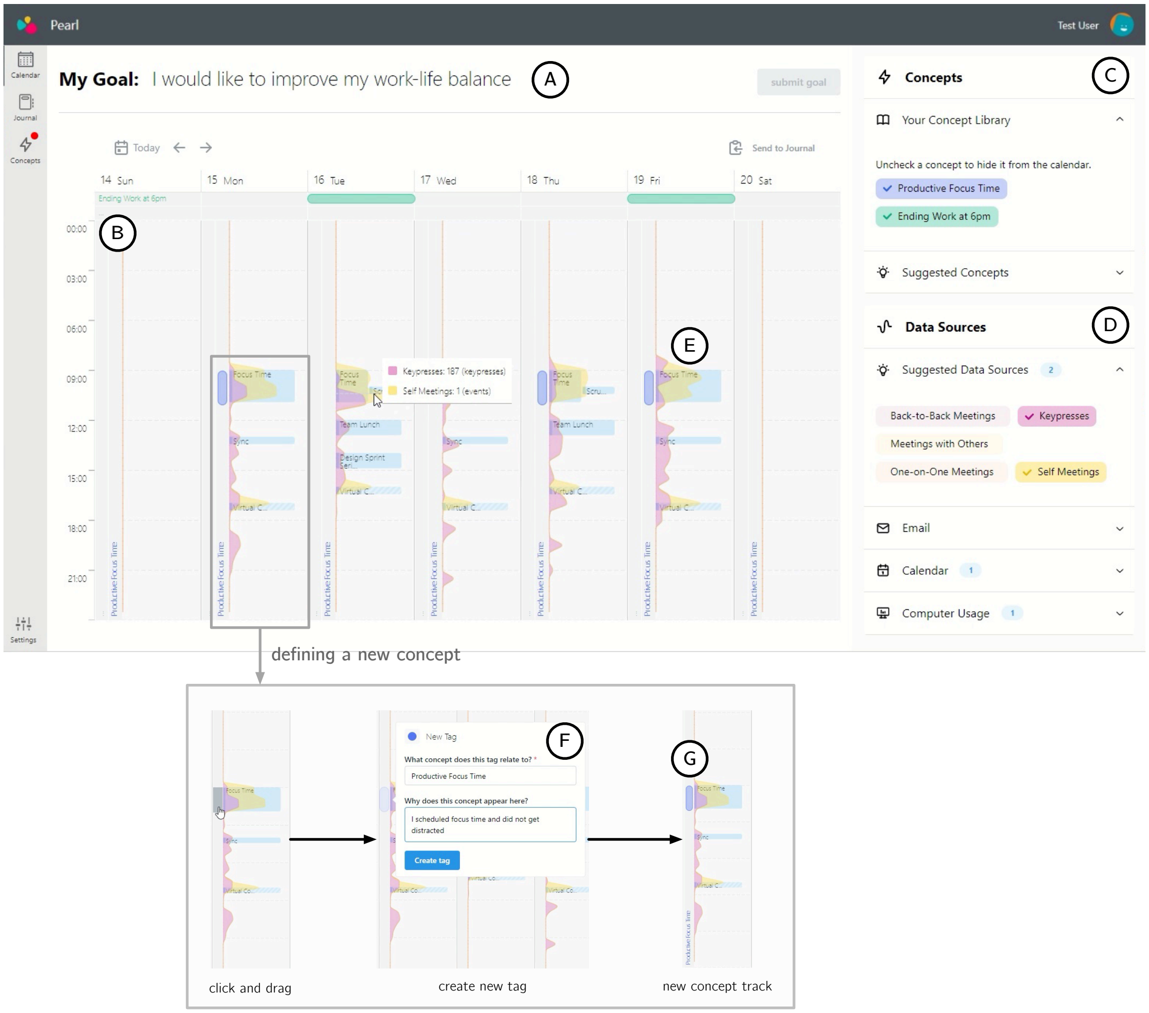
### 4.4 User Interface

In this section, we present the features provided by Pearl's interface and discuss how each feature relates to our design principles. Across all three views, the user's chosen wellbeing goal is displayed at the top of the screen. The calendar view supports visual data exploration on the calendar, as well as creating concepts by tagging regions of time. The concept view presents system-generated visualizations and statistics for user-defined concepts. Finally, the journal view consists of a digital notebook that allows users to save system-generated insights for later reflection and take free-form notes to document their learnings.

#### Goal

The interface displays a user's goal predominantly at the top of the screen and is always present across all three views (Fig. 3A). The goal can be updated at any time and serves as a reminder to guide data exploration and self-reflection **(DP-2)**. The goal statement is also parsed using natural language processing to suggest potentially relevant data sources

[4] https://pandas.pydata.org [5] https://spacy.io [6] https://scikit-learn.org [7] https://reactjs.org



My Goal: I would like to improve my work-life balance

Uncheck a concept to hide it from the calendar.

Keypresses: 187 (keypresses)

Self Meetings: 1 (events)

New Tag

What concept does this tag relate to? * Productive Focus Time

Why does this concept appear here? I scheduled focus time and did not get distracted

Fig. 3. **Calendar View**. (A) The goal area, containing a user's chosen wellbeing goal; (B) the calendar, which displays calendar events and data source visualizations; (C) the concept selector; (D) the data source selector; (E) a visualization of two data sources; (F) the concept definition pop-up menu; (G) a concept track containing one new concept.

**(DP-3)**. The similarity between the goal statement and the data sources is computed using a combination of exact-match heuristics and semantic similarity using non-contextual word embeddings[8].

## Calendar View

The calendar view is the default view of the interface and supports two main activities: (1) visual data source exploration **(DP-1)**, and (2) creating concepts by entering tags **(DP-2)**.

[8] We use the 300-dimensional non-contextual word embeddings provided by spacy's `en_core_web_lg` model (https://spacy.io/models/en)

*Data Source Visualization.* The user's calendar events are displayed on the calendar in the center of the calendar view (Fig. 3B). In the right panel of this view, the user can toggle the visibility of different data sources by using the data source selector (Fig. 3D) **(DP-1)**. By default, the "Suggested Data Source" tab is expanded, which contains five data sources that the system suggests from the user's goal statement **(DP-3)**. The data source panel also groups the sources into calendar, email, and computer usage categories. When a user hovers over a data source, a tooltip appears that contains a text describing that data source **(DP-1)**.

When a data source is selected, a vertical, filled line graph is overlaid on each column (i.e., day) of the calendar (Fig. 3E) similar to the presentations in [34]. When multiple data sources are simultaneously selected, they are displayed as a stacked streamgraph inspired by the data presentation in [5]. Raw sensor values for each data source are displayed on hover. Data sources are deliberately overlaid on top of calendar events to help provide context that can reduce the cognitive effort required to interpret raw signal data **(DP-1)**, a strategy that has also been proposed in prior work [34].

