Dealing with Information Overload in Multifaceted Personal Informatics Systems

Simon L. Jones[1], Ryan Kelly[1]

[1]University of Bath, UK

Authors' Mini-bios:

Simon L. Jones (s.l.jones@bath.ac.uk, http://go.bath.ac.uk/simonjones) is a computer scientist with an interest in personal informatics, data mining and information visualisation; he is a Lecturer in Human-Computer Interaction in the Department of Computer Science of the University of Bath.

Ryan Kelly (r.m.kelly@bath.ac.uk, http://go.bath.ac.uk/rmkelly) is a researcher in Human-Computer Interaction with an interest in visual representations of individual and collaborative action in digital technologies; he is a Research Associate in Human-Computer Interaction in the Department of Computer Science of the University of Bath.

Dealing with Information Overload in Multifaceted Personal Informatics Systems

Running Head: INFORMATION OVERLOAD IN PERSONAL INFORMATICS SYSTEMS

ABSTRACT

Personal informatics systems are tools that capture, aggregate and analyse data from distinct facets of their users' lives. This paper adopts a mixed methods approach to understand the problem of information overload in personal informatics systems. We report findings from a three-month study in which twenty participants collected multifaceted personal tracking data and used a system called 'Exist' to reveal statistical correlations within their data. We explore the challenges that participants faced in reviewing the information presented by Exist, and we identify characteristics that exemplify "interesting" correlations. Based on these findings, we develop automated filtering mechanisms that aim to prevent information overload and support users in extracting interesting insights. Our approach deals with information overload by reducing the number of correlations shown to users by ~55% on average, and increases the percentage of displayed correlations rated as interesting to ~81%, representing a 34 percentage point improvement over filters that only consider statistical significance at $p<0.05$. We demonstrate how this curation can be achieved using objective data harvested by the system, including the use of Google Trends data as a proxy for subjective user interest.

CONTENTS

1. INTRODUCTION

With many personal tracking technologies becoming mainstream consumer products or services, millions of people are now engaging in the collection of data about their lives. Personal informatics (PI) systems, which allow users to explore and analyse this data, have been shown to provide value in a variety of life settings, from supporting reminiscence (Cosley *et al.*, 2009; Elsden, Kirk & Durrant, 2015) to managing chronic medical conditions (Huckvale, Car, Morrison & Car, 2012; Karkar *et al.*, 2015). Typically, systems for collecting and analysing personal tracking data focus on a particular life *facet*, a single aspect of a person's life (e.g. health, wellbeing, or productivity), within which a user may have general interests or specific goals for understanding, monitoring or changing their behaviour. Recent studies have demonstrated that systems which extend personal tracking across multiple facets can lead to "holistic engagement with one's life" (Li, Forlizzi & Dey, 2010) and may foster sustained engagement with tracking technologies (Bentley *et al.*, 2013). These *multifaceted* systems are capable of providing insights into the associations between distinct aspects of a person's behaviour and activities (Bentley *et al.*, 2013). For example, combining data from physical activity trackers with self-reported measures of daily mood might expose the relationship between exercise and a person's mental wellbeing.

A growing number of mainstream personal informatics systems are now adopting a multifaceted configuration, capitalising on increased public interest in aggregating and analysing personal tracking data from diverse sources. Popular examples include Exist[1], TicTrac[2], Zenobase[3], and Gyroscope[4]. Each of these systems uses statistical analysis to explore the relationships among different sources of personal data, presenting results back to users in the form of data visualisations (e.g. scatter plots or line graphs), or natural language statements summarising statistical correlations, e.g. "you get more sleep when you do more exercise". The main benefits of such systems are twofold; first, they simplify the management of data by processing it within a single repository, and second, they automate the analysis of data to provide users with holistic insights that they could not easily derive themselves.

We suggest that the increasing ease with which diverse data can be aggregated and analysed by multifaceted PI systems facilitates exploratory use, whereby users do not necessarily collect data to address pre-defined goals or questions. Rather, these systems reduce the effort required to integrate and analyse any available data, allowing users to see if the system's output yields valuable and meaningful insights. While the benefits of

[1] https://exist.io
[2] https://tictrac.com
[3] https://zenobase.com
[4] https://gyrosco.pe

multifaceted PI systems have been widely discussed in academic literature (e.g. Li *et al.*, 2010; Bentley *et al.*, 2013; Rooksby, Rost, Morrison & Chalmers, 2014), little work has focused on identifying and addressing problems that users face when using multifaceted PI systems in this way, and exploring the outputs that they produce.

We contend that there are two unexplored issues that arise as a result of the diversity and quantity of data that users can provide to multifaceted PI systems. First, users may find it difficult to review the output of PI systems due to the sheer volume of correlations presented, potentially giving rise to problems associated with information overload. As an example, a tool that explores pairwise correlations among 20 variables has the potential to report up to 190 relationships (Jones, 2015). We argue that making sense of many novel and potentially unanticipated observations is likely to require significant cognitive effort. Thus, there is a need to understand how we can support users in making sense of this data and how we can help them to easily identify information that is of value, given the context of exploratory use, i.e. that which is driven by the goal of procuring interesting outputs rather than specific relationships.

This gives rise to a second issue in that it is not clear what outputs users deem to be either valuable or interesting, particularly in the context of exploratory use. Previous studies have shown that some results of automated analysis by PI systems are considered more useful than others, and that 'obvious' observations offer little value to users (Bentley *et al.*, 2013; Jones, 2015). Bentley *et al.* (2013) describe an 'obvious' observation as one which "simply made sense with [users'] lives, whether they had previously considered it or not", e.g. "being happier on weekends". They recommended that systems in this domain should allow users to hide observations that are 'too obvious', and automatically prioritise information that users might find less obvious.

A common approach to the automatic prioritisation or filtering of information is to use machine learning (Witten & Frank, 2005), which is prevalent in applications such as recommender systems (Ricci, Rokach & Shapira, 2011). However, automatically filtering results which are of most interest to the user remains an open challenge in PI systems. With a high number of potential outputs and a context that lacks specific goals, it may not be easy to determine what users want to see. We argue that addressing this challenge is increasingly important given the rise of personal tracking as a mainstream activity. Hence, this paper aims to support the task of presenting users with interesting insights by developing a technical intervention that automatically filters information on the basis of its *Interestingness* to the user, thus mitigating the likelihood of information overload and improving user experience with PI systems.

1.1. Overview of the Paper

This paper adopts a mixed methods approach to understand and alleviate the problem of information overload in a multifaceted personal informatics system via automated filtering on the basis of user interest. We seek to address the following research questions:

(RQ1) What do users find interesting within the correlational outputs from their personal data, given the context of exploratory use?

(RQ2) How can we algorithmically curate interesting insights and alleviate information overload for users?

We present findings from a 3-month study of Exist, a commercial tool designed for the aggregation and analysis of multifaceted personal data. Participants in the study provided Exist with diverse personal data, comprising daily measures of physical activity, sleep, productivity and distraction, mood, calendar events, social media interactions, music listening and local weather conditions.

Qualitative analyses are used to investigate participants' experiences with the Exist output. We find that, when viewing their personal data for the first time, participants encounter a range of issues that we interpret as related to information overload. We consider how such overload might be alleviated by filtering information outputs on the basis of interestingness, and participants' comments allow us to identify six features that initially characterise "interesting" outputs. These include the extent to which correlations are surprising, easy to deduce, unique, matched to a user's expectations, supportive of practical action, or pertinent to the individual's aims in understanding their life.

Quantitative approaches are used to investigate users' subjective ratings of the Exist output, accounting for dimensions that include how interesting, accurate, novel, stable, surprising, unique, useful and positively/negatively valenced the insights are. We find that participants were more likely to rate correlations as interesting when they were seen to be surprising or useful; when they presented between- rather than within- data category relationships (i.e were multifaceted rather than unifaceted); when they were not associated with uninteresting data categories (e.g. weather); or when they exhibited low p-values (high confidence). These findings support our use of supervised machine learning to automate the filtering of outputs on the basis of interestingness, thereby reducing overload on the user. Our approach reduces the number of correlations shown to users by ~55% on average, and increases the percentage of interesting correlations within this output to ~81% (which represents a 34 percentage point improvement over filters that only consider statistical significance at $p<0.05$). In lieu of subjective data from users, we show how such curation can be achieved by drawing on measures of 'Interestingness' from data mining research and contextual data from external sources, including search term popularity from Google Trends.

This paper aims to build on previous work which has indicated the prospective value of, and desire for, multifaceted personal informatics systems (e.g. Bentley *et al.*, 2013), but which has not explored the aforementioned challenges associated with their use in great depth. The contributions of this paper include new insights into the challenges of sensemaking and information overload in personal informatics systems; an understanding of the value that multifaceted systems provide, including what constitutes interesting information in this context; and objective data that can be used to drive filtering mechanisms that may prevent information overload in future PI systems.

2. BACKGROUND

In this section we discuss previous research in personal informatics in order to situate our contributions. We also explore information overload in the field of Human-Computer

Interaction, with a view to understanding its relevance as a potential design problem for multifaceted PI systems. We conclude by examining the use of information filtering as a mechanism for dealing with information overload.

2.1. Personal Informatics Systems

Recent technological advances have made it easier to engage in digitised tracking of biological, physical, behavioural, and environmental data (Swan, 2013). Millions of people are now equipped with tracking technologies via smartphones with integrated sensors, and which function as signal receivers for a growing range of data-logging devices. A growing body of research in the area of human activity recognition is continuously advancing our ability to recognise and quantify many everyday activities across many facets of life (Lara & Labrador, 2013), from general fitness and lifestyle (McGrath & Scanaill, 2013) to acute healthcare and rehabilitation (Patel, Park, Bonato, Chan & Rodgers, 2012).

Access to automatically captured behaviour data increases the potential for latent patterns and signals to be discovered and used to inform future actions (Shah, 2015). Personal Informatics systems aim to support users in exploring this data for gaining self knowledge and enhancing self reflection (Li, Forlizzi & Dey, 2010). While the notion of leading a "data-driven life" (Wolf, 2010) has often been associated with the specialist demographic of "quantified selfers" (Choe *et al.*, 2014), the use of PI systems is now growing amongst mainstream technology users. Choe *et al.* (2014) reported that the most common motivation for self-tracking is to improve health and other aspects of life (e.g. to find a balanced lifestyle, or cure or manage a health condition).

Li, Dey & Forlizzi (2010) presented a model of the stages involved in the use of PI systems. In the first stage, *preparation*, individuals establish which data they require and identify suitable tools to support the data *collection* stage. The *integration* stage then involves combining and transforming data such that it can be processed in the subsequent *reflection* stage. During reflection, people explore, interpret and consider insights provided by the data. This underpins the *action* stage, in which people take courses of action based on the knowledge they have gained. This model has been used to support identification and categorisation of problems that users experience when interacting with PI systems. For example, the stage of *reflection* can be hampered by time constraints and poor visual representation of data (Tollmar, Bentley & Viedma, 2012; Cuttone, Petersen & Larsen, 2014; Chung, Cook, Bales, Zia & Munson, 2015).

Li *et al.* (2011) identified six categories of question to which people seek answers from their data during *reflection*, and suggested that systems should tailor their features to support users for each category accordingly. These categories include: *goals*, *status*, and *discrepancies* (e.g. how much physical activity should I be doing today, have I currently done enough, and how much more should I do to meet my goal?); *history* and *trends* (e.g. how much physical activity did I do last month, and how has this amount changed over time?); as well as *context* and *factors* (e.g. in what circumstances did I do the most physical activity, and what factors affect my tendency to do physical activity?).

Questions relating to *context* and *factors* logically require a multifaceted approach to tracking, whereby supplementary streams of data can provide contextual information and insight into factors that affect, or are affected by, certain behaviours (Li *et al.*, 2011). A number of researchers have built systems that correlate distinct sources of personal data, enabling users to obtain meaningful and practical inferences from the data. For example, Kay *et al.* (2012) developed the Lullaby system to capture environmental information, e.g. sound, light and temperature, to help users understand the factors affecting the quality of their sleep. Li (2011) used contextual information to improve awareness of factors affecting physical activity, and Bentley *et al.* (2013) built the *Health Mashups* system to identify connections between health and wellbeing factors.

It is commonly suggested within the quantified self research community that tracking is most interesting and useful when it takes a holistic view of a user's life (e.g. Dingler, Sahami & Henze, 2014). Haddadi & Brown (2014) suggest that the ability to relate data across different facets is likely to result in more "appealing inferences" for users, and increase engagement in collecting and using personal data. Bentley *et al.* (2013) showed that revealing correlations between wellbeing factors did indeed result in increased user engagement in their Health Mashups system. The insights generated were highly variable between individuals, but complex relationships were easily understandable when represented as natural language statements.

Despite the benefits of a multifaceted approach to tracking, Li *et al.* (2012) identified that, typically, mainstream personal informatics tools are uni-faceted, functioning as stand-alone trackers of atomic behaviours. Users of uni-faceted systems are faced with the task of piecing together fragmented information from disparate sources if they wish to understand the effects of different factors on their behaviour (Li *et al.*, 2010). Multifaceted data aggregation and analytics systems eliminate the need to hold information in memory to make comparisons and perform multivariate analysis. Variables can be viewed in the context of others, and relationships, such as correlations, between distinct aspects can be uncovered within a single system.

Although data aggregation and analysis tools have been explored in previous research, there remain a number of open research questions and design challenges associated with this type of multifaceted PI tool. For example, several researchers question the best way to represent correlations to users, particularly to those who may be less familiar with data analysis (Choe, Lee & Schraefel, 2015). Others describe how inaccuracies in tracking data may affect the reliability of the insights that are provided (Rapp & Cena, 2014). We wish to focus on information overload as an additional challenge and open issue in personal informatics research.

2.2. Information Overload Challenges in HCI

Information overload has long been recognised as a potential problem for the design of interactive computing technologies (e.g. Hiltz & Turoff, 1985) and has been noted as a recurring issue for users of IT systems (Bawden & Robinson, 2008). While there is no universally accepted definition of information overload, Eppler & Mengis (2004) state that, in everyday terms, overload relates to the experience of "receiving too much

information" (p. 326). The notion that an individual can receive 'too much' information is premised on a vision of the human information-processing system as a limited capacity resource. Reflecting this, Galbraith (1974) states that information overload occurs when the information-processing requirements of a task exceed the information-processing capacity available to an individual. Eppler & Mengis (2004) note that historical research interest in information overload stems from a desire to understand how a person's performance varies in line with the amount of information to which he or she is exposed. This has been characterised as an inverted U-curve in which there is a "sweet spot" between the amount of information presented to an individual and the decisions made based on that information; adding information beyond the sweet spot causes overload and a decrease in the quality of decision-making (Chewning & Harrell, 1990). In general, information overload characterises situations in which receiving information becomes a hindrance, despite the information being potentially relevant and useful to the task at hand (Bawden & Robinson, 2009).

