Decades of data: will it be useless?

Lie Ming Tang
University of Sydney
Sydney, Australia
ltan8012@uni.sydney.edu.au

Bob Kummerfeld
University of Sydney
Sydney, Australia
bob.kummerfeld@sydney.edu.au

Judy Kay
University of Sydney
Sydney, Australia
judy.kay@sydney.edu.au

Abstract
Personal sensors allow people to effortlessly amass data such as step count, heart-rate and sleep patterns. Decades of this data ought to be useful. We identify questions the data should be able to answer and the challenges we need to address.

Author Keywords
Authors' choice; of terms; separated; by semicolons; include commas, within terms only; required.

ACM Classification Keywords
H.5.m [Information interfaces and presentation]: Miscellaneous; See [http://acm.org/about/class/1998/]

CHI'18, April 26–May 1, 2018, Montreal, Canada.

DOI string from ACM form confirmation

Introduction and background

Consumer activity trackers have now been widely available for a decade. It is now timely to consider how long term data can be useful.

The next section characterises important questions that physical activity sensor data should help people answer. Then we discuss the challenges yet to be overcome if people are to be able to harness the decades of data they are amassing to answer those questions.

What is long term data about physical activity be useful for?

We identify three important classes of questions and illustrate each with examples for the case of physical activity tracking. In the final section, we discuss how this relates to other sensor data types.

Learning about trends & patterns

This is the simplest class of question – it enables a person to discover trends over the long term, answering questions like these:

- Have I become less active now than I was 10 years ago?
- Am I generally more active on weekends?
- The newest advice is to get 60 active minutes a day - have I been achieving that?

Personal Hypothesis

Long term data should also enable people validate beliefs about factors affecting their level of activity. They may have also tried various strategies to achieve long term goals.

- I believe that when I moved to the suburbs a year ago, I became more active.
- Taking public transport to work (as I do on Mondays) means I am more active than the 3 years I had gym membership period.

Remembering & Reminiscence

We also see potential for benefits from reminiscence around personal long term health data as in a study of 15 long-term users [2] which reported the ways that participants interacted with their data with storytelling and reminiscence. In this case, the question is: what memories does my data evoke?

Key challenges and research directions

If people are to be able to answer questions like those above, there are several challenges. We first describe core interface challenges for exploring and understanding long term data, then technical challenges and their associated interface challenges.

Interfaces for interpreting data

One key challenge is to make the data available in a suitable form. For example, we created a calendar visualization [8] like that shown in Figure 1. User studies indicated that this interface enabled people to answer all three classes of questions above. It also highlighted some of the challenges we describe below, by referring to the figure. It shows data for a hypothetical user Alice, who had collected physical activity data over 10 years, from 2015 - 2024. The dark cells indicates days she met her 10K steps goal, light cells 50% and white cells below 50%. The grey cells are days with no data recorded. The change in cell colour after 2016 reflects when she changed device (from a Fitbit). Superimposed on the figure are icons indicating key challenges in Alice's life and context. Alice started tracking in 2015 as part of a University study where she had a Fitbit Zip. At the time, she was physically active and participated in many student activities. In 2017, she began her working career, entered a serious relationship and became more conscious of her weight gain and lower physical activity. So, she bought her own Android wear smart watch device along with Google