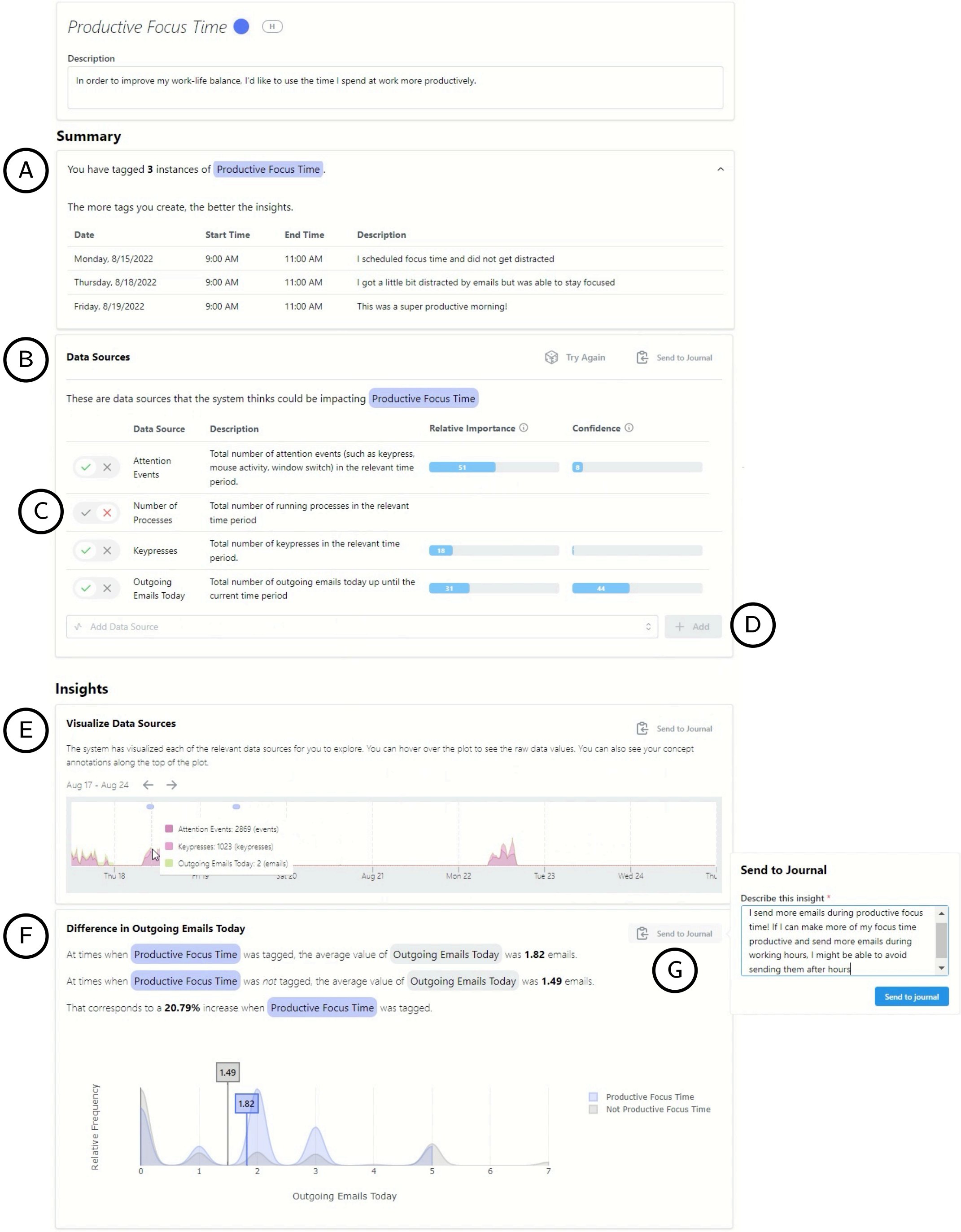
*Concept Definition.* Concepts are high-level categories that are relevant to a user's goals **(DP-2)**. Users can provide a collection of tags for a given concept that correspond to regions of time when a concept was present **(DP-3)**. For example, if "productivity" is the concept of interest, a user could tag individual meetings or focused work sessions that they believe they used productively. Concepts can have tags that correspond to entire days (daily concepts) or regions of time within a given day (hourly concepts). To simplify the way our system learns about concepts, we do not allow the same concept to be both hourly and daily. To create a new hourly tag, the user can click-and-drag along the gray vertical track next to each calendar column. Similarly, the user can click and drag along the horizontal track below the top calendar row to create a new daily tag. After clicking and dragging, a pop-up menu appears asking the user to provide the concept name (or choose among existing concepts from a drop-down menu) and to provide an optional description as to why this tag is indicative of the chosen concept (Fig. 3F) **(DP-2)**. The user can also choose a color to be associated with this concept.

For each new concept, the system inserts a parallel track that the user can use to quickly enter new tags for that concept. Individual tags are displayed as colored bars within each concept's track (Fig. 3G). Concepts also appear in the right panel of this view in a concept selector (Figure 3C). This selector allows the user to toggle the visibility of different concepts in the calendar view **(DP-1)**.

### Concept View

After defining a concept and entering tags in the calendar view, users can explore data sources that may be relevant to concepts, as well as system-generated visualizations and data summaries in the concept view (Fig. 4). The concept view exposes the interactive ML system to the user.

The summary panel contains a card showing how many tags the user has entered, which can be expanded to show the timestamps and descriptions for all tags (Fig. 4A). It also contains an interactive table containing potentially relevant data sources (Fig. 4B) to the currently selected concept. These data sources are initially chosen by the ML system (the procedure by which they are chosen is described below) **(DP-3)**. Each row of the table contains a selector to include/exclude the data source, a text description of that data source, the relative importance of those data sources, and the system's confidence in that data source (Fig. 4C). If a user does not think a system-selected data source is relevant, they can deselect that data source to exclude it from the system **(DP-3)**. The user can also add data sources missed by the system using the drop-down menu at the bottom of the table (Fig. 4D). **(DP-3)**



*Productive Focus Time*

Description

In order to improve my work-life balance, I'd like to use the time I spend at work more productively.

**Summary**

You have tagged **3** instances of Productive Focus Time.

The more tags you create, the better the insights.

| Date | Start Time | End Time | Description |
|---|---|---|---|
| Monday, 8/15/2022 | 9:00 AM | 11:00 AM | I scheduled focus time and did not get distracted |
| Thursday, 8/18/2022 | 9:00 AM | 11:00 AM | I got a little bit distracted by emails but was able to stay focused |
| Friday, 8/19/2022 | 9:00 AM | 11:00 AM | This was a super productive morning! |

**Data Sources**

These are data sources that the system thinks could be impacting Productive Focus Time

| | Data Source | Description | Relative Importance | Confidence |
|---|---|---|---|---|
| ✓ ✕ | Attention Events | Total number of attention events (such as keypress, mouse activity, window switch) in the relevant time period. | 51 | 8 |
| ✓ ✕ | Number of Processes | Total number of running processes in the relevant time period | | |
| ✓ ✕ | Keypresses | Total number of keypresses in the relevant time period. | 18 | |
| ✓ ✕ | Outgoing Emails Today | Total number of outgoing emails today up until the current time period | 31 | 44 |

**Insights**

**Visualize Data Sources**

The system has visualized each of the relevant data sources for you to explore. You can hover over the plot to see the raw data values. You can also see your concept annotations along the top of the plot.

Attention Events: 2869 (events)
Keypresses: 1023 (keypresses)
Outgoing Emails Today: 2 (emails)

**Difference in Outgoing Emails Today**

At times when Productive Focus Time was tagged, the average value of Outgoing Emails Today was **1.82** emails.

At times when Productive Focus Time was *not* tagged, the average value of Outgoing Emails Today was **1.49** emails.

That corresponds to a **20.79%** increase when Productive Focus Time was tagged.

**Send to Journal**

Describe this insight *

I send more emails during productive focus time! If I can make more of my focus time productive and send more emails during working hours, I might be able to avoid sending them after hours

Fig. 4. **Concept View.** (A) A summary card showing the user's tags, (B) the interactive data source selection table, (C) a deselected data source, (D) the menu for adding data sources, (E) an insight card containing a streamgraph visualization of all selected data sources, (F) an insight card containing data summaries and a kernel density plot for a given data source, (G) a send to journal button and pop-up menu.

The insights panel contains visualizations and data summaries for each of the data sources selected by the ML system and/or the user. The first visualization is a stacked streamgraph of all data sources, with tags overlaid along the top of the plot (Fig. 4E). This visualization aims to facilitate the visual inspection of potentially relevant data sources and patterns over time **(DP-1)**. Next, for each of the selected data sources, the system generates a card containing a text description of the difference in the average value of that data source conditioned on the selected concept being present vs. not present, as well as a kernel density plot of the conditional distributions of the selected data source (Fig. 4F). This visualization aims to facilitate insight and reflection on the role that a data source could have on a concept and its connection to a wellbeing goal **(DP-1)**, **(DP-2)**.