The experience of overload is thought to be affected by the time available to an individual (Schick, Gordon & Haka, 1990) as well as qualitative properties of the information that needs to be processed (Eppler & Mengis, 2004). For example, the degree of novelty, ambiguity, uncertainty, intensity, or complexity of information can either reduce or amplify the experience of overload (Schneider, 1987). Keller and Staelin (1987) also cite the "usefulness" of the information as impacting overload (p. 202).

In general, information overload is problematic because it has been associated with diminished reasoning ability and decision quality, poorer memory recall and feelings of confusion, stress, and anxiety (Schick, Gordon & Haka, 1990). Various attention deficit problems are thought to be associated with information overload. One such example is *continuous partial attention* (Stone, 2008), whereby users pay superficial attention to a wide assortment of information and do not give their full attention to any single piece of information. A further example is *attention deficit trait* (Hallowell, 2005), a distractibility and impatience due to excessive mental stimuli.

In the context of Human-Computer Interaction, problems of information overload are known to detract from positive user experiences with interactive systems, resulting in frustration, dissatisfaction, and a lack of user engagement (Koroleva, Krasnova & Günther 2010). For example, recent research has addressed issues of overload in social media applications such as Facebook and Twitter, where status and information streams threaten to become overwhelming to users (Koroleva, Krasnova & Günther, 2010; Bernstein *et al.*, 2010), and in email communication systems (Dabbish, 2005), where overload is related to increased stress and decreased job satisfaction (Mano & Mesch, 2010). However, few studies have examined issues of information overload in multifaceted personal informatics systems, despite anecdotal evidence that it may present a problem. For example, Choe *et al.* (2014) reported that "quantified selfers were often too ambitious" and "tried to track too many things", which resulted in 'tracking fatigue' and failure to effectively analyse and make sense of the data collected.

An additional challenge imposed by information overload in PI systems is that it may become difficult—or perhaps even impossible—to engage in sensemaking of one's data. Broadly, sensemaking pertains to finding meaning in a situation (Paul & Morris,

2009). In HCI, it refers to the cognitive act of understanding information (Whittaker, 2008). Although several detailed models of sensemaking have been proposed, (e.g. Russell, Stefik, Pirolli & Card, 1993; Weick, 1995; Lee *et al.*, 2016), the act of sensemaking can generally be characterised as a process which involves searching for and creating structures and representations in information, organising and encoding the information to place new knowledge in the context of what is already known, modifying representations, and consuming information for use in performing a task. Models often emphasise that sensemaking is an iterative process, whereby information processing may occur repeatedly until sensemaking is successful. Hence, the act of sensemaking is also characterised as explicit and effortful (Mamykina, Smaldone & Bakken, 2015). Previous work has shown that sensemaking can become more difficult in situations of overload, or when information is unfamiliar and appears in large quantities. For example, Kelly & Payne (2014) reported that users of a system designed for collaborative information seeking struggled to make sense of search returns due to the sheer volume of pages gathered and the presence of large amounts of irrelevant content. Similarly, Lee *et al.* (2016) found that, in a qualitative study of various information visualisations, users often 'floundered' when trying to make sense of visual representations that were unfamiliar.

The recurring nature of information overload-related challenges in HCI provides reason to believe that similar challenges are likely to affect the use of multifaceted PI systems, particularly given the likelihood that these systems involve the presentation of large volumes of information that is novel to the user and requires effort to interpret and understand.

2.3. Strategies for Dealing with Information Overload

The most simple strategies for dealing with information overload are those of *information avoidance* (Sweeny, Melnyk, Miller & Shepperd, 2010) and *information withdrawal* (Savolainen, 2007), whereby information is simply ignored, or only a limited number of sources are considered. These strategies are affectively oriented, guided by individuals' adverse emotional responses to excessive information (Savolainen, 2007). More sophisticated strategies for dealing with overload can also be adopted. For example, *queueing*, where information is considered in smaller chunks at intervals, and *satisficing* (Schwartz *et al.*, 2002), whereby only a small amount of information is examined, on the basis that it provides just enough information to meet a need. Phillips, & Battaglia (2003) discuss an alternative approach: that of training sensemaking skill, which focuses on enhancing recipients' information processing capabilities, rather than altering the information presented, such that they become better equipped to deal with information, even when its quantity may be challenging.

The most commonly adopted approach for overcoming information overload, however, is to apply information filtering (Hanani, Shapira & Shoval, 2001). Filters are intended to determine whether information is relevant to the user according to some scheme of priorities, and weed out information that is presumed to be irrelevant. Examples of information filters can be found in many everyday applications, including Internet search results, social media content (Rader & Gray, 2015), personal email, and

news. All share the goal of automatically redirecting user attention to the most valuable information and using their limited time effectively (Hanani *et al.*, 2001).

Information filtering relates to a variety of processes that involve the selection, omission or ranking of information for people who need it (Belkin & Croft, 1992). It is most commonly used to support the management of large information flows and to expose users to only the information that is relevant to them (Hanani, Shapira & Shoval, 2001). Previous research has distinguished between *active* filtering systems, which actively seek information that is likely to be of interest to a user, and *passive* systems, which omit irrelevant items from incoming streams of information (Shapira, Hanani, Raveh & Shoval, 1997).

In the context of personal informatics, Li *et al.* (2010) note that systems must strike a balance between automatic and manual data processing, so as to alleviate demands on users, whilst allowing control to remain with the user. Hence, information filtering in this context should also seek to find a balance between reducing explicit user input and maximising the accuracy of representations of users' interests.

Interestingness Measures

In terms of understanding the characteristics of the information that is processed, *Interestingness* measures have played a key role in previous data mining and information filtering systems (Geng & Hamilton, 2006). These measures are intended to support the identification of general features and patterns in terms of their potential interest to users, across a wide variety of information types and contexts. Interestingness measures can be used to prune uninteresting patterns or actively select interesting patterns so as to reduce the information space (Agrawal & Srikant, 1994).

Although much work has been conducted in this area, there is currently no widespread agreement on a formal definition of interestingness. Geng & Hamilton (2006) argue that it is best treated as a broad concept that encapsulates a variety of measures that include *conciseness*, *coverage*, *reliability*, *peculiarity*, *diversity*, *novelty*, *surprisingness*, *utility*, and *actionability*. In a review of previous data mining literature, they provide a comprehensive summary of objective measures which capture particular aspects of interestingness. For example, peculiarity measures account for patterns that are infrequent and significantly different from the rest of the data (Zhong *et al.* 2003), and which may be unknown to the user, hence interesting. It is widely accepted that no single measure is superior to all others, or suitable for all applications.

2.4. Summary and Research Questions

Personal Informatics systems allow people to collect and process diverse personal data about themselves. As with other interactive systems that present streams of novel and complex data, PI tools have the potential to create situations of information overload. In the present work, we aim to understand information overload in the use of Exist and explore information filtering as a way of negating overload in PI systems more generally.

Our first research question (RQ1: What do users find interesting within the correlational outputs from their personal data, given the context of exploratory use?)

seeks to derive criteria upon which information may be filtered. Building on prior data mining research (e.g. Geng & Hamilton, 2006), we focus on qualities that make particular correlations *Interesting* as a way of determining what is likely to be valuable to users during initial, exploratory use of a PI system. This leads to our second research question (RQ2: How can we algorithmically curate interesting insights and alleviate information overload for users?), which is intended to result in better user experiences by providing a practical approach to selecting interesting information for presentation to users of PI systems.

3. EXIST USER STUDY

To address our research questions we designed a three-month study which would allow us to collect personal tracking data from a diverse set of participants, reveal statistical correlations between different facets of their data, and investigate challenges associated with exploration of this correlational information. We focus on Exist because it exemplifies a growing number of popular personal informatics systems that process varied, multifaceted personal data. Furthermore, Exist incorporates and expands upon many features present in systems put forward by the research community, e.g. the Health Mashups (Bentley *et al.*, 2013) and Mobile Health Mashups (Tollmar, Bentley & Viedma, 2012) systems. We use Exist to address the broader question of what constitutes interesting correlational information, as well as how one might filter this information on the basis of interestingness. These are issues that are likely to apply to any PI system. However, we recognise that some design features of Exist may contribute to a user's experience of information overload. We attempt to account for the particulars of its design in our study and analysis of results, such that our results can be extrapolated to other systems that have been the focus of researchers' interest in the personal informatics literature. Finally, from a pragmatic point of view, Exist permits full API access to user data, allowing us to use it as a ready-made tool for data collection.

3.1. Exist Features and Functionality

Exist is a commercial tool designed for the aggregation and analysis of data from a range of personal tracking technologies. It is a typical example of a multifaceted PI system in that it aggregates data from multiple self-tracking services and discovers statistical correlations present within the data. The service presents correlational information to its users as graphical visualisations and natural language statements (see Figure 2), e.g. 'You're more productive when you have more events', or 'You have a better mood when you listen to more music'.

The Exist platform advertises itself as a general tool to "track everything in one place" and "understand your life" (Exist, 2016), and, like many other personal informatics systems, enables exploratory use, whereby users can volunteer as much data as possible in order to see if interesting insights emerge. The developers of Exist reported having 581 active users at the end of 2015 (Hello Code, 2015), each paying a monthly subscription fee of $6 USD.

Figure 1. Data Categories and Attributes Integrated with Exist

Data Category	Attribute	Description	Tracking Technology
Weekday	day_of_week	Weekend or Weekday	n/a
Events	events events_duration	Number of events in current day calendar schedule Duration of events in current day calendar schedule	Google Calendar / iCal
Mood	mood_score	Mood rating for the day, 1=Terrible, 5=Perfect	Exist
Music	tracks	Number of tracks scrobbled by Last.fm	Last.fm
Physical Activity	steps steps_active_min steps_distance	Number of steps taken Number of minutes spent being physically active Distance walked	Fitbit
Productivity	distracting_min neutral_min productive_min	Number of mins. spent using 'distracting' apps/websites Number of mins. spent using 'neutral' apps/websites Number of mins. spent using 'productive' apps/websites	RescueTime
Sleep	sleep sleep_start sleep_awakenings sleep_end time_in_bed	Number of minutes spent sleeping Time of sleep Number of awakenings during sleep Time of awakening Total number of minutes spent in bed	Fitbit
Social Media	instagram_comments instagram_likes instagram_posts tweets twitter_mentions	Number of comments received on Instagram posts Number of likes received on Instagram posts Number of posts made on Instagram Number of tweets made on Twitter Number of mentions received on Twitter	Twitter / Instagram
Weather	weather_cloud_cover weather_precipitation weather_temp_max weather_temp_min weather_wind_speed	Percentage cloud cover Percentage chance of precipitation Maximum weather temperature Minimum weather temperature Maximum wind speed	Forecast.io

Figure 1 lists the categories of data and individual data attributes which could be tracked by Exist at the time of our study. These included daily measures of: physical activity and sleep (both recorded by a wearable Fitbit sensor); productivity and distracting time (recorded by RescueTime logging software); mood (self-reported mood scores on a 5 point Likert-scale ranging from 1 (Terrible) to 5 (Perfect), collected by daily emails); events (automatically retrieved from online calendars); social media interactions (from Twitter and Instagram); music listening (recorded by Last.fm 'scrobbling' from music players such as Spotify and iTunes); and local weather conditions (from Forecast.io).

Given that Exist presents correlations between each of these attribute pairs (of which there were 26 in total) it has the possibility to present up to 325 distinct correlations (mathematically, $\binom{26}{2}$ or *26 choose 2*) for users tracking everything. Each of these correlations is presented using a representation like that shown in Figure 2, which displays an actual correlation from our dataset. This includes a '% related' measure, which corresponds to the correlation r-value, a 'confidence' star rating of 1 to 5 stars, which corresponds to the correlation p-value for statistical significance (with a linear mapping between number of stars and p-value, where 1 star = $p<0.2$ and 5 stars = $p<0.05$), and an indication of the period (number of days' data) from which the correlation is calculated. Correlations between data attributes are only presented within Exist, or returned via the API, if they have a p-value < 0.2 and thus our data only includes correlations that fall beneath this threshold. Hence, Exist has a basic level of filtering for the purpose of reducing spurious correlations, however at least one spurious

correlation is almost guaranteed in the filtered dataset when making as many 325 comparisons at the $p<0.2$ threshold (family-wise error rate: $1-(1-0.2)^{325} \approx 1$).

Figure 2. Example Exist Correlation using Productivity and Events

3.2. Participants and Recruitment

Our study was advertised on our university noticeboards (physical and online). Prospective participants were offered the opportunity to try a range of tracking technologies, each of which would be connected to Exist, for a period of up to three months. Participants were given a brief introduction to Exist and were informed that they would be able to see its output following a period of data collection using the tracking technologies on offer. Participants were also offered entry into a prize draw for one £50 (~$70) Amazon voucher as an incentive to participate in the study.

Twenty participants were recruited in total. Ten individuals were recruited in May 2015 and a further ten in October 2015. All participants lived in and around the city of Bath in the United Kingdom, due to the requirement for them to be able to meet the researchers in person to collect their tracking devices, receive support in configuring user accounts, and participate in face-to-face interviews after the study.

Participants were between 21–60 years old (M= 33.1, SD= 11.8), comprising a range of occupations and an equal mix of males and females. The sample included eight members of the general public who had heard about the study via word of mouth, as well as twelve university staff and students. Seven of the participants had previous experience using a tracking technology (e.g. Nike+ Fuelband, Garmin GPS Watch, Weight Watchers food logging app, Last.fm music 'scrobbler'). During an initial briefing session, participants were asked to describe their reason for wanting to participate in the study. We found that all of the participants expressed curiosity and interest in using tracking technologies to uncover information about themselves, but did not have clearly defined problems to address, nor hard-set goals for behaviour change. Beyond the £50 prize draw, participants' stated motivations for participating in the study included: "I'm interested to see what it [Exist] tells me about myself" [Participant 4, Office Administrator, F, 26], "I want to know what tracking my life can offer me" [Participant 5, Postgraduate Student, F, 25], "To discover the truth about how I really am" [Participant 8, Postgraduate Student F, 21], and "Interested to see if I need to make any changes to my lifestyle" Participant 14, Teacher, M, 52].

We argue that these participants reflect a growing proportion of 'exploratory' personal informatics users, following the emergence of personal tracking as a mainstream activity (Lupton, 2013; Gurrin, Smeaton & Doherty, 2014; Rapp & Cena, 2016). We consider potential issues of information overload to be worthy of investigation for this type of user, because it may be harder to predict in advance the type of information and insights that they are looking to gain from their use of the system.