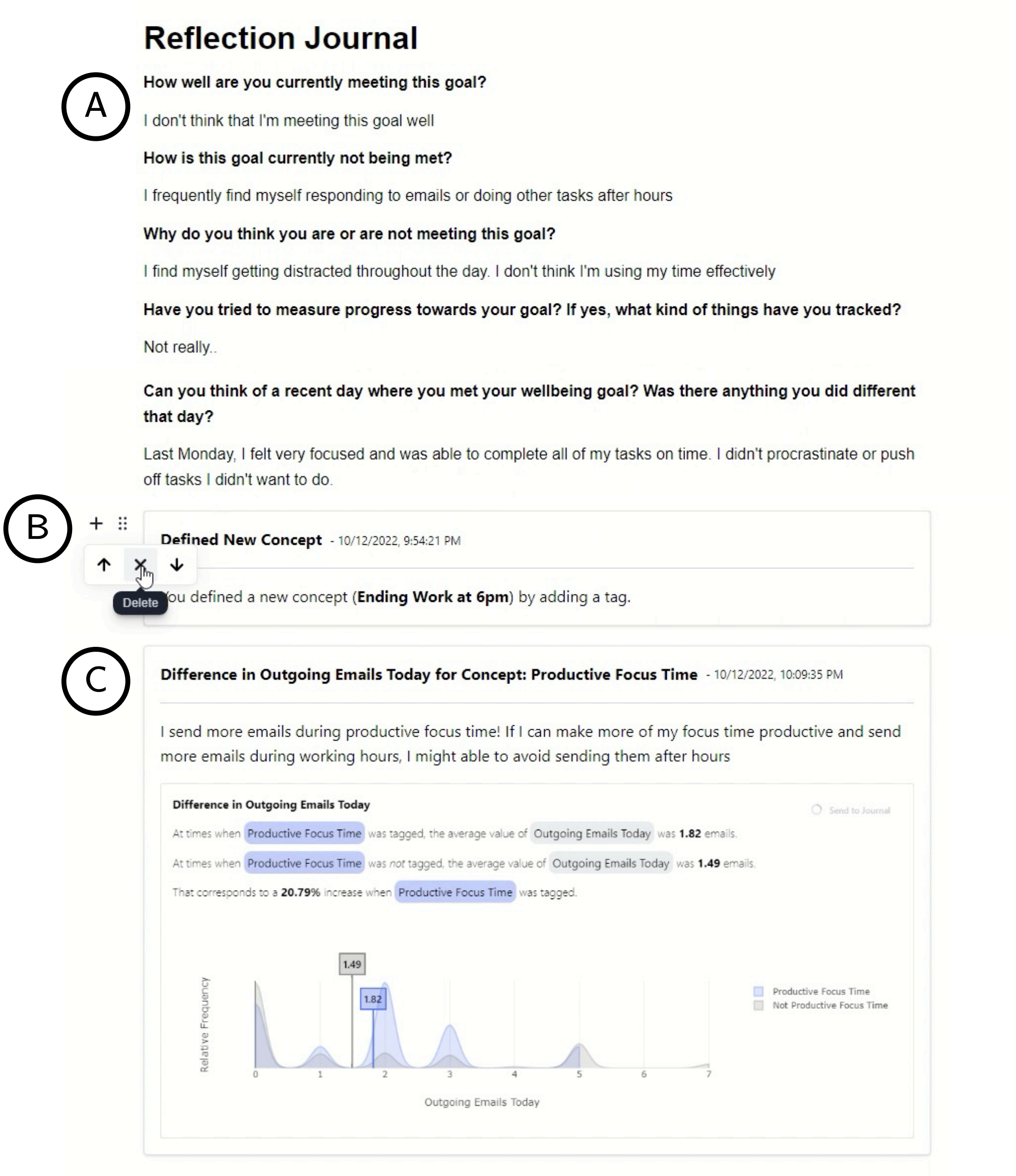
*Concept Learning.* From a machine learning perspective, we define a concept as an unknown function $f$ mapping raw signal data $\mathbf{x} \in \mathbb{R}^d$ to a binary outcome variable $y = f(\mathbf{x})$. The user's tags represent positive instances of a given concept: tuples $(\mathbf{x}_i, y_i)$ where $y_i = 1$. To train a machine learning model to approximate $f$, we first mine negative examples by randomly sampling an equal number of examples from regions of the user's data that do not contain a tag for that concept. We exclude regions of time with missing data from our sampling procedure such that the negative examples do not include times of inactivity or sensor error. If a user has not performed any data source selection, we first perform $\ell_1$-regularized logistic regression to select a small number of potentially relevant data sources. If the user has included or excluded any data sources, we perform $\ell_2$-regularized logistic regression on the subset of included features only.

In particular, for our low data setting (approximately 5–20 labels per concept), we found this process to be highly noisy, which could lead to unnecessary confusion and eventual distrust. To reduce the stochasticity of the model and quantify the uncertainty over the model's weights, we train an ensemble of logistic regression models by repeating the above procedure 100 times. The relative importance scores are computed using the softmax of the absolute value of the average feature weights across all models in the ensemble. The confidence scores are computed using a 95% confidence interval of the weight distribution across all models in the ensemble.

In the spirit of a technology probe, we note that this ML system is purposely straightforward. Although our logistic regression ensembles likely to do not have high predictive accuracy, logistic regression has a number of desirable properties for our setting: it is robust in a low-data setting, it is very fast to train, and it provides direct access to interpretable feature weights. Moreover, we emphasize that the purpose of our ML system was to reduce the search space of potentially relevant data sources such that a user could more efficiently arrive at insightful conclusions **(DP-3)**. Our purpose is not to construct a concept model with high predictive accuracy, which would have led to different modeling and interaction design decisions. For this purpose, we found that our logistic regression ensemble was sufficient to fuel a design probe.

### Journal View

The journal view (Fig. 5) consists of a digital notebook that supports documenting insights for self-reflection. The journal consists of a list of cells that can contain plain text notes and snapshots of elements in the user interface, which the user can freely re-arrange, add, or delete (Fig. 5B). Throughout Pearl, any interface element that can be saved to the journal contains a "Save to Journal" button (Fig. 4G) **(DP-1)**. This functionality aligns with the notion of Visual Mementos in [61]. When clicking this button, the user is prompted to reflect on why this insight is relevant to their goal. This prompting is designed to capture the user's thinking in the moment while encouraging them to connect this insight to their goal **(DP-2)**. After entering their reflection, a screenshot of the element is saved to the journal along



Reflection Journal

**How well are you currently meeting this goal?**

I don't think that I'm meeting this goal well

**How is this goal currently not being met?**

I frequently find myself responding to emails or doing other tasks after hours

**Why do you think you are or are not meeting this goal?**

I find myself getting distracted throughout the day. I don't think I'm using my time effectively

**Have you tried to measure progress towards your goal? If yes, what kind of things have you tracked?**

Not really..

**Can you think of a recent day where you met your wellbeing goal? Was there anything you did different that day?**

Last Monday, I felt very focused and was able to complete all of my tasks on time. I didn't procrastinate or push off tasks I didn't want to do.

**Defined New Concept** - 10/12/2022, 9:54:21 PM

You defined a new concept (**Ending Work at 6pm**) by adding a tag.

**Difference in Outgoing Emails Today for Concept: Productive Focus Time** - 10/12/2022, 10:09:35 PM

I send more emails during productive focus time! If I can make more of my focus time productive and send more emails during working hours, I might able to avoid sending them after hours

**Difference in Outgoing Emails Today**

At times when Productive Focus Time was tagged, the average value of Outgoing Emails Today was **1.82** emails.

At times when Productive Focus Time was *not* tagged, the average value of Outgoing Emails Today was **1.49** emails.

That corresponds to a **20.79%** increase when Productive Focus Time was tagged.

Fig. 5. **Journal View.** (A) The goal specification questions used in our evaluation study, along with example responses; (B) a system-generated journal cell for defining a new concept, showing the menu for moving or deleting a cell; (C) a visualization insight that user has saved from the concept view, along with their description.

with a timestamp and the user’s description (Fig. 5C). Furthermore, the system automatically adds a cell to the journal whenever a new concept is defined **(DP-3)**.

## 5 EVALUATION STUDY

We used Pearl as a technology probe with 12 participants without data science expertise. We aimed to address the following research questions:

**RQ-1:** How did participants gain insights into their wellbeing while interacting with Pearl?

**RQ-2:** How did participants make use of machine assistance to reflect on their data and generate insights?

**RQ-3:** What kinds of data did users want to explore with Pearl?

In evaluating Pearl as a probe, our primary motivation was to initiate conversations with participants about the ways our system facilitated (or hindered) reflection and insight generation, as well as opportunities for future systems to better support their needs and desires. Our choice of these research questions reflects a desire to understand the process by which participants gain insights into their wellbeing through self-reflection on personal data and the utility of machine assistance at various stages in this process.

### 5.1 Participants

We recruited employees from a large technology company by sending emails asking for participation that targeted a diverse population of information workers. Our participation selection criteria required participants to (1) not be employed as data scientists nor have formal training in data analysis, and (2) be open to doing work that improves their wellbeing. We recruited 12 participants for participation in our study.

Among study participants (6 female, 6 male), 4 considered themselves data science novices, 4 reported basic experience, and 4 reported intermediate data science experience. None of our participants reported expertise in data science. Most of the participants were familiar with or used personal data collection devices such as fitness or sleep trackers. Our participants included different job roles (3 project managers, 3 people managers, 3 researchers, 1 engineer, and 2 with other non-specified roles). We compensated each participant with a $75 Amazon gift card. Participants in this study did not participate in the previous formative study. The study and the experimental instruments were approved by the Microsoft Research Institutional Review Board.

### 5.2 Study Protocol

Our study consisted of an initial 10-minute onboarding session to install the required logging software and a 90-minute main study session in which they interacted with Pearl.

*Onboarding.* In the onboarding session, participants installed the system's data logging software on their work devices. We also explained exactly which data source would be collected as part of the study and the granularity at which they would be aggregated, as well as anonymization and privacy-preserving features that are built into the logger. At the end of the session, participants completed a short questionnaire.

*Tutorial & Goal Specification Activity.* The main study session was scheduled to take place at least 10 business days after the onboarding session so that participants had sufficient data to explore and reflect on during the session. On average, participants collected 21.3 days (16 business days) worth of data at the time of the main study session. Participants did not have access to Pearl during this data collection phase. At the beginning of the session, we showed participants a 10-minute tutorial video for Pearl and walked them through each of its features using a synthetic dataset. After the tutorial, we asked participants to choose a workplace wellbeing goal and enter it into Pearl. To facilitate the process, we provided participants with a set of example goals that they could choose from or use as inspiration (Table 1). After

| **Example Goals** |
|---|
| I would like to improve my work-life balance.<br>I would like to feel more energetic at work.<br>I would like to have more flexibility over when and where I work.<br>I would like to be more productive at work.<br>I would like to be in greater control of the time I spend working. |
| **Participants' Goals** |
| I would like to be in greater control of the time I spend working. (2x)<br>I want to be more productive during business hours. (2x)<br>I would like to take advantage of the flexibility I'm provided to balance the time I spend working.<br>I want to have more productive meetings.<br>I want to be more focused at work.<br>I want to be more productive at work.<br>I would like to feel more prepared for the day/week ahead.<br>I want to be more efficient with my time during business hours.<br>I would like to have a better work-life balance.<br>I want to be more productive between meetings. |

Table 1. (Top) Example goals presented to participants at the beginning of our study. (Bottom) Participants' actual goals that they entered into Pearl and explored during our study.

choosing a goal, we instructed participants to go to the system's journal view to answer five goal-specification questions (shown in Fig. 5A). These questions were intended to help participants think critically about their goal and make the goal more specific, actionable, and decomposable. Participants wrote their responses in the journal, which persisted throughout the session.

*Guided Think-Aloud.* Next, participants thought aloud while interacting with Pearl as it presented their data. We structured the session in a way that balanced free exploration while nudging participants to perform each of the following activities:

(1) Visualize data sources on the calendar
(2) Define a concept with several tags
(3) Select data sources in the concept view
(4) Reflect on a system-generated insight
(5) Save an insight to the journal

We only nudged participants to complete a particular task when they asked what to do next or when time was running short. When participants had questions about the meaning of a data source, we encouraged them to read the tooltip text description and answered their question if they could not answer it themselves.

*Post-Study Feedback.* At the end of the session, participants were asked to return to their journal and reflect on their saved insights. Participants were asked the following reflection questions:

(1) What have you learned about your workplace wellbeing that you did not know before?
(2) How do you think that exploring your data has informed your wellbeing goal?
(3) Given what you have learned today, are there any actions you might take now to improve your wellbeing?

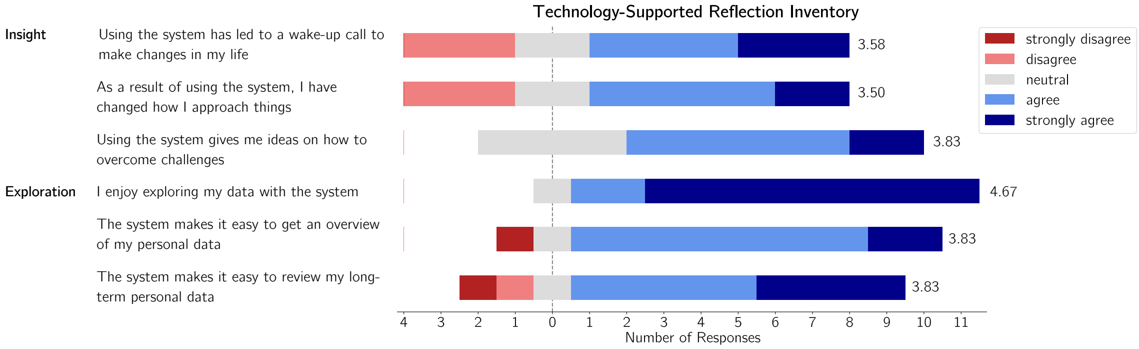


Fig. 6. Participant responses to the insight and exploration dimensions of the Technology-Support Reflection Inventory. Each question was scored on a 5-point Likert scale. Each horizontal bar contains responses for 12 participants, with average values displayed on the right.

The session ended with a post-study survey containing open-ended feedback questions and a subset of the Technology-Supported Reflection Inventory (TSRI) [10]. We included the insight dimension, which measures the degree to which a technology provides users with insights, and the exploration dimension, which measures the ease and enjoyment of exploring personal data with the system. We excluded the comparison dimension, as Pearl does not support reflection through social activity, and scored each question on a 5-point Likert scale.

## 6 RESULTS

In this section, we first report a quantitative analysis of our interaction logs and survey data to confirm that participants were successfully able to use Pearl to gain insights into their wellbeing. We then report our findings for each research question, informed by our quantitative data and a thematic analysis of our interview transcripts.

### 6.1 System Usage & Usability

*Log Analysis.* Our interaction logs confirm that participants successfully completed all five activities in our guided study protocol. Participants defined 2.25 concepts on average during their session, the majority of which (22/27) were hourly concepts. Two participants only defined daily concepts and only one participant defined both hourly and daily concepts. Participants entered 14.75 tags on average, corresponding to 7.08 average tags per concept. Participants saved an average of 3.17 insights to the journal, the majority of which (36/38) were difference in mean insights (Fig. 4F).