3.3. Pre-Study Data Collection and Exist Setup Procedure

Ethics, Briefing and Consent

The study outline was reviewed by an independent researcher within the Department of Computer Science at the University of Bath, in accordance with a local code of ethics. Each participant was briefed individually about the purposes of the study, the data that would be collected by each of the tracking technologies, and their right to withdraw and/or request the deletion of their data at any time prior to the study's completion. They were also informed that any data collected by the researchers would be fully anonymised prior to its storage and use for publication. Those who chose to participate in the study gave written informed consent and were then able to selectively 'opt-in' to their use of each tracking technology. Participants were also informed that the minimum planned study period was one month (the amount of time required for Exist to begin producing correlational information), but that they would be given the opportunity to continue their participation for an ongoing period of up to three months (90 days) if they chose (the maximum period for which Exist calculated correlations). Section 3.4 provides details on the duration of each participant's involvement in the study.

Pre-Study Phase

Participants completed a pre-study interview and questionnaire which obtained demographic information, accounts of their motivations for participating in the study, and previous experiences of using tracking technologies.

As part of the pre-study questionnaire, participants were asked to consider which aspects of their lives they believed were correlated with others. This was done to enable post-study comparisons between users' mental models of the correlations present within their behaviour and those that could be detected by the Exist system. This would allow for an investigation into whether the congruence between users' mental models and the output of the system had any bearing on participants' assessments of the interestingness of the correlational information provided.

Participants were given explanations and examples of positive, negative and uncorrelated outcomes between the variables that could be recorded by Exist, such that they were clear on their meanings. The participants were then required to make predictions about the nature of correlations between different facets of their lives. These were: Events, Physical Activity, Productivity, Mood, Sleep, Schedule, Social Media Usage, Music Listening and Weather.

Predictions were elicited by presenting a set of four statements for each of the pairwise combinations of facets. The first statement described a positive correlation (e.g. "When my mood is better, my sleep is better"), the second described a negative correlation (e.g. "When my mood is better, my sleep is worse"), the third described no correlation (e.g. "My mood is not correlated with my sleep"), and the fourth statement read "I do not know if…" (e.g. "...my mood and sleep are correlated"). Participants were instructed to circle one of the four statements based on their understanding of typical patterns of behaviour in their lives. These statements were designed to match the natural language correlation statements that would be produced by the Exist system.

Exist Setup Procedure

Twenty paid Exist user accounts were created for participants by the researchers. Each participant was invited to connect their pre-existing social media accounts to Exist, e.g. Twitter and Instagram, and calendar applications, e.g. Google Calendar and iCal, since it is necessary to leverage existing social infrastructure for these types of services. The use of existing accounts also served to minimise disruption to participants' normal behaviour during the study, such that the data collected was reflective of their typical activities. Separate accounts were created for Fitbit, RescueTime, Last.fm, Forecast.io, and Swarm by Foursquare, where users did not already have accounts of their own. Services such as RescueTime, Last.fm, and Forecast.io were not considered disruptive, since they could operate surreptitiously during the study period to capture productivity, music listening activity, and weather data. The wearable Fitbit device presented minimal disruption to users' typical physical activity. The Swarm by Foursquare account was created as a means to provide location data to the Forecast.io service, such that weather information could be collected based on each participant's location throughout the study.

Following completion of the pre-study questionnaire, participants were given a Fitbit device and charger. The researcher demonstrated their functions and provided guidance on data synchronisation, charging and maintenance. Participants were then guided through the installation of software necessary for viewing and synchronising data (e.g. Fitbit app, RescueTime client, Last.fm scrobbler) on personal devices which they had brought along. Instructions were provided for subsequent installation on any devices that could not be brought to the session.

Participants who had opted to provide daily mood scores selected either email or mobile app methods for submitting their scores. Both methods involved a scheduled notification being sent to the participant at a time of their choosing, containing the Likert scale from which a mood score for the day could be selected. Participants were assisted in setting a time to receive this notification and were shown how to submit a mood score.

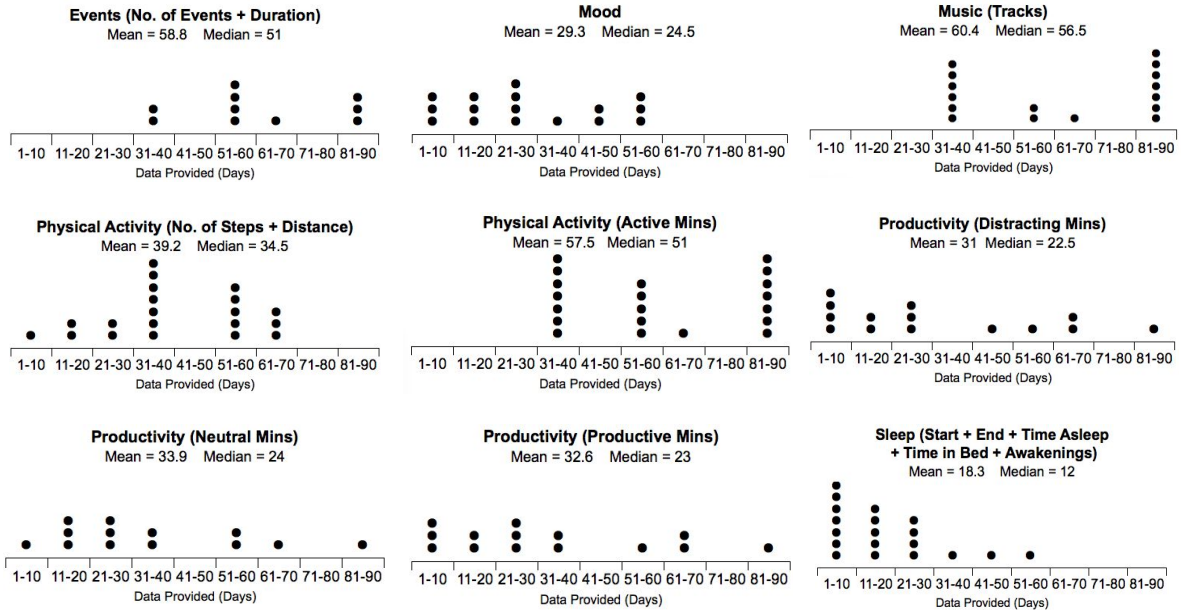
Participants were provided with login details and instructions to access each of the tracking services that they would be using during the study. Hence, they could view the services' respective data dashboards and analysis features, and use them to explore information relating to goals, status, discrepancies, history and trends. Furthermore, participants were exposed to the various mechanisms that each service uses to encourage continual data collection (e.g. Fitbit badges, Rescuetime goals, etc.). Participants were

therefore able to benefit from their use of these services, irrespective of their connections to Exist. Participants were not given login details for the Exist account and were unable to see the results of the analysis of their data (i.e. correlations between facets) during the study. They were informed that they would be able to view the output of the Exist system in a post-study session with the researcher. This was done so that the researchers could record participants' initial reactions to the correlational information provided.

3.4. Exist Study Data Collection

During the study, participants provided data to Exist by using the tracking technologies that had been given to them. The researchers used Exist's Python API to request all of the data collected from each of the connected services, as well as any data relating to the correlations that Exist had calculated (i.e. correlation coefficients (r-value), statistical significance (p-value), and time period (number of data points)). We observed that Exist calculated its correlations based on the most recent 28–90 days' data, hence our participants were invited to participate for a similar period. Thirteen out of twenty participants opted to continue beyond the minimum one-month period. Of the seven participants that concluded the study after one month, three cited the reason of being unavailable to meet with the researchers or consistently track themselves in the coming months (e.g. due to travel and vacation arrangements). Two participants felt that device maintenance (primarily for the Fitbit) presented a burden, and two gave no reason for wishing to conclude the study after one month.

Figure 3. Dot plots showing the amount of data provided (in days) for each of the recorded attributes. Each dot represents one participant.



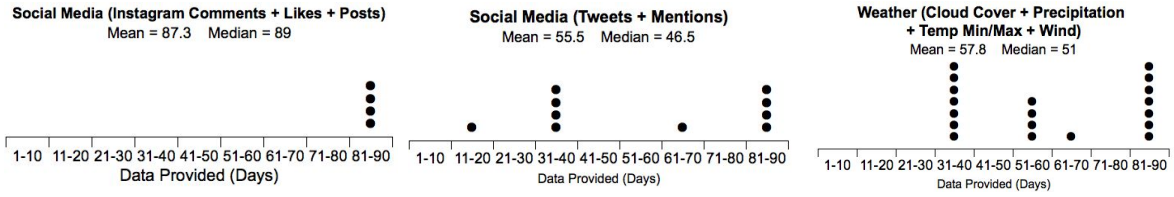


Figure 3 shows the amount of data collected for each attribute by each participant during the study (i.e. the number of data points provided to Exist, where each data point is a value for a particular day). For example, Exist received 50–60 days of *Mood* data from three participants and 30–40 days of *Music* listening data for six participants.

It can be seen that the majority of participants recorded fewer days' data for *Sleep* than *Physical Activity*, due to their removing the Fitbit at night. Three of the participants contacted the researchers during the study to inform them that they were finding the Fitbit uncomfortable and inconvenient to wear during sleep. Several other participants mentioned either forgetting to wear the device, or treating the nighttime as a regular opportunity to charge the device, which accounts for the paucity of data in this category. Similarly, mood data was diminished due to the manual nature of its collection, with participants occasionally forgetting or being too busy to enter a mood rating every day.

Many data attributes were automatically collected, with no additional effort required from the participants, e.g. those relating to Weather and Music. As such, data was provided for the full duration of the participants' involvement in the study.

3.5. Post-Study Data Collection

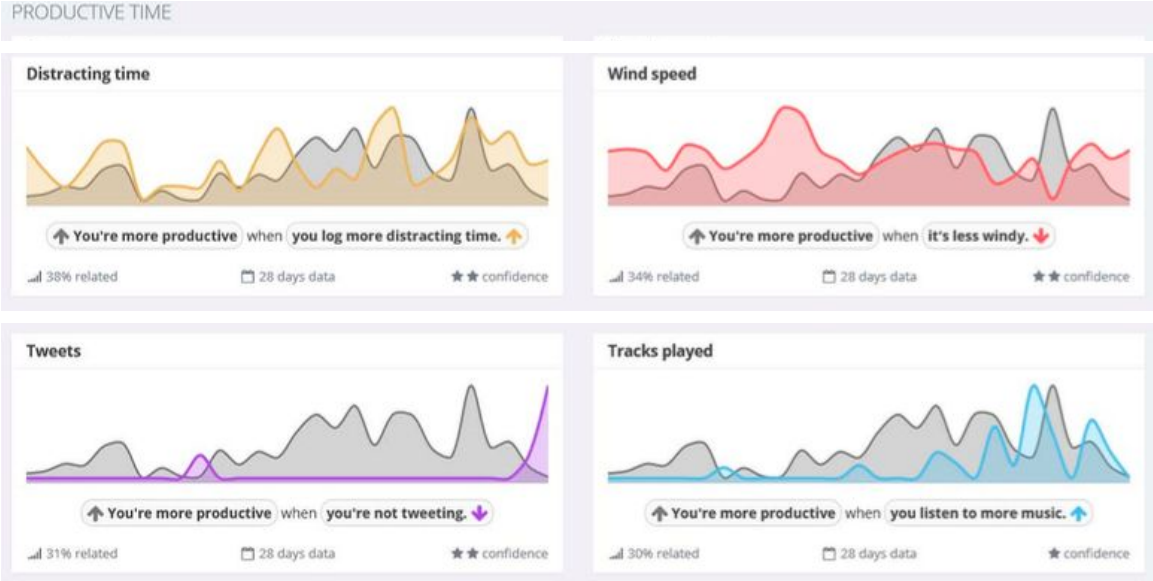
Post-Study Interviews

We conducted a face-to-face semi-structured interview at the end of each participant's data collection period. Interviews lasted between 39–111 mins (M=78.5, SD=21.5). Participants were first informed that they were free to end the interview session, or to pause for a break and refreshment, at any time. We then asked open-ended questions relating to each individual's tracking experiences during the study. This conversation was intended to gain some insight to any problems that may have occurred during the data collection period and served as a warmup to get the participants talking.

Next, participants were shown printed screen captures of all of the correlational information revealed by the Exist system. Figure 4 shows a sample of the post-study correlation output from Exist for one of our participants. The number of unique correlations ranged from 12–113 (M=62.4, SD=30.6). All correlations relating to a particular attribute were presented on a single page, precisely as they appeared within a single screen in the Exist interface. Printed screenshots were used to enable participants to make notes whilst participating in the session, which could be used to support our interview analysis. Oulasvirta, Hukkinen & Schwartz (2009) note that several previous

studies have presented user interfaces to users in paper form and that "there is no effect reported that casts into doubt the validity of paper form" (p. 518).

Figure 4. Sample of Post-Study Correlation Output from Exist



This was participants' first opportunity to see the correlations that had been derived from their data. Participants were asked to think aloud whilst reviewing the output, commenting in particular on aspects of the output that they did not understand; on things that they thought were particularly interesting or insightful; on correlations that they deemed obvious or uninteresting; and on identified outcomes that they did or did not expect to see within the results. They were also asked how they might utilise or act upon the information provided. Participants were free to explore the correlations in any order they chose, switching from one page of attributes to another if they so wished. The interviewer probed for deeper explanations of the output until no new information seemed to come from participants' responses, and all correlations had been examined. At the end of each interview session, the participant returned all of the tracking devices that they had been using, and were assisted by the researcher in uninstalling any tracking applications and software which they no longer wished to use in their own time. The interviews were later transcribed by the first author and were analysed inductively, using phases of open coding to identify concepts within the data, and axial coding to identify relationships among the concepts (Corbin & Strauss, 2008).

Post-Study Questionnaires

To identify the factors associated with interestingness of correlational information in Exist, we asked participants to rate a random sample of ten correlations within their data, according to nine statements. We decided that ten correlations presented a reasonable balance between collecting a sufficient amount of data for analysis and the burden of time and effort that would be placed on participants to complete 90 ratings, following pilots with two volunteers from our research group. Both volunteers were using Exist with their own data, but were not participating in the study.

The nine statements presented to participants were generated based on subjective measures of interestingness taken from data mining research. Four measures were adapted from Geng & Hamilton (2006), comprising: Utility (whether a correlation can be used to support progress towards a goal), Surprisingness (the quality of being unexpected or contradicting a person's expectations), Novelty (whether the correlation presents a finding that was previously unknown), and Accuracy/Reliability (whether the correlation is perceived to capture the true nature of their behaviour). These were accompanied by four additional measures that arose from our own discussions about further factors that might influence users' assessments of the correlations, namely: Valence (i.e. the intrinsic attractiveness (positive valence) or aversiveness (negative valence) of a correlation), Uniqueness (the perceived likelihood that others users might receive the same correlation), and Stability (the likelihood that a correlation might change over time or in different contexts).