*Technology-Supported Reflection Inventory* Participants reported an average (median) score of 10.9 (11) out of 15 for the insight subscale (Fig. 6). This suggests that participants (on average) were able to gain insights into their wellbeing using Pearl. However, we note that 3 participants expressed disagreement with the first two items, indicating that the system could better support taking action on their wellbeing insights. Participants reported an average (median) score of 12.3 (13) out of 15 for the exploration subscale. In particular, almost all (11/12) participants enjoyed using the system to explore their personal data and most (10/12) found the system easy to use for these purposes.[9]

[9] We had one outlier participant who experienced several usability issues and expressed low motivation to act towards improving their wellbeing.

## 6.2 Insight Generation Process (RQ1)

All twelve participants successfully completed the five activities in our guided study protocol and our post-study reflection questions indicate that all participants were also able to articulate insights about their wellbeing by using Pearl. Although all participants successfully used the system to gain insights into their wellbeing, they differed in the depth and nature of their insights. In this section, we discuss which features of Pearl facilitated or hindered insight generation and the kinds of insights afforded by different interactions. We organize this section using the five activities in our guided study protocol (Section 5.2).

*Visual Data Exploration.* We observed two major types of insights participants gained through visual data exploration. The first type of insight occurred as participants were deciding which data sources to visualize. We observed participants decomposing broad concepts (e.g., meetings) into specific concepts (e.g., meetings with myself) in relation to why a particular data source was relevant to their goal: *"Meetings with myself are probably going to be more productive. But then I'd also, as a measure of my own accountability, I would want to look at application engagement during those time periods. [...] I want to know when I'm doing what—P07.* Oftentimes, this insight did not generate new knowledge, but was used to recall existing knowledge about a user's goal and their behaviors in a way that made it amenable to decomposition. For instance, as P08 clicked on the number of keypresses and remarked, *"Yeah, I was just curious because when I'm really focused, I'm using my keyboard a lot."* Participants frequently formed hunches or rough theories about why a particular data source might be relevant to their goal and tested these theories by visualizing the data source on their calendar. Other times, they left these hunches unquestioned.

Another class of insights occurred when participants visually explored their data sources on the calendar and were able to further interpret their behaviors. Nearly all participants used calendar events to provide external context that helped them interpret raw signal values, which also encouraged value judgments (e.g., distraction), often with respect to how well participants were meeting their goal: *"Yeah, I was feeling distracted at that time. I know because I was double booked here, triple booked there. [...] I was watching the instant messages for one meeting while I was trying to be active in the other and trying to straddle that line—P10."* While all participants accessed and viewed their calendars daily, some participants expressed that having the opportunity to reflect on their meeting load and cadence (particularly when overlaid with metrics such as computer activity) gave them the space to recognize salient patterns in their work activity. For some participants, the act of consciously reflecting on their calendars and interpreting their experience may have been a wellbeing intervention in itself [2].

During the visual exploration process, many participants scrutinized the data to verify that it matched their expectations before trusting that the data source could be used reliably to extract insights. Participants generated questions and tested whether the collected data conformed to their recollections. Upon examination of the raw values and patterns, many participants expressed that having quantitative confirmations of existing hunches was valuable, and some were led to honest realizations about their wellbeing: *"It's just sad. [...] And that just really goes to show the only time I have control over my time is when we have, almost contractually agreed moments where we're not going to bother each other so that we can eat—P09."* Scrutinizing data sometimes led participants to catch errors in the data[10]. These findings were disruptive, as these participants spent additional time trying to assess whether the data was accurate before trying to determine how it related to their goal.

[10] Originating from malfunction of the data collection logger.

*Defining Concepts & Entering Tags.* Participants differed in how they defined concepts related to their wellbeing and subsequently the kind of support they found relevant in doing so. Some participants were easily able to think of concepts that they wanted to teach the system about, and effortlessly entered tags into the system: *"I didn't have any meetings [...] So this time period I was really productive. I was getting a lot of editing done. So I would say I want to tag this time period [as productive]—P14."* Other participants struggled to define concepts due to missing data (e.g., location) or insufficient granularity (e.g., meetings with a particular person), and others expressed the desire to define concepts using a rule, which we could not support. Another group wanted to use more expressive teaching languages to provide information to the system, such as directly categorizing meetings (as opposed to regions of time), selecting meetings with keywords in the titles, or teaching the system to ignore personal meetings (e.g., focus time or to-do's) in its analysis.

For all participants, choosing a concept to tag involved translating their goal into concrete, measurable units that they thought were worth keeping track of. Thus, while defining a concept could require cognitive effort, the act of choosing a concept was a form of sensemaking and gave insight into their wellbeing goal: *"I'm just trying to remember which of these meetings kind of correlate with our tag and which don't. But [...] I can tag out other types of meetings as well. Like, these are customer calls. But I'm like, does that really get me any closer to understanding what I want to understand about my goal?—P09"* Participants made value judgements about their behavior when choosing and naming concepts and modified their concept definitions as they explored. For example, P04 caught themselves making these assumptions, renaming their concept from "Too Many Meetings" to "Meetings with Others" and remarking *"I wouldn't like to judge that I have too many meetings. I'd just say that I have meetings with others at this point of time and then see over a period of time if it is too many meetings or not."*

*Data Source Selection.* When scanning through the list of suggested data sources on the concept page, participants could be divided into two groups. The first group scanned through the list as a sanity check, mostly accepting the data sources the system had suggested at face value: *"I don't think the number of processes is kind of relevant because there could be other processes that's running that's not directly related to the work I do. [...] All right, incoming emails, average meeting durations... Kind of. Okay, mail today, attention events, maybe not so much. Everything else sounds okay to me—P02."* These participants were more interested in examining the visualizations in system-generated insight cards than in teaching the system which data sources they considered to be relevant or irrelevant. However, as discussed below, this did not mean that they were unbothered by Pearl selecting potentially irrelevant or several closely related data sources—instead, this group generally expected the relevant data sources to be accurate and distinct.

The second group systematically examined each data source and tried to explain why it was or was not relevant. While this group gained some insights through this act of reflection, they also faced challenges when interacting with Pearl. Sometimes the model's relevance scores did not align with these participants' own sense of relevance, leading to confusion. The practice of reflection also led some participants to challenge the notion of a data source being "relevant" to a particular concept. Some participants made distinctions between data sources that were causes or effects of their concept of interest, as this would change the kind of actions they would take in response. Others wanted direct information about the directionality of the relationship.

*System-Generated Insights* Compared to Pearl's other capabilities, system-generated difference-in-mean cards (Fig. 4F) most frequently led to insights that contain new knowledge for two reasons. First, the cards *aggregated* data with respect to a concept that was novel and relevant to the participant's goal. Second, the cards *compared* regions where the concept was tagged with regions where it was not tagged, directly facilitating participants comparing the aggregated values: *"That corresponds to a 342% increase with the multitasking tag. I mean, I also suspected this as well, but to see this* *in a card, I think that we all complain about how back-to-back meetings are so disruptive, but they're also driving disruptive behaviors for me, which is the multitasking. [...] Yeah, that speaks volumes—P09."*

When participants encountered system-generated information and visualizations about their concept of interest, it could either match their expectations or contradict them. When system-generated cards matched their expectations, participants were generally uncritical: *"This one is interesting, right? And it makes sense. It's almost so intuitive that I wouldn't necessarily put it in the journal. It's like, okay, yeah. If I have more meetings, I'm going to have more back-to-back meetings—P07."* When participants encountered data that contradicted their expectations, most would not scrutinize the system's output to assess its accuracy or relevance. Instead, participants would accept the statistics at face value and find alternative explanations for the data. For example, P03 wrote *"Surprised to find that there was a higher volume of incoming emails when I was accomplishing tasks. Surprising because I typically find myself distracted by emails,"* as they saved an insight to their journal, later adding that *"I wonder if sometimes my task accomplishment may be tied to those emails, either because I am sending out delegation of tasks and therefore people are responding and taking action."* Such reinterpretations were frequently perceived to be the most insightful by participants, even when they had reasons to question the system's accuracy. For example, P02 saved several email-related insights to the journal despite having previously encountered email data errors.

A system-generated statistic could allow multiple interpretations depending on whether the participant attributed a difference to their own actions or to the actions of others. For instance, P03 found that they received more incoming emails when working after hours, prompting them to reflect, *"maybe I'm the culprit because I am engaging, because I'm engaging after hours, so I'm getting a lot of incoming emails."* P03 could have easily been working after hours because they were receiving more emails, a hypothesis they left unquestioned.

*Journal Reflection.* We found that saving insights to the journal slowed down interactions and was a highly effective method in relating insights to their goal. For example, in relation to the goal of "I would like to take advantage of the flexibility I'm provided to balance the time I spend working," P02 learned about after-hours email behavior: *"The biggest insight that I had is I knew that I was working after hours, I knew that I was sending emails after hours. I just didn't know the volume of emails that I was sending. Right, because even though I have a disclaimer in my signature that, hey, I'm not expecting any response from anyone, but I'm sure people are noticing and it's probably not a good idea not setting the right expectation and example—P02."* In our post-study reflection activity, participants used the insights saved in their journal to reflect on their learnings and how they might change their behavior to make progress towards their goal. For example, towards the goal of "I want to be more productive during business hours", P06 decided to start re-allocating their work: *"So one of the things I learned is the last one, the meetings vs. non-meetings day. [...] The day with non meetings, I'll start pushing some of that work towards there—P06."* Participants also provided several high-level insights and new intentions during this activity that they had not mentioned previously.

### 6.3 Role of Machine Assistance (RQ2)

In comparing and contrasting participants' interactions with the various forms of machine assistance Pearl supports, we identify two roles machine assistance plays in the self-reflection process: one as an accelerator and one as a decelerator.

*Accelerator.* Machine assistance accelerated both the visual data exploration and the self-reflection process by mechanizing a task that would otherwise have been tedious. Suggesting relevant data sources allowed participants to quickly find data sources to visualize without having to sift through a long list. Automatically proposing insights bootstrapped participants into higher levels of reflection [27] by providing them with visualizations and statistics that naturally encouraged them to explore relationships between data sources and concepts. These capabilities reduced cognitive effort for participants and decreased the amount of time needed to arrive at an insight into their wellbeing. However, this acceleration came at the cost of participant scrutiny in evaluating whether the system's automations were accurate and factual. As discussed in the previous section, participants generally did not question system-generated insights and reinterpreted them in a way that made intuitive sense to them without deeply considering alternative interpretations. Thus, designers of future intelligent wellbeing support system should exercise caution when accelerating the reflection process—faulty machine inferences and suggestions can lead to flawed and unscrutinized wellbeing insights. In the best case, such faulty machine support could erode trust in the system over time; in the worst case, it could lead to counterproductive behaviors and mindsets.

*Decelerator.* Pearl required participants to define a goal, define relevant concepts, enter tags for those concept in order to gain insights from the system. Participants could not receive system insights just by exploring data sources as they pleased. In requiring participants to conform to this particular interaction, the system decelerated the process of visual data exploration because the act of translating a goal into concepts that can be tagged on the calendar inherently requires cognitive effort. In fact, it was the only activity that occasionally required the study facilitators to remind them of their goal specification questions in the journal. However, translating goals into concepts also carried wellbeing-related benefits because participants had to think critically about their goal and what they cared about measuring. Our notion of deceleration is reminiscent of *slow technology* [31], a design strategy in which slowness is deliberately used to create necessary space for reflection and sensemaking [7].

In order for deceleration to be effective, designers must be mindful that cognitive effort is both beneficial for reflection and inherently aversive [40]—too much deceleration may result in an interaction that is too difficult and effortful to be useful for reflection. One design strategy is to provide immediate feedback in response to user effort, which can help users recognize wellbeing benefits and sustain effort over time. For instance, Pearl automatically notifies users of new insights in response to entering concept tags.

### 6.4 Insightful Data Sources (RQ3)

Participants explored a wide variety of data sources while interacting with Pearl. Across all participants, nearly all the available data sources were visualized on the calendar. Among the insights that participants saved to the journal, the most popular data sources were attention events, keypresses, incoming emails, and outgoing emails. Aggregated by data source category, 8 of the saved insights corresponded to calendar data sources, 11 to email data sources, and 17 to computer usage data sources. The most requested data sources that our system did not support were instant messaging (e.g., Slack, Microsoft Teams), video conferencing, and incoming notifications.

Somewhat paradoxically, participants expressed the most confusion about attention events despite it being the data source with the most saved insights. Due to privacy and permissions issues, as well as implementation challenges, our system did not measure the amount of time participants spent in individual applications nor the number of times participants switched between applications. Instead, the data logger reported aggregate metrics about user "engagement" (the total amount of time spent in all applications in a 30-minute window) and "attention" (the total number of attention signals in a 30-minute window, such as keypresses, mouse clicks, or window switches). This data source was confusing for participants because they did not understand exactly what the data source was measuring and the description was complex. However, this data source was frequently considered insightful because participants wanted to know when they were actively working on their computer, which they most often mapped to attention events or keypresses. In hindsight, this technical constraint violated our design principles, as it limited participants' agency in determining what "engagement" or "attention" meant in their own terms. Depending on the nature of each participant's work, events such as keypresses or window switching may not correspond to attention, nor would time spent in a particular application correspond to engagement.

Participants also expressed frustration about the granularity of the system's data sources. Sources such as attention events and application engagement were too coarse, while data sources such as the number of processes were too fine. Instead, we found that participants desired data sources at the granularity with which they interact with a user interface. For example, data sources such as keypresses and mouse clicks were easy to understand, easy to interpret, and highly popular. Similarly, many participants wanted access to the amount of time spent on individual applications and expressed frustration with our aggregated metrics. Thus, while participants were highly interested in abstract concepts such as attention and engagement, we recommend that future systems allow users to define these concepts using interpretable, low-level signals at the level of individual interactions.

## 7 LIMITATIONS

Designing Pearl as a technology probe and conducting our evaluation in a single session bounded the scope of the questions we were able to ask. We were unable to study the experience of participants using Pearl over an extended period of time, nor its effect on their long-term behavior and wellbeing. We also note that our system suffered from data-collection errors in email data for some participants, which affected their insight generation process.

Since all of our participants were employees at a large technology company, we note that most are likely exposed to data summaries, statistics, and visualizations in their day-to-day work activities. Thus, even self-reported data science novices likely possessed some basic data literacy, although this is not atypical of information workers. We did not observe noticeable differences in participants' ability to generate insights based on their self-reported data science experience.

Despite the simplicity of our interactive ML system, we found that participants produced meaningful insights into their wellbeing and wellbeing goals. However, our study design prevented us from determining whether these insights truly reflected their actual work habits or if they merely believed these insights to be true. Future work is needed to assess whether participants would receive greater insight from a more sophisticated machine learning system or if simple systems are not only sufficient but a design strategy for encouraging reflection. Similarly, our data presentations used standard approaches in information visualization—future work could help clarify whether novel visualization strategies could assist users in appropriately scrutinizing machine-generated insights.

The data sources supported by Pearl constrain the kinds of work activities that our system can capture to those that are bound to a computer. Extensions to Pearl that include data sources from wearables, mobile devices, or environmental sensors could feasibly expand the kinds of activities that can be captured. We also believe that our findings may be applicable beyond workplace wellbeing support, although future studies are needed to confirm the extent to which our findings generalize to other domains (e.g., personal health, fitness).