Additionally we included one statement to capture users' overall interest in seeing a particular correlation within Exist: "This is a correlation that I am interested in seeing when using this system". All statements were presented in the form: "This is a correlation that…", followed by an affirmative statement about the subjective measure, i.e., "...is surprising", "...is pleasing", "...makes me unique", etc. Participants were asked to respond to each of the statements on a Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). Since users were asked to rate the randomly sampled correlations after they had reviewed and discussed their entire Exist output with the researcher, we were confident that participants fully understood the correlations they were rating, and that the ratings provided were likely to account for relative comparisons against the set of all other correlations they had seen.

4. EXIST STUDY FINDINGS

In this section we first report findings from our interview analyses. Our focus is on issues that arose when participants were faced with the task of interpreting their correlations, and which speak to the challenges of dealing with information overload[5] rather than their more general experiences with self-tracking. Participants' claims give us reason to believe that information overload was a genuine problem for some individuals when encountering their Exist data. Subsection 4.2 explores what makes a correlation generally interesting to users, such that filtering mechanisms designed to reduce information overload can take these factors into account.

4.1. Information Overload in Exist

When viewing their data for the first time, six participants considered the presentation of numerous correlations to be a positive characteristic of the system, e.g. *"There's so many. Yay! That's really cool." [P9]*. These participants valued the prospect of having many outputs to explore, seemingly because they believed that this would

[5] Additional findings from this analysis, including issues associated with information presentation, transparency, and fragmentation, are reported in Jones & Kelly (2016).

correspond to the insights they could derive from the system, e.g. *"I've been really excited to find out how much it has to display about me." [P13].*

However, it was apparent that, for at least four of the participants, the initial satisfaction of receiving many correlations gave way to frustration with regard to the cognitive effort required to review and reflect upon all of the outputs: *"There's so many it's hard to reflect on what's missing." [P1], "I hope this teaches me enough to justify the effort of going through all of this." [P5], "Actually, this is tiring!" [P20],"You've really got to have your brain in gear to go through these!" [P17].* These four participants were not presented with appreciably more correlations than others (*Range*=39-105, *M*=66.0, *SD*=30.2), suggesting that it was not only the highest quantities of output that presented an overload challenge.

The volume of unfamiliar correlations seemed to impinge on participants' ability to make sense of their data. One participant [P1] referred to his inability to form rational conclusions as 'analysis paralysis': *"Is that right? Oh I can't figure it out, I'm getting analysis paralysis. I've got it big time… I'm struggling to figure out what these things mean now." [P1].* The quantity of outputs also led to difficulties in cross-referencing correlational information with earlier insights. One of the supposed benefits of multifaceted systems is the ability to enable links between different aspects of a user's life. One participant found it difficult to recall correlations that he wanted to revisit, due to the amount of information presented to him: *"This correlation seems to be related to that one I found interesting earlier. Which one was it?… there's a lot here" [P14].*

Difficulties interpreting the outputs of the system were compounded by the unfamiliarity of correlational information of this kind, with P17 claiming *"I'm not used to this sort of thing".* In addition, some insights were deemed to be duplicates of one another because of the level of granularity at which they were interpreted. For example, Participant 3 felt that having multiple variables within a single weather data category *"goes into more depth than is needed"* and suggested that the system could collapse variables such as 'cloud cover', 'wind speed' and 'precipitation levels' into a single 'weather' category, making a large number of outputs redundant (e.g. *"I would say rain, whether it's dry, is the only one I would need." [P3])*.

An additional source of frustration was the presence of many insights that were considered "obvious", e.g. covering more distance when taking more steps: *"There's quite a lot, I'd take out all the "it's rainier when it's cloudier", "you sleep more when you're in bed more"...all of that obvious stuff... Too much to deal with." [P4]. "There are some silly things, like you're more active when you have more steps. That's kinda obvious." [P10].* Bentley *et al.* (2013) reported similarly negative reactions to "obvious" insights in Health Mashups, noting that their presence did not make sense to some users and created a tension between telling people what they already know and educating them with new insights. These 'junk' correlations likely offer little value and may contribute to the feeling of being overloaded with information.

Finally, there was evidence that one participant was inclined to avoid or withdraw from the data because there was *"a lot to sift through..."*, suggesting that third parties

could play a role in filtering on their behalf: *"...I'd rather just take to this to my doctor and get him to make sense of it." [P12]*.

In summary, our participants encountered a number of issues that we interpret as related to information overload. While not all of our participants experienced overload, our results give us sufficient cause to believe that it could be a problem for some users, motivating our interest in exploring how overload could be avoided by filtering outputs that are of little interest.

4.2. What Makes a Correlation Interesting?

In this section we draw on additional interview analyses and ratings of sampled correlations to identify features of interestingness that could be used to support information filtering in PI systems. We identify qualities associated with the interestingness of correlations, which in turn informs our subsequent selection of measures that might be used in information filtering.

Qualitative Findings on Correlation 'Interestingness'

Our analysis identified six initial properties that participants suggested were related to the interestingness of a correlation. They were influenced by whether the correlation was in some way surprising; was perceived to be unique to the user; matched a user's expectations; appeared to have practical utility; whether it was obvious/easy to deduce; or whether the correlation spoke to the user's specific interests in understanding their life.

First, participants were particularly interested in correlations that were unexpected and therefore surprising. e.g.*"Really? That's so surprising. I love it." [P8]*, *"I didn't actually expect it to show that." [P20]*. Correlations were surprising for a number of different reasons, including that the strength of a correlation was greater than expected ("*...It's a strong correlation by the looks of it. Surprising." [P2])*, that it did not match a participant's existing model of their own behaviour, or because a participant was impressed that a relationship was able to be uncovered by the technology that they had been using (*"That's amazing how it picked that up. I'm impressed." [P17])*.

When receiving information that was not expected, three participants considered the extent to which similar correlations might appear for other people. They were keen to understand whether surprising information might imply deviations from 'normal' behaviour (*"Really! So...is that normal?" [P15], "See, I knew I was weird!" [P19])* or differences from the 'average' person (e.g. *"What is an average person like?" [P9])*.

Although surprising correlations were often a focus of participants' interest, additional utility was found in seeing correlations that confirmed prior expectations, e.g. *"That's useful because it reaffirms what I already thought anyway." [P3]*. For others there was limited interest in information that was already known e.g. *"I'm not sure if this is all that useful. I could already explain most of them." [P4]*

We found evidence to suggest that perceived practical utility acts a determinant of interest, e.g. *"That's useful. Maybe now I could look for an indoor exercise activity."*

[P3]. Three of the participants found value in the potential to identify possible changes that could be enacted within their lives (*"I seem to sleep worse when I have a lot going on the next day. Maybe I could try to spread my meetings a bit more throughout the week.." [P13],* or, *"Quite interesting that I'm more productive when I listen to more music.... ok maybe I'll have to start listening to music." [P8])*. Another three participants explicitly stated that they struggled to identify how they might make use of certain information within their lives (e.g. "*I mean...what am I gonna do with that? I'm not sure it helps me to see this" [P6])* and hence they were less interested to see it presented within the system. For at least three participants, assessments of how information could be put to practical use appeared to shape their overall impression of the Exist service. Some found the outputs to be useful, e.g. *"This is really cool. I'm definitely going to try and act on all of these insights" [P14]*. One participant was less impressed, e.g. *"I don't think it really tells me enough useful stuff that I don't know already for it to justify the effort." [P16]*.

Two participants were particularly interested in relationships that they could not easily deduce without the support of tracking data, or without secondary sources of information to guide them: *"I think these are valuable… about the weather, because I can't work this out myself. I can only guess this stuff." [P1], "The most interesting thing about it is the sleep. Because you don't really know how well you sleep. You can't really measure it yourself" [P8]*. Participant 3 suggested that some facets, whilst containing particular correlations that were interesting, did not necessarily produce universally interesting correlations, and that it was the particular combination of facets which mattered: *"That particular one is interesting. But not all physical activity ones are interesting. Some relationships I've just wondered about a bit more because, like, they'd be useful to know." [P3]*.

Finally, participants were interested in correlations that spoke to their prior interests. For example, one participant had previously investigated the relationship between sleep and music by seeking information online: *"I've looked into that relationship before actually. I wanted to know if there's music to help me sleep better." [P8]*.

Quantitative Analysis of Correlation 'Interestingness'

Building on our qualitative analysis, we explored features of Interestingness based on the randomly sampled correlations for which participants provided Likert-scale ratings (see subsection 3.5). The aim of this was to determine which variables were significant predictors of interestingness, such that these variables can be used as a basis for information filtering. A linear mixed effects model was constructed to predict the Interestingness rating (dependent variable) based on a combination of subjective and objective measures (independent variables). The subjective ratings encapsulated the qualities of correlations previously discussed in subsection 3.5 (Utility, Surprisingness, Valence, etc). The objective measures came from three sources. First, *Data Categories*, which comprise the high level categories of data tracked in the study (see Figure 1). These were: Weekday, Events, Mood, Music, Physical Activity, Productivity, Sleep, Social Media and Weather. Next, *Correlation Characteristics*, comprising correlation confidence (p-value), correlation coefficient (r-value), and the number of days data

analysed (period), as well as a label denoting whether the correlation was uni- or multi-faceted; and *Predictions*, indicating whether participants' pre-study correlation predictions matched with the actual correlation outcomes in the data.

We anticipated that multi-faceted correlations (inter-correlations between two distinct data categories, e.g. *Physical Activity* and *Sleep)* would be more interesting than uni-faceted correlations (intra-correlations between attributes from the same category, e.g. *Sleep*: time to bed vs. *Sleep*: time spent asleep). Previous studies have shown that users are often interested in insights that span multiple different types of data, and that multi-faceted analysis encourages engagement with personal informatics technologies (e.g. Dingler, Sahami & Henze, 2014; Bentley *et al.*, 2013). Hence, we differentiate between uni-faceted and multifaceted correlations in our linear mixed model analysis.

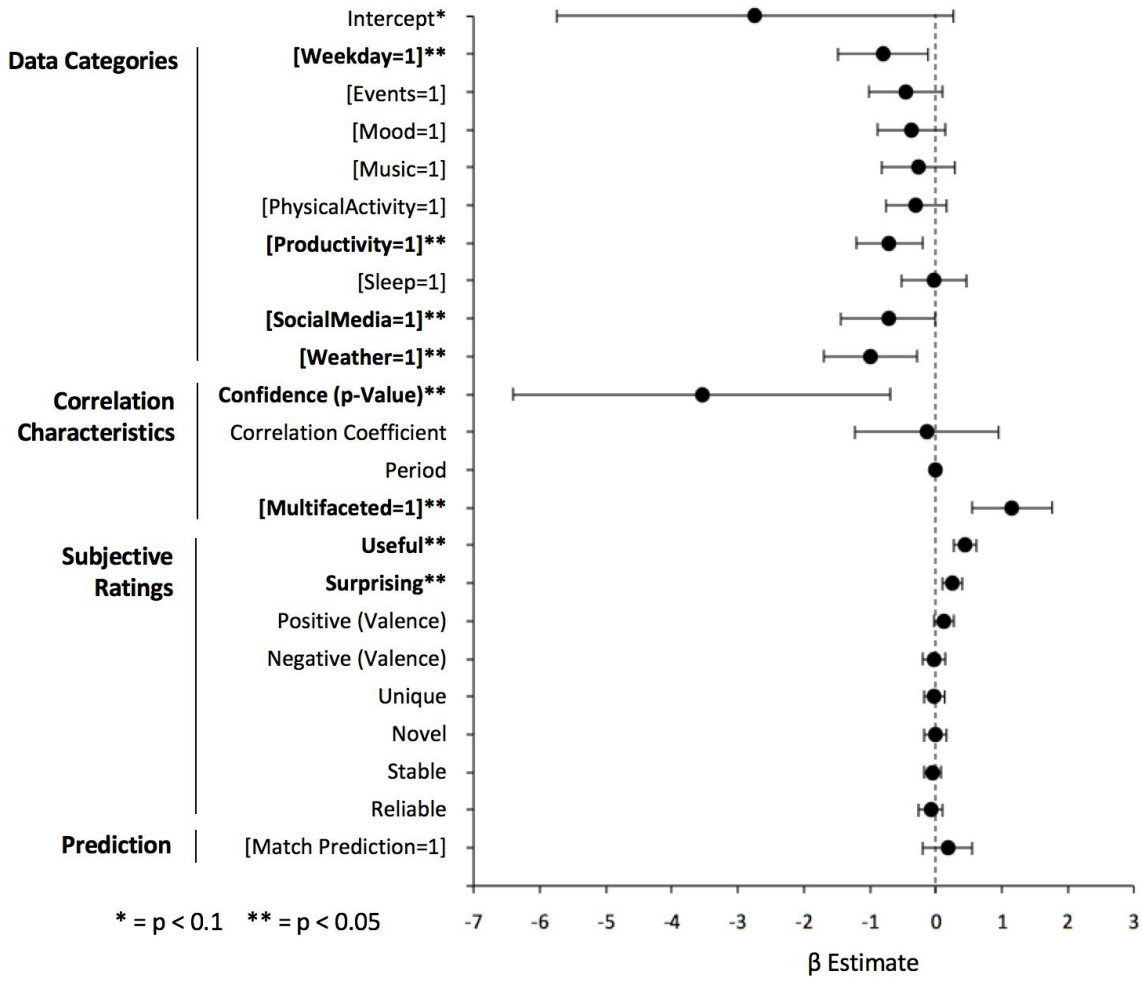
We also anticipated that participants' interest in a correlation may be influenced by its predictability. Comparisons between participants' predictions and their actual results revealed that they had correctly anticipated 46.7% of the correlations in Exist. 37.8% of correlations contradicted pre-study predictions, and 16.5% had no associated predictions, because participants were unable to forecast an outcome. We discuss these prediction-match findings further in Sections 4.3 and 6.

Linear Mixed Effects Model Results

We conducted a linear mixed effects analysis (in IBM SPSS Statistics 23) using restricted maximum likelihood (REML) estimation methods. Mixed effect models were required because our data contained multiple, inter-dependent observations for each participant. We treated participant identity as a random effect to account for the repeated measures nature of the data and variation due to individual differences in interest ratings. Fixed effects for all other independent variables were added one at a time and retained if they improved the fit of the model according to Akaike's Information Criterion (AIC). Our model contained 200 observations; 10 from each of the 20 participants. An association between the dependent variable and independent variable was considered to be significant in the final model at the level $p<0.05$. Figure 5 shows the standardised β estimates, significance values, and confidence intervals for each variable included in the final model. Binary variables (such as *Data Categories*, *Uni/Multi-faceted* and *Prediction Match*) were encoded as 0=False and 1=True within the data. Hence, a β estimate of ~+1 for [Multi-faceted=1] indicates an increase in interestingness when correlations are multi-faceted rather than uni-faceted.