Moreover, our focus on information workers and passively sensed data determines the kinds of workers our system is able to support. Specifically, our system is able to support workers who can collect personal data easily without the risk of privacy violations and who have the agency to make changes to their work habits. In light of growing concerns over workplace surveillance [6, 36], we encourage designers of future workplace wellbeing support system to be mindful of the sociotechnical context they are designing for and which forms of support are most appropriate for their population. Designers should be cautious that the data collection risks posed by systems such as Pearl might outweigh its potential for benefit in certain populations, particularly for historically marginalized groups that are differentially impacted by surveillance technology. Simultaneously (as mentioned in the following section), having objective documentation of otherwise invisible work activities can also be a tool for self-advocacy and worker empowerment. Lastly, we mentioned that our work is culturally specific to information workers at technology companies in the United States, and further work is required to assess the extent to which our findings generalize across different industries, cultures, and regions of the world.

## 8 EXPANDING THE DESIGN SPACE

An important role of the probe is to inspire new ideas about future technologies. In this section, we present several dimensions of the design space for intelligent reflection support systems that emerged from our conversations with participants. Each of these dimensions points to opportunities for future work.

*Past, Present, and Future.* PEARL deliberately supports reflection on past data only. Participants expressed interest in using such a system over time for self-tracking, planning, and goal-setting. The continuous use of a wellbeing support system would allow users to reflect on their data more frequently, examine important trends over time, and assess their progress toward their goal. Participants also expressed a number of opportunities for machine assistance in this process. A system could use concept models to display tag predictions on past data to reduce labeling effort or to display tag predictions for future data to aid in planning [15]. Concept models could also be actively learned by querying the user for labels at opportune moments [37]. Concept model predictions could be used as triggers for experience sampling prompts [62], e.g., asking the user to rate their mood when they leave the office.

*Closing the Loop: From Reflection to Action.* While PEARL can provide insights into a user's wellbeing, making progress towards wellbeing goals ultimately involves the user changing their behavior and taking action [42]. The insights provided by PEARL may provide clues as to which actions a user can take to enact change, but do not provide concrete suggestions. Future intelligent wellbeing support tools could better support users in translating insights into tangible action plans. These action suggestions might be built into the system (e.g., incorporating behavior change interventions designed by experts), crowdsourced from users with similar goals, or machine-generated using large language models. Once a user decides to take action, the system could provide support for monitoring progress or intelligently nudging a user back on track when they deviate from the plan. When a user is unsure which of several actions are effective, the system could support and provide guidance for effective self-experimentation [22].

*Teaching Languages.* PEARL supports interactive machine teaching through direct manipulation and labeling interactions on the calendar. Participants expressed interest in a more expressive teaching language and teaching interactions. For example, the use of natural language could enable chatbot-based conversations about goals [39] or open-ended visualization [50]. While interviews with experts in our formative study indicated that label-based concept definitions were more promising than rule-based ones, we found that most participants in our evaluation study expressed a desire to define rule-based concepts—in alignment with remarks from [51]. Experts in our formative study correctly identified that many high-level wellbeing concepts (e.g., productivity or stress) may not admit rule-based definitions. However, participants desired rule-based concepts (e.g., "the number of messages I received from my boss while I'm in a meeting") as a cognitive scaffold for data exploration and as precursors to higher-level concepts.

One might also imagine incorporating intelligent wellbeing support functionality directly into a user's calendar application, or other applications such as email or instant messaging. Participants reported many different uses of their calendars, such as color-coding meetings into categories, blocking focus time by scheduling self-meetings, or adding calendar items as a to-do list. Integrating interactive machine teaching functionality directly into the user's workplace tool could enable machine learning systems to utilize user's existing actions as implicit supervision.

*Social Sharing.* Many participants expressed a desire to share and compare their wellbeing data with others. Surprisingly, three participants wanted to share the data with their managers to initiate conversations about their work and meeting load. Two participants in management roles reflected on whether they were being "good examples" for the employees on their team, while one discussed wanting to see their employees' data so that they could better support them. Two wanted to compare averages or baselines for different data sources to assess whether their data was in "normal" ranges or compare relevant data sources and insights with people who had similar goals. While social dimensions of wellbeing support can bring clear benefits [8], we caution that implementing such features responsibly is a challenging design problem in light of workplace surveillance and privacy violations [6].

*Navigating Privacy Trade-offs.* Particularly among people with limited data literacy, users of data-driven wellbeing support systems may not be aware of the privacy risks associated with data collection. Future wellbeing support systems need to support users in making more informed privacy decisions. For example, systems could clearly communicate the benefits and risks of logging a particular data source and provide granular settings to turn data collection on and off. We also recognize the potential for interactive machine teaching systems to help illuminate which data sources are predictive of particular behaviors, which may help educate end-users about their privacy trade-offs. We also caution that this may also coerce users into collecting or sharing more data than they would have otherwise, and encourage designers to thoughtfully navigate these tensions.

## 9 CONCLUSION

In this work, we investigated the potential for reflection support systems to help people form connections between personal data and their wellbeing goals using interactive machine teaching. Through a formative study, we developed a set of design principles that informed the design of Pearl, a technology probe for machine-assisted reflection on personal data. Our evaluation shows that participants without data science expertise were able to use Pearl to gain insights into their workplace wellbeing and in our analysis, we studied the ways in which machine assistance fostered self-reflection. Finally, we contribute a set of design principles for intelligent reflection support systems that serve as inspiration for future work.

## REFERENCES

[1] 2021. Workplace Stress. https://www.stress.org/workplace-stress

[2] Charles Abraham and Susan Michie. 2008. A taxonomy of behavior change techniques used in interventions. *Health psychology* 27, 3 (2008), 379.

[3] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13.

[5] Bon Adriel Aseniero, Charles Perin, Wesley Willett, Anthony Tang, and Sheelagh Carpendale. 2020. Activity river: Visualizing planned and logged personal activities for reflection. In *Proceedings of the International Conference on Advanced Visual Interfaces*. 1–9.

[6] Kirstie Ball. 2010. Workplace surveillance: An overview. *Labor History* 51, 1 (2010), 87–106.

[7] Eric PS Baumer. 2015. Reflective informatics: conceptual dimensions for designing technologies of reflection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 585–594.

[8] Eric PS Baumer, Sherri Jean Katz, Jill E Freeman, Phil Adams, Amy L Gonzales, John Pollak, Daniela Retelny, Jeff Niederdeppe, Christine M Olson, and Geri K Gay. 2012. Prescriptive persuasion and open-ended social awareness: expanding the design space of mobile health. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 475–484.

[9] Eric PS Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing reflection: on the use of reflection in interactive system design. In *Proceedings of the 2014 conference on Designing interactive systems*. 93–102.

[10] Marit Bentvelzen, Jasmin Niess, Mikołaj P Woźniak, and Paweł W Woźniak. 2021. The Development and Validation of the Technology-Supported Reflection Inventory. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–8.

[11] Marit Bentvelzen, Paweł W Woźniak, Pia SF Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting Reflection in HCI: Four Design Resources for Technologies that Support Reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27.

[12] Kirsten Boehner, William Gaver, and Andy Boucher. 2012. 14 Probes. *Inventive Methods: The happening of the social* 185 (2012).

[13] Kirsten Boehner, Janet Vertesi, Phoebe Sengers, and Paul Dourish. 2007. How HCI interprets the probes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1077–1086.

[14] Jenna Butler and Sonia Jaffe. 2021. Challenges and gratitude: A diary study of software engineers working from home during covid-19 pandemic. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 362–363.

[15] Eun Kyoung Choe, Saeed Abdullah, Mashfiqui Rabbi, Edison Thomaz, Daniel A Epstein, Felicia Cordeiro, Matthew Kay, Gregory D Abowd, Tanzeem Choudhury, James Fogarty, et al. 2017. Semi-automated tracking: a balanced approach for self-monitoring applications. *IEEE Pervasive Computing* 16, 1 (2017), 74–84.

[16] Eun Kyoung Choe, Bongshin Lee, et al. 2015. Characterizing visualization insights from quantified selfers' personal data presentations. *IEEE computer graphics and applications* 35, 4 (2015), 28–37.