The linear mixed effects model reveals that participants were more likely to rate correlations as interesting when they were seen to be surprising or useful; when they presented between- rather than within- data category relationships (i.e multifaceted rather than unifaceted); when they were not associated with uninteresting data categories (i.e. weather, weekday, productivity and social media); or when they exhibited low p-values (high confidence).

Figure 5. Linear Mixed Effects Model (Items in bold indicate significant predictor variables within the model. X-axis indicates the β coefficient estimate for each variable. Error bars show upper and lower 95% confidence intervals.)



4.3. Discussion of Exist Study Results

Our study of Exist produced several key findings. Initial qualitative analyses revealed that participants were able to find interesting correlations within Exist's output, reiterating the potential for PI systems to support positive self-reflection (Li, Forlizzi & Dey, 2010; Elsden, Kirk & Durrant, 2015). However, we found that participants struggled with the task of reviewing their correlations, and that this was due in part to the sheer volume of information produced. Although our study makes use of a particular PI system, we see the issue of overload as one that is likely to arise in any multifaceted PI system, given the potential for such systems to produce a large number of outputs.

Our subsequent analysis sought to address the problem of overload by exploring qualities that made particular correlations interesting to participants (as per RQ1), with a view to using these as a basis for information filtering. First, our interview data revealed six qualities that participants associated with interesting correlations. These qualities broadly corresponded to the surprisingness of a correlation; its match to prior expectations; how likely a similar outcome might be for other users; its perceived

practical utility; its connection to facets of their life that they were particularly interested in; and its apparent obviousness.

These results dovetail with those of our quantitative analysis. Specifically, two subjective qualities, *surprising* and *useful,* were found to be significant predictors of interestingness, corroborating comments about these qualities from the interviews. In addition, the category of data producing the correlation was revealed as a significant predictor, which may speak to participants' interview comments that expressed particular interest in some life facets over others. While interview comments also suggested that the specific combination of facets was an important factor, it was beyond the means of our study to collect subjective interestingness ratings for every pairwise combination of facets (due to the myriad of possible combinations) to include in our quantitative model.

Our mixed model analysis also revealed significant predictors of interestingness that were not explicitly singled out within our interviews, namely that multi-faceted correlations were typically more interesting than those from a single facet, and that lower correlation p-values (higher confidence levels) were of greater interest. Participants' interest in multifaceted correlations may reflect a desire to map the relationships between distinct life facets (Li, 2011), and can be linked to interview comments which indicated that uni-faceted correlations were sometimes considered to be obvious (e.g. you sleep more when you spend more time in bed, or you get more steps when you spend more time active). Participants' interest in the confidence of correlations may speak to a desire for outcomes that are seen to be trustworthy and which provide sufficient evidence for taking action.

We hypothesised that the surprisingness of a correlation might be most closely related to participants' ability to predict its presence in advance. However, this was not a significant predictor within the mixed effects model, and comments from interviews suggested that surprisingness was also associated with other factors such as the perceived likelihood of detecting such an outcome from the data, and whether certain outcomes might appear for other users.

Both analyses contribute a characterisation of qualities that make correlations interesting in the context of exploratory use. We suggest that these qualities can used as a basis for filtering outputs, such that users are given a more manageable and digestible set of outcomes for self-reflection. That being said, our representation of interestingness is not yet wholly tractable for practical use in building an operational filter. Firstly, this is because of the extent to which it requires explicit input from users. Two of the significant predictor variables in the model, *surprisingness* and *utility*, are purely subjective measures, requiring access to the user's opinions about their data. Collecting such data in a PI system is feasible in theory but would undermine the goal of reducing the effort required to identify interesting insights.

Secondly, the model is limited in terms of its generalisability and scalability with respect to the possible inputs that users may wish to integrate and analyse. Our results reveal that levels of interest are influenced by the categories of data being correlated, and the interactions between these categories. Hence, data processed by PI systems other than Exist may differ from the categories captured within our model. Even within Exist,

the developers periodically add new categories, and our model could not account for potential user interest in these without requiring the collection of further 'ground truth' interestingness data from a sample population of users.

We argue that the aforementioned problems can be resolved by the identification of additional objective data measures, which are readily available to the system and its filtering mechanisms, and which provide suitable substitutes for subjective measures associated with interestingness. Thus, in the next section, we present two sources for such measures, namely: *Interestingness* measures *(Generality, Diversity and Peculiarity)*, derived from previous literature on data mining, and *Google Trends* measures (*Mean Google Trend Score* and *Google Trend Rank)*, which we use as a proxy for interest in particular combinations of data categories.

5. AUTOMATED FILTERING USING OBJECTIVE DATA

In this section we address RQ2 by exploring how we might practically build a filtering mechanism in the form of a machine learning classifier that can operate on the basis of objective data. We incorporate the objective measures from within Exist presented thus far, and exclude subjective measures which necessitate input from the user. We present two additional sources of objective data that may embody the qualities associated with the interestingness of correlations (as discussed in Section 4.3). These are *Interestingness* measures (specifically: Generality, Diversity, and Peculiarity), derived from Geng & Hamilton (2006), which capture aspects of surprisingness and considerations about the comparisons against other users, and *Google Trends* data, which captures qualities relating to the inherent interestingness of different data categories and their combinations.

5.1. Additional Objective Data for Predicting Interestingness

Objective Interestingness Measures

Data mining research has provided a variety of objective measures of the 'Interestingness' of patterns within a dataset (see Geng & Hamilton (2006) for a comprehensive review of these measures). Based on the results of our former qualitative and quantitative analyses, we identify existing measures which may help to reveal 'surprising' correlations in particular. We select measures that allow the identification of atypical correlations, or which provide an indication of a correlations' prevalence amongst other users. This is apropos of interview comments, which indicated that participants considered the extent to which correlations might also appear for other users, or the likelihood that they might be found within their data, when considering their interestingness.

Specifically, we identify three objective measures to add to our machine learning classifier. First, *Diversity*, which refers to the range of observed outcomes for a particular correlation, amongst all users. The relationship between a pair of attributes has three possible outcomes: positive, negative, or no correlation. The observed outcomes for the relationship can be considered diverse if some users have a positive correlation, some

have negative, and some have none. Conversely, a relationship is not diverse if all users have the same outcome. We hypothesize that correlations with diverse outcomes are more surprising (and hence, interesting), since they may be harder to predict in advance.

Diversity for a pair of attributes is calculated as: the number of observed unique correlation outcomes for a particular attribute pair (e.g. Events and Instagram Comments), divided by the number of possible outcomes (i.e. 3). If only 1 outcome is found between *Events* and *Instagram Comments*, the diversity score for *Events* and *Instagram Comments* is 1/3 = 0.33.

The second measure is *Generality,* which refers to how common a correlation outcome is amongst all users. Generality differs from diversity in that it captures the proportion of all users that obtain the most common outcome. Although patterns in data that are highly general are often characterised as interesting (Geng & Hamilton, 2006), we hypothesize that low generality correlations are surprising and therefore interesting.

Generality is calculated as: the maximum number of users with a matching correlation outcome (positive, negative or no correlation) for a particular attribute pair, divided by the total number of users recording data for that attribute pair. For example, if 17 users have positive correlations, 2 have negative correlations and 1 has no correlation within their recorded *Mood* and *Sleep* data, the generality score for *Mood* and *Sleep* is 17/20 = 0.85.

Finally, *Peculiarity* refers to the distance of a user's correlation outcome from that of other users (Geng & Hamilton, 2006). Peculiar correlation outcomes represent outliers within the data. Peculiarity builds on generality by taking into account not only the extent to which users have matching outcomes, but also whether a given user is in a minority or majority. Hence, peculiarity is calculated for each attribute pair, per user. We hypothesise that the surprisingness of correlation outcomes may be reflected in their peculiarity.

The peculiarity of a particular correlation outcome is calculated by taking a particular user and pair of attributes (e.g. Mood and Sleep). We calculate the number of users with the same correlation outcome as the selected user, divided by the total number of users recording data for that attribute pair. This value is then subtracted from 1 such that more peculiar correlation outcomes are closer to 1, and less peculiar outcomes are closer to 0. This means that if 17 users have positive correlations, 2 have negative correlations, and 1 has no correlation between *Mood* and *Sleep*, the peculiarity score for the single user with no correlation is high: 1-(1/20)=0.95. The peculiarity score for the 2 users with negative correlations is slightly lower: 1-(2/20)=0.90. For the 17 users with positive correlations the peculiarity score is low: 1-(17/20)=0.15.

Google Trends Data

As highlighted by our linear mixed model results (Section 4.2), the level of interest in correlations varied between data categories. Furthermore, our qualitative analysis revealed that the particular combination of data categories involved in a correlation was important for determining its interestingness. In this section we describe a source of objective data which may serve as proxy for users' interest in understanding

relationships between data categories. Specifically, we consider the popularity of Google searches related to each of the category pairs from our study.

Previous research has demonstrated that search query data can provide reliable indications of current interest in topics and can be used to forecast near-term values. For example, Choi & Varian (2012) showed that the volume of queries relating to different travel destinations acted as a reliable indicator of actual visits to that destination and was helpful in predicting future visits. Similarly, Preis, Moat & Stanley (2013) reported that changes in Google query volumes for terms related to finance, corresponded with actual changes in the stock market. These studies both used Google Trends[6], which returns data on the volume of worldwide Google searches that contain specified keywords. Values relating to the number of searches are provided for all weekly intervals over a defined period. All values are provided as relative percentages of the maximum value returned, across all queries and all weeks.

We wish to explore the use of search volume data as an indicator of general interest in, and the potential utility of, information about the relationship between particular life facets. We examine whether search volume data can be reliably used for filtering in lieu of subjective interest ratings from users. In practical terms, this means that PI systems could use emerging search trends as a basis for filtering correlations. We seek measures of the popularity of search queries that reference two different facets, thus reflecting general levels of interest in understanding or exploring the relationships between these facets. That is to say, we wish to capture the frequency of searches such as "does my sleep affect my mood?", "can listening to music make you more productive?", "using social media at the weekend", etc. In order to identify search terms that could represent each of the facets, two researchers independently generated a list of synonyms for each data category label (e.g. Physical activity: exercise, fitness, workouts, and so on). Each of the synonyms were then queried using the Google Trends service to determine their mean popularity score in the 10-year period from 2006-2016. The most popular term was subsequently selected as the representative term for that data category. Our decision to select a single term to represent a category arises from the varying number of synonyms associated with each of the categories, and the potential bias that may be introduced in using unequal numbers of search terms to collect data relating to each of the facets. Hence, each facet is represented by the single most popular term associated with it.

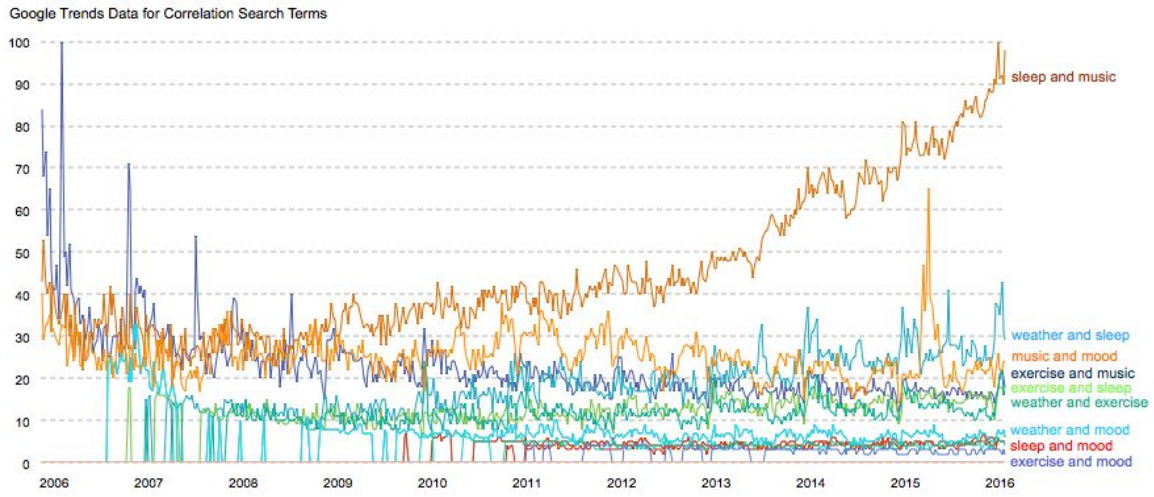
Next, queries were performed for each pairwise combination of the selected terms. For example, to assess the popularity of searches relating to the facets of physical activity and sleep, we used the query "exercise AND sleep". This provided quantitative data on the volume of searches that included both of the words 'exercise' and 'sleep', or minor variations thereof (e.g. exercising, exercised, sleeping etc.). That is to say, it incorporated searches like: "does exercise impact sleep", "sleeping after exercise", "can't sleep on days when I exercise". Since results may also include data for more obscure searches, which are not directly related to exploring the relationship between the included facets (e.g. "*exercise* my right to *sleep*"), we adapted our queries to avoid the inclusion of common 'nuisance' searches in the following way.

[6] https://www.google.com/trends/

For each query, the Google Trends service displays a list of the most common associated searches. Two researchers independently inspected and coded this list (Cohen's Kappa k=0.92 for intercoder reliability) to identify any potential nuisance searches that could be skewing the search volume results, i.e. topics which incorporated the facet-related search terms, but which were clearly not associated with exploring the relationships between the facets. For example, the query "weather AND music" yielded a popular related query: "weather channel music", which was associated with a music streaming service as part of the 'Weather Channel'[7]. In cases such as this, queries were modified with logical operators to exclude common nuisance searches (e.g. adapting the Google Trend query to become "(weather AND music) NOT channel").

Figure 6 visualises the results returned for our set of facet pair search queries. The average relative values within the defined period are also provided by the Google Trends service. We refer to this as the Mean Google Trend Score (MGTS). We also calculate the rank for each term, based on the MGTS, referred to as the Google Trend Rank (GTR). Figure 7 shows the MGTS for the top 10 ranking terms.

Figure 6. Google Trend Results for Facet Pair Queries



It is worth noting that Google Trends has an undisclosed minimum threshold for search volume, below which results are not returned. Hence, only the facet pair terms yielding results above this threshold were obtained. All other results were assigned an MGTS of 0 and the lowest possible GTR ranking.

Figure 7. Top Ranking Google Search Terms (Based on Mean Google Trend Score)

Rank (GTR)	Search Terms	Mean Google Trend Score (MGTS)
1	music and sleep	44.973
2	mood and music	25.227
3	exercise and music	23.043

[7] http://www.theweatherchannelmusic.com/

4	sleep and weather	16.259
5	exercise and sleep	10.884
6	exercise and weather	9.781
7	mood and weather	6.335
8	mood and sleep	2.685
9	busy and weather	2.385
10	music and social media	1.908

In the following section we examine the performance of machine learning classifiers, acting as information filters, with all of the previous objective data inputs used within the mixed effects model analysis (Correlation Characteristics and Data Categories), as well as the additional input measures of Diversity, Coverage, Peculiarity, Mean Google Trend Score and Google Trend Rank.

5.2. Analysis Method

Weka (version 3.6.13), an open source suite of machine learning algorithms (Frank *et al.*, 2005), was used to classify correlations according to their interestingness, using the aforementioned objective data. A comparison of correlation filtering performance was carried out between seven different supervised learning classifiers: Bayes Net, Decision Table, J48, Naïve Bayes, Random Forest and Random Tree, all of which are commonly used in recommender systems and predictive classification research (Hall *et al.*, 2009). We also report results from two different baseline classification approaches. The first, Baseline 1, is a naïve algorithm which assumes that every correlation in Exist (i.e. with a p-value < 0.2) should be classified as 'Interesting' and therefore shown to the user. This reflects the approach of the current Exist system. Hence, comparisons against Baseline 1 allow us to assess the potential benefits of automated filtering mechanisms in comparison to Exist. The second, Baseline 2, addresses the possibility that Exist may be inducing an information overload problem by setting its only filtering threshold (based on statistical significance) above the standard level of $p<0.05$. Hence, Baseline 2 measures the performance of a system that only presents correlations that are statistically significant at the $p<0.05$ level.

Since our classification algorithms can only predict the value of a categorical variable, our Interestingness measure was transformed into a categorical variable by assigning Likert values to one of two categories: 'Uninteresting' (for correlations that should not be shown to the user), and 'Interesting' (for correlations that should be shown to the user). Values 1 (Strongly Disagree), 2 (Disagree), and 3 (Neither Disagree nor Agree) were assigned to the Uninteresting category. Values 4 (Agree) and 5 (Strongly Agree) were assigned to the Interesting category.

We used 'leave-one-participant-out' cross-validation to perform all classification. This cross-validation approach holds out one participant's data as a testing set at each of 20 iterations, training the classifier on the remaining 19 participants' data. This approach simulates the real world task of predicting the interestingness of a new user's correlations from a model based on existing users, and prevents observations from a single participant from appearing in both the training and testing folds of the dataset. This is a

common performance evaluation method, used in similar classification research (e.g. Soleymani et al., 2012).

5.3. Classifier Results

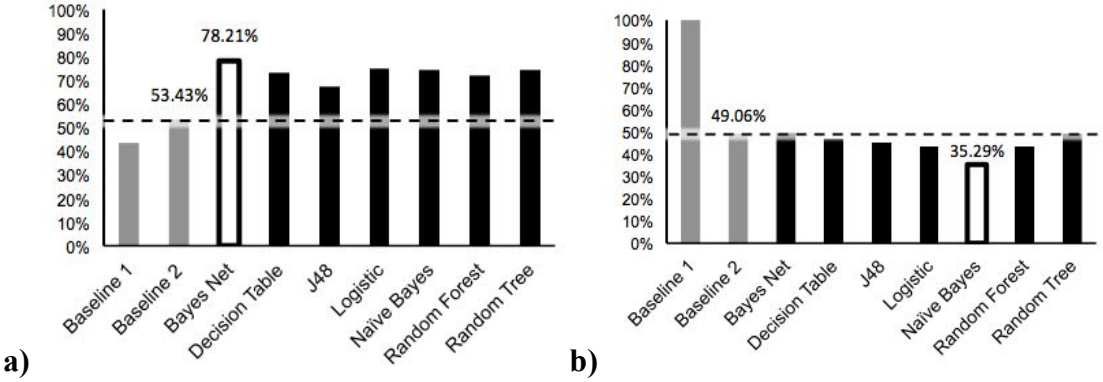
The results of the classifier testing for all of the algorithms are shown in Figure 8. This includes summary measures of classification performance and a breakdown for weighted true positive (recall), false positive, and precision means, across both of the Interesting and Uninteresting classes.

Figure 8. Machine Learning Classifier Performance Measures

		Baseline 1	Baseline 2	Bayes Net	Decision Table	J48	Logistic	Naive Bayes	Random Forest	Random Tree
	Correctly Classified Instances	0.435	0.534	**0.782***	0.744	0.748	0.733	0.674	0.720	0.744
	Incorrectly Classified Instances	0.565	0.466	**0.218***	0.256	0.252	0.267	0.326	0.280	0.256
	Kappa Statistic	0.000	0.066	**0.501***	0.460	0.456	0.429	0.338	0.410	0.484
	Mean Absolute Error	0.492	0.434	0.311	0.267	0.307	0.372	0.355	0.330	**0.258***
	Root Mean Squared Error	0.496	0.480	0.405	0.401	**0.386***	0.432	0.496	0.395	0.447
Weighted Average Accuracy	True Positive Rate / Recall	1.000	0.534	**0.771***	0.744	0.748	0.733	0.674	0.720	0.744
	False Positive Rate	1.000	0.462	0.273	0.273	0.275	0.265	0.306	0.306	**0.251***
	Precision	0.435	0.533	0.805	0.807	0.764	0.796	0.763	0.750	**0.809***
	F-Measure	0.408	0.533	**0.771***	0.744	0.735	0.736	0.672	0.712	0.748
	ROC Area	0.495	0.531	0.779	**0.869***	0.831	0.805	0.659	0.861	0.745

Figures 9a and 9b provide bar chart representations of the correct classification percentages and the proportion of correlations categorised as Interesting, respectively. Baseline measures are shown with grey bars. The best performing baseline level (in terms of highest accuracy and lowest number of correlations shown to users) is shown with a dashed horizontal line. The best performing classifiers are shown with white bars.

Figure 9 a) Correctly Classified Correlations b) Correlations Categorised as Interesting



a) **b)**

Our baseline results illustrate that the current Exist system has the lowest performance in terms of the proportion of correlations that are correctly classified as being of interest to the user. Figures 8 and 9a, show 43.5% of all correlations displayed by Exist (Baseline 1) were rated as interesting by our participants. Baseline 2, which displays only correlations at the $p<0.05$ level, provides a small improvement in correct classifications, with 53.4% being correctly identified as interesting or uninteresting.

While filtering on the basis of statistical significance (using the more commonly accepted threshold of $p<0.05$) reduces the likelihood of information overload by presenting fewer correlations, the resulting classification accuracy remains close to that of a random selection filter, or a system applying no filtering at all. Hence, there is significant room for improvement with regards to this approach.

Figures 8 and 9a show the F-Measure (F1-Score), ROC Area, and the percentage of correct classifications for both the Interesting and Uninteresting classes combined. Each of these measures gives a general indication of overall classification performance. However, in practical terms, only correlations classified as Interesting are actually shown to the user. Research has shown that, in many information retrieval settings, users care most about performance within the class of results that they actually see (van Rijsbergen, 1975; Kay, Patel & Kientz, 2015). Hence, the acceptability of accuracy in a retrieval setting is often more significantly influenced by *precision* within the results displayed (the proportion that users are actually interested in, amongst those shown), than the *recall* (the proportion of *all* interesting correlations within the data that are actually displayed) (van Rijsbergen, 1975). Our analysis therefore pays particular attention to performance measures within the Interesting class from this point forward.

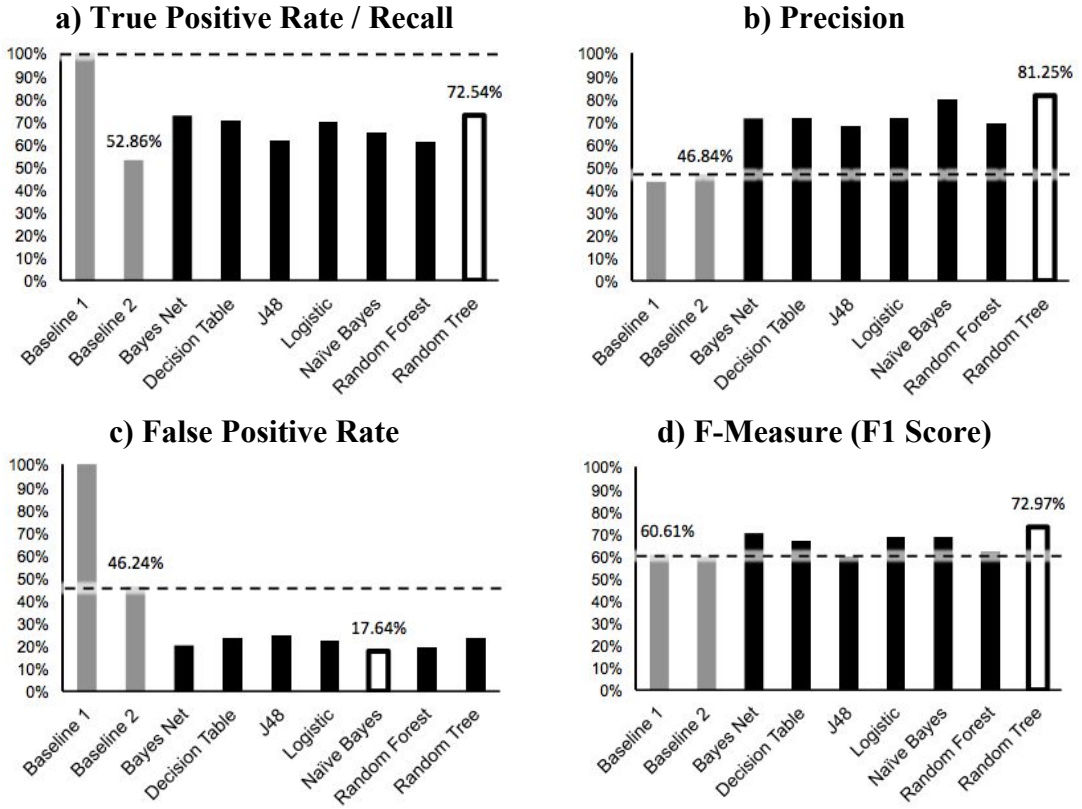
Figure 9b shows the proportion of all correlations assigned to the Interesting class by each of the ML classifiers. A mean average of 44.6% (*SD*=4.43%) of correlations are classified as Interesting across all classifiers. This means that in practice, applying these classifiers in a system such as Exist could reduce the number of correlations shown to the user by more than half (55.4%). The Naïve Bayes classifier provides the greatest reduction in information to be shown, however all other classifiers provide a similar reduction in the quantity of outputs, which are on a par with Baseline 2. For filtering to be beneficial, however, a reduction in quantity must also coincide with an increase in quality (classification accuracy).

Figure 10 shows the performance breakdown for correlations classified as Interesting. Figures 11 a,b,c illustrate the true positive rate (recall), precision and false positive rate results within the Interesting class, respectively.

Figure 10. Classifier performance breakdown for 'Interesting' class

		Baseline 1	Baseline 2	Bayes Net	Decision Table	J48	Logistic	Naive Bayes	Random Forest	Random Tree
Interesting Class Accuracy	True Positive / Recall	1.000	0.529	0.724	0.702	0.614	0.694	0.647	0.607	0.725
	False Positive	1.000	0.462	0.202	0.234	0.246	0.221	0.176	0.193	0.232
	Precision	0.435	0.468	0.713	0.713	0.678	0.715	0.796	0.690	0.813
	F-Measure	0.606	0.597	0.702	0.671	0.600	0.688	0.685	0.623	0.730
	ROC Area	0.506	0.570	0.790	0.805	0.696	0.831	0.869	0.861	0.745

Figure 11. Performance measures for Interesting class



Our results reveal precision scores of up to 81.3% within the Interesting class, using the ML classifiers (*M*=73.1%, *SD=4.8%)* (see Fig. 9b). That is to say, of the correlations presented to the user post-classification, 81.3% are actually interesting. The Random Tree classifier provides the highest precision value, which represents a 34.4 percentage point improvement over the best performing baseline (Baseline 2) and a 37.8 percentage point improvement over the current Exist approach (Baseline 1).

The true positive rate (recall) results (see Fig. 9a) reveal that our filtering approaches retrieve up to 72.5% of all interesting correlations within the data, such that they are displayed to the user (*M=67.4% SD=4.7%* for all classifiers). The Random Tree classifier also provides the best performance with respect to true positive rate. Since Baseline 1 offers no filtering, it provides a true positive rate of 100%, counteracted by a false positive rate also at 100% (see Fig. 9c). Baseline 2 offers the poorest true positive rate, retrieving 52.9% of all interesting correlations, whilst also incorrectly retrieving 46.2% of all uninteresting correlations (false positives). Using our classifiers, as few as 17.6% (Naive Bayes) of all uninteresting correlations within the data are falsely classified as interesting and shown to the user (*M=21.5% SD=2.3%* for all classifiers), at the same time as achieving improved true positive rates.

The results thus far relate to classifiers that include *all* available objective input data. Figure 12 provides results for the best performing classifiers (based on the Interesting class F-Measure), when trained and tested with each possible permutation of objective data as input, namely Correlation Characteristics (CC), Data Categories (DC), Interestingness Measures (IM), and Google Trends (GT). For example, a classifier with CC and GT only uses data relating to Correlation Characteristics and Google Trends.