[17] Eun Kyoung Choe, Bongshin Lee, Matthew Kay, Wanda Pratt, and Julie A Kientz. 2015. SleepTight: low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 121–132.

[18] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding self-reflection: how people reflect on personal data through visual data exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 173–182.

[19] Thomas W Colligan and Eileen M Higgins. 2006. Workplace stress: Etiology and consequences. *Journal of workplace behavioral health* 21, 2 (2006), 89–97.

[20] Victor P Cornet and Richard J Holden. 2018. Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of biomedical informatics* 77 (2018), 120–132.

[21] Vedant Das Swain, Victor Chen, Shrija Mishra, Stephen M Mattingly, Gregory D Abowd, and Munmun De Choudhury. 2022. Semantic Gap in Predicting Mental Wellbeing through Passive Sensing. In *CHI Conference on Human Factors in Computing Systems*. 1–16.

[22] Nediyana Daskalova, Eindra Kyi, Kevin Ouyang, Arthur Borem, Sally Chen, Sung Hyun Park, Nicole Nugent, and Jeff Huang. 2021. Self-E: Smartphone-Supported Guidance for Customizable Self-Experimentation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 227, 13 pages. https://doi.org/10.1145/3411764.3445100

[23] Chris Elsden, Abigail C Durrant, and David S Kirk. 2016. It's Just My History Isn't It? Understanding smart journaling practices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2819–2831.

[24] Chris Elsden, David S Kirk, and Abigail C Durrant. 2016. A quantified past: Toward design for remembering with personal informatics. *Human–Computer Interaction* 31, 6 (2016), 518–557.

[25] Daniel A Epstein. 2015. Personal informatics in everyday life. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. 429–434.

[26] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. 39–45.

[27] Rowanne Fleck. 2012. Rating reflection on experience: A case study of teachers' and tutors' reflection around images. *Interacting with computers* 24, 6 (2012), 439–449.

[28] Bill Gaver, Tony Dunne, and Elena Pacenti. 1999. Design: cultural probes. *interactions* 6, 1 (1999), 21–29.

[29] Lars Grammel, Melanie Tory, and Margaret-Anne Storey. 2010. How information visualization novices construct visualizations. *IEEE transactions on visualization and computer graphics* 16, 6 (2010), 943–952.

[30] Hayley Guillou, Kevin Chow, Thomas Fritz, and Joanna McGrenere. 2020. Is your time well spent? reflecting on knowledge work more holistically. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.

[31] Lars Hallnäs and Johan Redström. 2001. Slow technology–designing for reflection. *Personal and ubiquitous computing* 5, 3 (2001), 201–212.

[32] Kristina Höök. 2000. Steps to take before intelligent user interfaces become real. *Interacting with computers* 12, 4 (2000), 409–426.

[33] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[34] Dandan Huang, Melanie Tory, and Lyn Bartram. 2016. A Field Study of On-Calendar Visualizations. In *Proceedings of the 42nd Graphics Interface Conference*. 13–20.

[35] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24.

[36] Jodi Kantor and Arya Sundaram. 2022. The Rise of the Worker Productivity Score. https://www.nytimes.com/interactive/2022/08/14/business/worker-productivity-tracking.html

[37] 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.

[38] Harmanpreet Kaur, Daniel McDuff, Alex C Williams, Jaime Teevan, and Shamsi T Iqbal. 2022. “I Didn’t Know I Looked Angry”: Characterizing Observed Emotion and Reported Affect at Work. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[39] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for workplace reflection: a chat and voice-based conversational agent. In *Proceedings of the 2018 designing interactive systems conference*. 881–894.

[40] Wouter Kool and Matthew Botvinick. 2018. Mental labour. *Nature human behaviour* 2, 12 (2018), 899–908.

[41] Carol Collier Kuhlthau. 1999. The role of experience in the information search process of an early career information worker: Perceptions of uncertainty, complexity, construction, and sources. *Journal of the American Society for information Science* 50, 5 (1999), 399–412.

[42] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 557–566.

[43] Ian Li, Anind K Dey, and Jodi Forlizzi. 2011. Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th international conference on Ubiquitous computing*. 405–414.

[44] Madelene Lindström, Anna Ståhl, Kristina Höök, Petra Sundström, Jarmo Laaksolathi, Marco Combetto, Alex Taylor, and Roberto Bresin. 2006. Affective diary: designing for bodily expressiveness and self-reflection. In *CHI’06 extended abstracts on Human factors in computing systems*. 1037–1042.

[45] Stephen M Mattingly et al. 2019. The Tesserae Project: Large-Scale, Longitudinal, In Situ, Multimodal Sensing of Information Workers. In *CHI Ext. Abstracts*.

[46] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. 2012. AffectAura: an intelligent system for emotional memory. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 849–858.

[47] Andre N Meyer, Gail C Murphy, Thomas Zimmermann, and Thomas Fritz. 2017. Design recommendations for self-monitoring in the workplace: Studies in software development. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–24.

[48] Kiran Mirchandani. 1998. Protecting the boundary: Teleworker insights on the expansive concept of “work”. *Gender & Society* 12, 2 (1998), 168–187.

[49] William C Murray and Adam Rostis. 2007. Who’s running the machine? A theoretical exploration of work stress and burnout of technologically tethered workers. *Journal of individual employment rights* 12, 3 (2007), 249–263.

[50] Arpit Narechania, Arjun Srinivasan, and John Stasko. 2020. NL4DV: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 369–379.

[51] Felicia Ng, Jina Suh, and Gonzalo Ramos. 2020. Understanding and supporting knowledge decomposition for machine teaching. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1183–1194.

[52] Donald A. Norman. 1994. How Might People Interact with Agents. *Commun. ACM* 37, 7 (jul 1994), 68–71. https://doi.org/10.1145/176789.176796

[53] Afarin Pirzadeh, Li He, and Erik Stolterman. 2013. Personal informatics and reflection: a critical examination of the nature of reflection. In *CHI’13 Extended Abstracts on Human Factors in Computing Systems*. 1979–1988.

[54] Bernd Ploderer, Wolfgang Reitberger, Harri Oinas-Kukkonen, and Julia van Gemert-Pijnen. 2014. Social interaction and reflection for behaviour change. , 1667–1676 pages.

[55] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. 2011. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proceedings of the 13th international conference on Ubiquitous computing*. 385–394.

[56] Gonzalo Ramos, Chris Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. 2020. Interactive machine teaching: a human-centered approach to building machine-learned models. *Human–Computer Interaction* (April 2020). https://www.microsoft.com/en-us/research/publication/interactive-machine-teaching-a-human-centered-approach-to-building-machine-learned-models/

[57] Amon Rapp and Federica Cena. 2016. Personal informatics for everyday life: How users without prior self-tracking experience engage with personal data. *International Journal of Human-Computer Studies* 94 (2016), 1–17.

[58] Patrice Y Simard, Saleema Amershi, David M Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, et al. 2017. Machine teaching: A new paradigm for building machine learning systems. *arXiv preprint arXiv:1707.06742* (2017).

[59] Petr Slovák, Christopher Frauenberger, and Geraldine Fitzpatrick. 2017. Reflective practicum: A framework of sensitising concepts to design for transformative reflection. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2696–2707.

[60] Poorna Talkad Sukumar, Anind Dey, Gloria Mark, Ronald Metoyer, and Aaron Striegel. 2021. Triggers and Barriers to Insight Generation in Personal Visualizations. In *Graphics Interface 2022*.

[61] Alice Thudt, Dominikus Baur, Samuel Huron, and Sheelagh Carpendale. 2015. Visual mementos: Reflecting memories with personal data. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 369–378.

[62] 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.

[63] Ariane G Wepfer, Tammy D Allen, Rebecca Brauchli, Gregor J Jenny, and Georg F Bauer. 2018. Work-life boundaries and well-being: Does work-to-life integration impair well-being through lack of recovery? *Journal of Business and Psychology* 33, 6 (2018), 727–740.