Figure 12. Classifier performance with different data input permutations

	Classifier Input Data														
	CC	DC	IM	GT	CC DC	CC IM	CC GT	DC IM	DC GT	IM GT	CC DC IM	CC DC GT	DC IM GT	CC IM GT	All
Best Classifier	Logistic	Random Forest	Random Tree	Logistic	Logistic	Random Forest	Decision Table	Logistic	Logistic	Decision Table	Random Forest	Logistic	Bayes Net	Logistic	Random Tree
Interesting Class %	0.398	0.435	0.441	**0.366***	0.460	0.379	0.429	0.441	0.373	0.429	0.416	0.404	0.453	0.398	0.492
True Pos./Recall	0.586	0.657	0.671	0.614	0.714	0.543	0.671	0.671	0.629	0.671	0.657	0.657	0.686	0.686	**0.725***
False Positive	0.253	0.264	0.374	**0.176***	0.264	0.253	0.242	0.264	**0.176***	0.242	0.231	0.209	0.275	**0.176***	0.232
Precision	0.641	0.657	0.580	0.729	0.676	0.623	0.681	0.662	0.733	0.681	0.687	0.708	0.658	0.75	**0.813***
F-Score	0.612	0.657	0.623	0.667	0.694	0.580	0.676	0.667	0.677	0.676	0.672	0.681	0.671	0.716	**0.730***
ROC	0.707	0.767	0.667	0.790	0.775	0.720	0.715	0.764	0.796	0.715	0.781	0.802	0.670	**0.805***	0.745

The results presented in Figure 12 reveal that the classifier performs best with *all* data in terms of achieving a high true positive rate, precision and F-score, but that reasonable results are obtainable with reduced data input. Notably, Google Trends (GT) performs best of the single input classifiers, reducing the number of correlations to display to 36.6%, whilst achieving a true positive rate of 61.4%, and outperforming both baselines in terms of precision (72.9%) and false positive rate (17.6%). Hence, using Google Trends data alone can provide an improvement over conventional filtering approaches, such as filtering on the basis of $p<0.05$, with regards to presenting interesting insights. Given the potential for Google Trends data to be obtained for data categories beyond those within a particular PI system, we believe that this result offers a promising approach that can be generalised across a range of PI systems. It also has the potential to scale as the number of possible data inputs to a particular PI system increases. Overall, comparisons against our two baselines suggest that any of our classifiers, with any combination of the objective data inputs put forth, could provide value in terms of increasing the interestingness of correlational information presented to the user.

6. GENERAL DISCUSSION

In this paper we have reported findings from a study of a multifaceted personal informatics system designed for aggregation and analysis of personal data. One of the benefits of systems like Exist is that they provide users with a unique opportunity to investigate latent relationships between different aspects of their everyday lives. All of the participants within our study found some value in using Exist to reflect on their behaviour and lifestyle, and were generally positive about the overall system. Several of our participants encountered information that they intended to exploit beyond the confines of our study. These included attempts to increase productivity by listening to music whilst working [P8]; to curb a sedentary lifestyle by seeking new forms of indoor exercise [P3]; and reevaluating the management of their daily schedule so as to improve their sleep [P13]. Yet our study indicates that comprehension of the relationships between diverse personal data may be difficult in practice, and that this is due in part to feelings of information overload that stem from the large number of outputs that can be derived from personal data. We have explored the potential to reduce the likelihood of

overload by filtering information on the basis of 'Interestingness' to the user, given the context of exploratory use.

Our proposed filters were derived from qualitative and quantitative analyses, in which we found that interestingness was most strongly associated with correlations that were surprising, useful, statistically significant, linked to particular data categories, and which provided insights between, rather than within, life facets. We sought objective inputs that might capture aspects of surprisingness and utility; for example, because they exposed potentially surprising differences and peculiarities between users (i.e. generality, diversity, peculiarity), or because they identified particular combinations of facets which were often the focus of people's interest. The value ascribed to each of these features may be explained in different ways. Surprisingness and utility make sense given the context of exploratory use, in which users may be looking for outcomes that challenge their existing beliefs, or which help them to reflect on their behaviour in an unanticipated way. The association of statistical significance with interestingness may be because users value outputs that they feel confident about using—a strong indication of confidence may indicate that the correlation is "ready for use". Finally, the fact that correlations between different life facets were rated as especially interesting speaks directly to claims made elsewhere in the personal informatics literature regarding the potential value of multifaceted tracking (e.g., Li, Forlizzi & Dey, 2010; Bentley *et al.*, 2013; Dingler, Sahami & Henze, 2014; Haddadi & Brown, 2014). However, it is worth noting that a correlation does not necessarily require all of these attributes to be of interest; for example, an outcome may be seen as surprising and useful without necessarily being statistically significant. Hence, filtering on the basis of a single criterion, e.g. p-value, is unlikely to be as effective as considering multiple criteria.

One of the main contributions of our study is a demonstration that, despite the insights provided by Exist being highly individualised, and interest in these insights also varying between individuals, it is possible to develop filtering algorithms which have the capability to identify interesting information across a diverse set of users. The present study also demonstrates that such filtering can be done to a reasonable degree of accuracy. One of the practical benefits of our approach is that the mechanisms driving these filtering algorithms require no explicit input from users in order to function. Since our proposed filters leverage objective, machine interpretable data that is either already present within the personal informatics system (e.g. statistical significance) or which is easily accessible through interaction with external services (e.g. Google Trends), it will be possible for designers of personal informatics systems to make use of these algorithms without the labour-intensive task of collecting data from users. The objective data that we identified is not specifically tied to Exist, and hence it is likely that our findings could be applied to similar PI systems, and scaled-up as the number of inputs expands.

While the accuracy of our classifiers is not perfect, we see our work as a positive step towards the algorithmic curation of PI system outputs, which we believe will mitigate the likelihood of information overload and improve the overall experience of exploratory use. As a practical example of this, use of the filters we have proposed, for the average participant in our study, would result in a reduction from approximately 62

correlations to 27, with many of these 27 (~81%) likely to be amongst the most "interesting".

An interesting finding from our study relates to the extent to which Exist revealed information that was already known by its users. Approximately half (46.7%) of the correlational information provided by the system was predicted by participants ahead of the study. The rest either provided information that refuted earlier predictions (37.8%), or that provided new information which the participant had been unable to make a prediction about (16.5%). There is likely to be an expectation from users that information provided by systems is informative, contributing to their knowledge in a significant way. However, not all known information was considered uninteresting. Users found value in both known and unknown information, either for corroborating or extending their existing self-knowledge, respectively (cf. Choe et al., 2015). In light of this finding, filtering on the basis of user interest, rather than on existing user knowledge, appears to be justified. Furthermore, we have shown that predicting interest is feasible with minimal explicit input from users, the same of which cannot currently be said for predicting existing user knowledge.

6.1. The growing need for filters

We believe that the need for filtering becomes ever more pressing as designers of PI systems seek to expand the facilities they provide for tracking data. For instance, after our study was completed, Exist introduced the ability to track data about *Food, Drink*, *Finance*, *Health*, *Relaxation* and *Environment*. Additional attributes relating to existing facets have also been added: the *Physical Activity* category has been augmented with cycling distance and time*; Productivity* now includes measures of individual application usage, emails written, and software commits made; and artist-specific listening counts have been added to *Music*. The additions of application- and artist-specific data in particular means that Exist now has the ability to detect increasingly nuanced correlations, such as "you have a better mood when you listen to Red Hot Chili Peppers", or "you spend more time on Facebook when you have fewer events". Whilst these additions have the potential to reveal more fine-grained correlational information, they are also likely to exacerbate information overload problems by vastly increasing the volume of information produced by the system, particularly for users who listen to many different artists and use many different software applications. We believe that, as the number of possible inputs to personal informatics systems increases, designers will require an ability to understand which information users are most interested in seeing, and will need access to mechanisms that can support the filtering of this information accordingly. Filters like those we have proposed will give designers an immediate and practical outlet for addressing these needs.

6.2. Filtering trade-offs

It is important to note that information filters give rise to trade-offs when deciding what to present to users. Delivering results that are of 'known' interest to the user may come at the expense of serendipitous discoveries, and may discourage manual exploration of correlations for which the value provided to the user may be less

immediate. Although our results show that statistically significant correlations were generally deemed more interesting, filtering on the basis of p-value alone was not enough to effectively isolate interesting correlations. Comments made by our participants at interview provided evidence of cases in which more 'spurious' correlations were interesting. For example, when viewing a correlation between weather and productivity with p=0.09 ("you are more productive on days when it is windy") a participant told us: "*That's fascinating. Seems a bit odd though. I'd like to collect more data to see if the correlation stays the same*" [P9]. Thus, by increasing the selectivity over information to show to the user, it is possible that a system may lose some positive characteristics, including those which motivate users to engage in additional data collection and make further use of the system through exposure to new information (Pariser, 2011). However, if spurious correlations are to be presented to the user, we advise that statistical confidence is communicated clearly so that users are nudged towards monitoring, rather than trusting, the emerging relationship.

Filtering is also known to raise concerns about which of the involved parties should have the power to control the display of personal data. For example, one study of Facebook showed that users reacted negatively upon discovering that the automated filtering of newsfeeds resulted in some 'Friends' disappearing altogether from the content that they could see (Pegoraro, 2011). It is possible that similar concerns might be raised with respect to the mechanisms that we have explored within this paper. One implication, therefore, is that the ability to algorithmically curate outputs in PI systems should not necessarily eliminate the option of viewing all results without filtering applied. In future designs that build on our work, we envisage a two-tier solution in which filtered content is shown to the user at the point of entry to the system, supporting a glanceable view of the most interesting content. Then, the system could contain a second "repository" layer in which all of the user's correlations are available for analysis. A final consideration for designers should be to make the presence of filtering transparent, because user experience problems with algorithmic filtering and curation often stem from inaccurate beliefs about these automated system processes (Rader & Gray, 2015). Since systems like Exist act on data that is inherently "personal", i.e. it is both about the user and belongs to them, users may experience discomfort at the thought of any tampering with their view of data, making the need for explanations of algorithmic filtering to be particularly important. In our study we noticed that some participants paid attention to which correlations were missing from the output and inferred some meaning about the relationships between life facets, based on their absence. Therefore we advise that the use of algorithmic information filtering should be made explicit and the criteria on which filtering is performed made transparent, such that users can consider other possible reasons for the absence of certain correlations. Again, designers may wish to include an option to disable filtering, should users wish to explore the data in full.

6.3. Further improvements to filtering mechanisms

Our classifiers operate on objective data inputs that circumvent the need for subjective data from users. Thus, we sought to avoid information overload without

requiring additional user effort. We were able to identify *Interestingness* measures and *Google Trends* data which provided reasonable proxies for certain qualities associated with interesting correlations. Future work looking to provide further improvements to the filtering mechanisms proposed within this paper could focus on identifying additional sources of objective data, which might reflect further qualities described in Section 4.3. For example, web browsing history data (Teevan, Dumais & Horvitz, 2005), app usage patterns (Jones et al., 2015), and social media interactions (Kosinski, Stillwell, & Graepel, 2013) could be used to infer a user's specific life interests.

Our filtering approach demonstrates the potential to support users in dealing with information overload, despite the highly individual nature of correlation outcomes. This was previously reported by Bentley et al. (2013) and is further corroborated by the results of our study. However, several participants commented that they expected definitions of 'interesting' to also differ between individuals to some extent: *"I think each person will have a different answer about what's interesting for them." [P2], "Somebody might say I'm interested in productivity, whereas somebody else says I'm interested in steps." [P6]*. While our general model of users' interest fits our data for the participants included in our study, improvements in the performance of the filtering mechanisms may be achievable by offering a greater degree of individual control, for example by permitting individual users to select certain criteria to be included or excluded from the filtering process.

The classifier we have proposed was trained on subjective interestingness ratings from a diverse set of users, but it is also possible that a unique classifier could be shaped for each individual user, learning through a feedback loop in which users 'tag' the types of correlations that they find particularly interesting. A system could then show the user more of the correlations that they like, based on the data they have provided. Alternatively, this filtering might be collaborative, using algorithms such as those outlined by Herlocker, Konstan, Borchers & Riedl (1999), to profile and group users based on their similarity. Unique classifiers could then be trained for each group, with information indicating the nature of interesting correlations being pooled amongst its members.

Although the development of our automated filtering approach is primarily motivated by the problem of information overload, we believe that the application of filtering mechanisms may also be relevant in situations where overload is not necessarily a problem. For example, they could be used for prioritisation even within manageable quantities of information. We note that the Mobile Health Mashups system (Tollmar *et al.*, 2012) provided a daily feed of new correlational information for its users, together with notifications when new correlations were uncovered, delivering insights in a more easily digestible form. Prioritising these notifications may be beneficial, and could help to avoid interrupting a user when information is less likely to be interesting. We expect that such an approach may contribute to longevity in use and engagement with a PI system, since the value provided may be consistently higher. Furthermore, while a single service may not provide enough information to overwhelm its users, many services naturally compete against other information sources for users' attention. Filtering within

each of these services may help to alleviate a more general load that results from interaction with multiple services.

It is because of the transformative potential of personal informatics systems that there is an increasing emphasis on their use in health and wellbeing contexts. We consider the prevention of information overload to be especially important in such contexts, since problems in dealing with the insights provided might undermine the significant benefits that a system can provide. In particular, the negative impact of information overload on the sensemaking process, which is consonant with the so-called 'analysis paralysis' that one of our study participants described, raises significant concerns in a context where inappropriate decisions and incorrect courses of action could cause physical or mental harm (Barton, 2012). Whilst we have shown that automated filtering is possible, we believe there may be significant risks associated with a dependence on information filtering in these contexts. Our approach to filtering is founded on users' desire to see 'interesting' insights. However, "interesting" and "important" are not always interchangeable. It is likely that health and wellbeing situations necessitate the display of information which users *should* or *must* see for their own benefit, as opposed to only information that they *want* to see. In contexts where certain information is critical—for example, where a correlation indicates that something is having an adverse effect on health—alternative filtering criteria are likely to be required. A system in this context might additionally incorporate expert knowledge (e.g. from clinicians and health professionals) as part of the filtering process in order to flag correlations that, if present, should categorically not be ignored.

6.4. Limitations

One possible limitation of our study is that we required participants to view all of their correlations in a single session. This is intended to represent the experience of encountering a body of personal data for the first time, yet it may not reflect the way in which users generally interact with PI systems. At the time of writing, Exist employs a menu system to subdivide outputs by data category (i.e. one page per category). In practice, users may take a more episodic approach to viewing the information provided, accessing the system multiple times and viewing fewer correlations within each session. However, this does not necessarily dispel the problem of overload because users are still faced with the task of interpreting their data and determining what is interesting. Therefore, we argue that a filtering mechanism would be useful for increasing the efficiency of these shorter sessions by guiding exploration to insights that are more likely to be of interest, increasing the value derived from each session and motivating the user to return to the system.

A further limitation relates to the $p<0.2$ filter that Exist applied to all of the data we were able to obtain. It is possible that some interesting (but not statistically significant) relationships were filtered out by this threshold and were, as a result, not examined in our qualitative or quantitative studies, or used to train or test the classifiers.

Ongoing use of a system such as Exist raises several questions that have not been addressed by our study. Our filtering mechanisms were founded upon the identification

of factors that affected users' initial impressions of the correlations they were shown. It is possible that users' assessments of the correlations may have been influenced by their overall experiences with tracking technologies during the study. Furthermore, it is likely that the Interestingness of a correlation exhibits a temporal component, or decay factor. That is to say, what may be considered interesting now may not necessarily be interesting at a later time. This seems particularly likely when Interestingness is assessed predominantly in terms of ability to surprise. We speculate that users of Exist may not be equally surprised by the same information when re-encountering it in a subsequent interaction with the system. Hence, filtering algorithms may wish to account for users' continuing use of the system, perhaps by downgrading correlations that have already been seen, and prioritising new correlations that are classified as interesting by the system. In a similar vein, there is a need to recognise that user interests may become increasingly "personal" as they develop specific tracking goals. The use of a filter might therefore need to be adjusted to accommodate a transition from exploratory to esoteric use. Future work should delve deeper into the issues associated with filtering in practice, by examining them in the hands of users in realistic settings. Observing user interactions with a multifaceted personal informatics system may also provide more reliable data on users' interest in correlational information, by capturing actual interactions and views of the information, rather than by obtaining self-reported estimates of their interest from questionnaires.

Finally, we found that multifaceted correlations are often more interesting than unifaceted correlations. However, this may be limited by the data collected in our study. We are able to think of cases in which unifaceted correlations might produce interesting insights, e.g. subjective rating of sleep quality versus that recorded by a sensor, particularly if such an insight counters intuitive beliefs. This should motivate future work which looks more closely at specific aspects of a user's life, and the surprising insights that might be drawn from a wider range of data sources.

7. CONCLUSION

This paper builds on previous work which has indicated the prospective value of, and desire for, multifaceted personal informatics systems that uncover latent associations between different life facets, but which has not explored the challenge of dealing with information overload in this context. The contributions of this paper include new insights into problems of information overload in personal informatics systems, and an understanding of the value that multifaceted systems provide; including what constitutes interesting information, given the context of exploratory use. We have used this understanding to design filtering mechanisms that algorithmically curate information outputs that are of interest to users, whilst simultaneously alleviating information overload. We believe that these contributions have the potential to support the uptake of PI systems in the real world.

Although filtering is a common mechanism for dealing with information overload in many contexts, other approaches, such as training users to interpret data more effectively, or presenting information in alternative forms, such that it is easier to digest,

could be explored in future work, and assessed in terms of their ability to reduce information overload and support sensemaking of personal data.

Notes

Acknowledgments. We would like to thank Edward Bradley for his support with data collection and early pilot work. We are also very grateful to all of the participants involved in our research for generously sharing their time.

References

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large databases, VLDB* (Vol. 1215, pp. 487-499).

Barton, A. J. (2012). The regulation of mobile health applications. *BMC medicine*, 10(1), 46.

Bawden, D., & Robinson, L. (2008). The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of Information Science*, 35(2), 180–191.

Belkin, N. J., & Croft, W. B. (1992). Information filtering and information retrieval: Two sides of the same coin?. *Communications of the ACM*, *35*(12), 29-38.

Bentley, F., Tollmar, K., Stephenson, P., Levy, L., Jones, B., Robertson, S. & Wilson, J. (2013). Health Mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behavior change. *ACM Transactions on Computer-Human Interaction* (TOCHI), 20(5), 30.

Bernstein, M., Hong, L., Kairam, S., Chi, E., & Suh, B. (2010) A torrent of tweets: Managing information overload in online social streams. In *Workshop on Microblogging: What and How Can We Learn From It?* Held at the ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '10.

Chewning, E. G., & Harrell, A. M. (1990). The effect of information load on decision makers' cue utilization levels and decision quality in a financial distress decision task. *Accounting, Organizations and Society*, *15*(6), 527-542.

Choe, E. K., Lee, N. B., Lee, B., Pratt, W., & Kientz, J. A. (2014). Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 1143-1152). ACM.

Choe, E. K., Lee, B., & Schraefel, M. C. (2015). Characterizing Visualization Insights from Quantified Selfers' Personal Data Presentations. *Computer Graphics and Applications, IEEE*, *35*(4), 28-37.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, *88*(s1), 2-9.

Chung, C. F., Cook, J., Bales, E., Zia, J., & Munson, S. A. (2015). More than telemonitoring: Health provider use and nonuse of life-log data in irritable bowel syndrome and weight management. *Journal of medical Internet research*, *17*(8), e203.

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research.* Sage, London, UK.

Cosley, D., Akey, K., Alson, B., Baxter, J., Broomfield, M., Lee, S., & Sarabu, C. (2009). Using technologies to support reminiscence. In *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology* (pp. 480-484). British Computer Society.

Cuttone, A., Petersen, M. K., & Larsen, J. E. (2014). Four Data Visualization Heuristics to Facilitate Reflection in Personal Informatics. In *Universal Access in Human-Computer Interaction. Design for All and Accessibility Practice* (pp. 541-552). Springer International Publishing.

Dabbish, L. A., Kraut, R. E., Fussell, S., & Kiesler, S. (2005). Understanding email use: predicting action on a message. *In Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 691-700). ACM.

Dingler, T., Sahami T. & Henze, N. (2014). There is More to Well-being than Health Data: Holistic Lifelogging through Memory Capture. In *Proceedings of the CHI Workshop on Beyond Quantified Self: Data for Wellbeing, 2014.*

Elsden, C., Kirk, D. S., & Durrant, A. C. (2015). A Quantified Past: Toward Design for Remembering With Personal Informatics. *Human–Computer Interaction*, 1-40.

Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The Information Society,* 20(5), 325-344.

Exist. (2016). https://exist.io Last accessed: 31 March 2016

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L. (2005). Weka. In *Data Mining and Knowledge Discovery Handbook* (pp. 1305-1314). Springer US.

Galbraith, J. R. (1974). Organization design: An information processing view. *Interfaces* 3:28–36.

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, *38*(3), 9.

Gurrin, C., Smeaton, A. F., & Doherty, A. R. (2014). Lifelogging: Personal big data. *Foundations and trends in information retrieval*, *8*(1), 1-125.

Haddadi, H., & Brown, I. (2014). Quantified self and the privacy challenge. *Technology Law Futures*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.

Hallowell, E. M. (2005). Overloaded circuits: Why smart people underperform. *Harvard business review*, *83*(1), 54-62.

Hanani, U., Shapira, B., & Shoval, P. (2001). Information filtering: Overview of issues, research and systems. *User Modeling and User-Adapted Interaction*, 11(3), 203-259.

Hello Code. (2015). *2015 in Review.* http://blog.hellocode.co/post/2015-review/ Last accessed: 31 March 2016

Herlocker, J. L., Konstan, J. A., Borchers, A., & Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 230-237). ACM.

Hilderman, R. J., & Hamilton, H. J. (2001). Evaluation of interestingness measures for ranking discovered knowledge. In *Advances in Knowledge Discovery and Data Mining* (pp. 247-259). Springer Berlin Heidelberg.

Hiltz, S. R., & Turoff, M. (1985). Structuring computer-mediated communication systems to avoid information overload. *Communications of the ACM*, 28(7), 680-689.

Huckvale, K., Car, M., Morrison, C., & Car, J. (2012). Apps for asthma self-management: a systematic assessment of content and tools. *BMC medicine*, 10(1), 144.

Jones, S. L. (2015). Exploring correlational information in aggregated quantified self data dashboards. *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 (UbiComp/ISWC'15 Adjunct)*. ACM, New York, NY, USA, 1075-1080.

Jones, S. L. & Kelly, R. (2016). Sensemaking Challenges in Personal Informatics and Self-Monitoring Systems. In *Proceedings of WISH 2016: Workshop on Interactive Systems in Healthcare*. CHI 2016. 2016-05-07, San Jose. ACM.

Jones, S. L., Ferreira, D., Hosio, S., Goncalves, J., & Kostakos, V. (2015). Revisitation analysis of smartphone app use. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 1197-1208). ACM.

Karkar, R., Fogarty, J., Kientz, J. A., Munson, S. A., Vilardaga, R., & Zia, J. (2015). Opportunities and challenges for self-experimentation in self-tracking. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers* (pp. 991-996). *ACM.*

Kay, M., Choe, E. K., Shepherd, J., Greenstein, B., Watson, N., Consolvo, S., & Kientz, J. A. (2012). Lullaby: a capture & access system for understanding the sleep environment. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 226-234). ACM.

Kay, M., Patel, S.N., and Kientz, J.A. How Good is 85%? A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* - CHI '15, ACM Press, 347–356.

Keller, K. L., & Staelin, R. (1987). Effects of quality and quantity of information on decision effectiveness. *Journal of Consumer Research*, 14, 200–213.

Kelly, R., & Payne, S. J. (2014). Collaborative web search in context: a study of tool use in everyday tasks. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 807-819). ACM.

Koroleva, K., Krasnova, H., & Günther, O. (2010) 'STOP SPAMMING ME!' - Exploring Information Overload on Facebook. *Proceedings of the Americas Conference on Information Systems*, paper 447. http://aisel.aisnet.org/amcis2010/447

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802-5805.

Lara, O. D., & Labrador, M. A. (2013). A survey on human activity recognition using wearable sensors. *Communications Surveys & Tutorials, IEEE*, *15*(3), 1192-1209.

Lee, S., Kim, S.-H., Hung, Y.-H., Lam, H., Kang, Y.-a. & Yi, J. S. (2016). How do People Make Sense of Unfamiliar Visualizations?: A Grounded Model of Novice's Information Visualization Sensemaking. *Visualization and Computer Graphics, IEEE Transactions on* 22, no. 1, 499-508.

Li, I., Forlizzi, J., & Dey, A. (2010). Know thyself: monitoring and reflecting on facets of one's life. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems* (pp. 4489-4492). ACM.

Li, I., Dey, A., & Forlizzi, J. (2010). A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 557-566). ACM.

Li, I., Dey, A. K., & Forlizzi, J. (2011). Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th international conference on Ubiquitous computing* (pp. 405-414). ACM.

Li, I. (2011). Personal Informatics and Context: Using Context to Reveal Factors that Affect Behavior. PhD Thesis, Carnegie Mellon University, Pittsburgh. Retrieved from http://ianli.com/thesis/

Lupton, D. (2013). Understanding the human machine [Commentary]. *Technology and Society Magazine, IEEE*, *32*(4), 25-30.

Mamykina, L. A., Smaldone, M. & Bakken, S. R. (2015). Adopting the sensemaking perspective for chronic disease self-management. *Journal of Biomedical Informatics* 56, 406-417.

Mano, R. S., & Mesch, G. S. (2010). E-mail characteristics, work performance and distress. *Computers in Human Behavior*, 26(1), 61-69.

Oulasvirta, A., Hukkinen, J. P., & Schwartz, B. (2009). When more is less: the paradox of choice in search engine use. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 516-523). ACM.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Patel, S., Park, H., Bonato, P., Chan, L., & Rodgers, M. (2012). A review of wearable sensors and systems with application in rehabilitation. *Journal of neuroengineering and rehabilitation*, *9*(1), 1.

Paul, S. A., & Morris, M. R. (2009) CoSense: enhancing sensemaking for collaborative web search. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09). ACM, New York, NY, USA, 1771-1780.

Pegoraro, R. (2011). Facebook News Feed filtering can make friends vanish. *The Washington Post*.

Petersen, D., Steele, J., & Wilkerson, J. (2009). WattBot: a residential electricity monitoring and feedback system. In *Proc. 27th international conference extended abstracts on Human factors in computing systems*, Boston, MA, USA, 2847-2852.

Phillips, J. K., & Battaglia, D. (2003). Instructional methods for training sensemaking skills. *Phillips, JK, Klein, G., & Sieck. WR (2004). Expertise in judgment and decision*

making: a case for training intuitive decision skills. In DK Koehler and N. Harvey (Eds.). Blackwell Handbook of Judgment and Decision Making. Wiley-Blackwell.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific reports*, *3*.

Rader, E., & Gray, R. (2015). Understanding user beliefs about algorithmic curation in the Facebook news feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 173-182). ACM.

Rapp, A., & Cena, F. (2014). Self-monitoring and technology: challenges and open issues in personal informatics. In *Universal Access in Human-Computer Interaction. Design for All and Accessibility Practice* (pp. 613-622). Springer International Publishing.

Rapp, A., & Cena, F. (2016). Personal informatics for everyday life: How users without prior self-tracking experience engage with personal data. International Journal of Human-Computer Studies, 94, 1-17.

Ricci, F., Rokach, L., & Shapira, B. (2011). *Introduction to recommender systems handbook* (pp. 1-35). Springer US.

van Rijsbergen, C.J. (1975). Evaluation. In Information Retrieval. Butterworth & Co., 1975, 95–132.

Rooksby, J., Rost, M., Morrison, A., & Chalmers, M. C. (2014). Personal tracking as lived informatics. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 1163-1172). ACM.

Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The cost structure of sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems* (pp. 269-276). ACM.

Savolainen, R. (2007). Filtering and withdrawing: strategies for coping with information overload in everyday contexts. *Journal of Information Science*, *33*(5), 611-621.

Schick, A. G., Gordon, L. A., & Haka, S. (1990). Information overload: A temporal approach. *Accounting, Organizations and Society*, 15(3), 199-220.

Schneider, S. C. (1987). Information overload: Causes and consequences. *Human Systems Management*, 7(2), 143-153.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: happiness is a matter of choice. *Journal of Personality and Social Psychology*, *83*(5), 1178.

Shah, N. H. (2015). Using Big Data. In *Translational Informatics* (pp. 119-128). Springer London.

Shapira, B., Hanani, U., Raveh, A., & Shoval, P. (1997). Information filtering: A new two-phase model using stereotypic user profiling. *Journal of Intelligent Information Systems*, *8*(2), 155-165.

Soleymani, M., Pantic, M., & Pun, T. (2012). Multimodal emotion recognition in response to videos. IEEE transactions on affective computing, 3(2), 211-223.

Statista.(2016). http://www.statista.com/statistics/259372/wearable-device-market-value/

Stone, L. (2008). Continuous partial attention–not the same as multi-tasking. *Business Week*, 24.

Swan, M. (2013). The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. *Big Data,* 1 (2): 85-99.

Sweeny, K., Melnyk, D., Miller, W., & Shepperd, J. A. (2010). Information avoidance: Who, what, when, and why. *Review of General Psychology*, *14*(4), 340.

Teevan, J., Dumais, S. T., & Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 449-456). ACM.

Tollmar, K., Bentley, F., & Viedma, C. (2012). Mobile Health Mashups: Making sense of multiple streams of wellbeing and contextual data for presentation on a mobile device. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2012 6th International Conference on* (pp. 65-72). IEEE.

Weick, K. E. (1995). *Sensemaking in organizations* (Vol. 3). Sage.

Whittaker, S. (2008). Making sense of sensemaking. In *HCI remixed: Reflections on works that have influenced the HCI community*. MIT Press, Boston, MA, USA.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wolf, G. The data-driven life. *The New York Times*, 28, (2010).

Zhong, N., Yao, Y. Y., & Ohishima, M. (2003). Peculiarity oriented multidatabase mining. *Knowledge and Data Engineering, IEEE Transactions on*, *15*(4), 952-960.