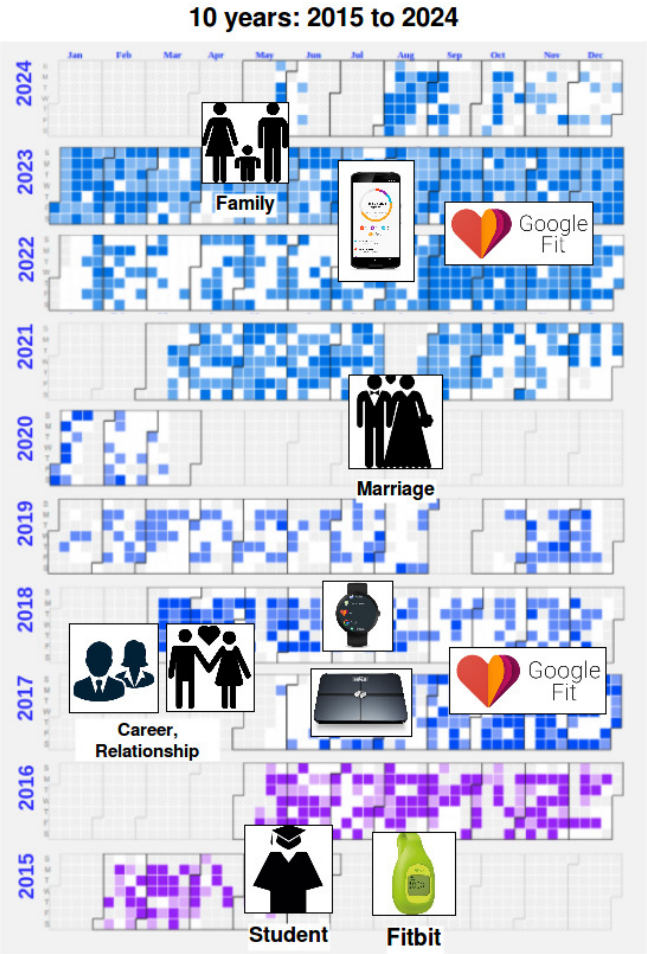

Figure 1: Illustrative overview of physical activity data (e.g., steps) of a hypothetical user, Alice, over 10 years, starting in 2015. Calendar chart view (steps / day): dark blue indicate reaching her goal (e.g., 10K steps); light blue 50% of goal, white less than 50%; grey means no data. Icons (overlaid) to indicate major life stages (e.g., student, marriage) and devices and vendor used (e.g., Fitbit, Google Fit, Smart watch, Smart Scale and Fitbit Zip) at different times.

Fit app and a smart scale. In 2020, her life changed dramatically with marriage and then in 2023 starting a family. From 2019, she stopped using her smart watch and scale and relied solely on the Google Fit app to keep track of her general health and physical activity.

Interpreting data - impact of data incompleteness

Alice's data shows the impact of days she did not wear her tracker. This is unsurprising and has been documented among long term trackers [8, 3, 7]. and considering decades, data is probably collected across different devices, tracking technologies and vendors. Figure 1 makes it easy to see whole days that had no data (the grey cells) but not the impact of wear time (hours per day). For example, if she only wore a tracker for 3 hours, this may give a white cell - but is 3-hours enough for the step count to be meaningful for questions about activity level? Interfaces need to account for this to avoid compromising user trust [8]. We need further work on how to ensure people can get meaningful and trustworthy answers to their questions, where the information available to the user include ways to assess the accuracy, accounting for incompleteness in people's long term data.

Interpreting data - scaffolding reflection

While this calendar visualisation proved quite effective for answering questions about physical activity, it highlighted the need for scaffolding to help users consider good questions. For example, many people are less active on weekends than weekdays without realising this [8]. A scaffolding interface could help people consider questions that would help them explore their data. Similarly, it could help people formulate hypotheses to explore.

Combining data from different stores

In the long term, even activity tracking may mean that data is stored by various vendors. This can make it very diffi-

cult for users to aggregate all their data and pose a serious problem as devices, companies and technologies evolve. It is not even clear if users own their own data and practically users have very little control over it. For example, users may not know what data is stored (steps/minute versus steps/day) and granularity and detail available is at the vendor's discretion.

Data privacy, security, provenance & management
Other major barriers relate to security and privacy [4, 1]. If data is centralised, this makes it easier to analyse but it becomes vulnerable. To make sense of data, it needs to be associated with the sources, in terms of who or what generated the data. If the data is shared, users should also be able to choose for example who can access it, for what purpose, when and for how long. Kuo et al [5] proposed using blockchain technologies as a promising approach to decentralising trust, managing data provenance and granular control of what and how personal data is used (e.g., smart contracts).

User interface to control & manage data
All the technical challenges of storage and privacy management have parallel interface challenges. Studies in online social network users point to how difficult these are, indicating the mismatch between intention and actual settings, for example one study revealed that privacy settings matched users' expectations only 37% of the time [6].

Conclusions and research agenda

Long term personal sensor data has the potential to enable people to answer important questions. We have discussed some of these for the case of physical activity data. There are corresponding and broader questions for the many other data types, such as the already widely tracked aspects such as sleep and heart-rate. These are likely to be even more sensitive than step data. We have outline the need for further work to identify the questions people should be able to answer, the interfaces that will enable them to do so and the infrastructure to support them in managing their data as they wish.

REFERENCES

1. Tim Althoff. 2017. Population-Scale Pervasive Health. *IEEE Pervasive Computing* 16, 4 (2017), 75–79.
2. C Elsden, D.S. Kirk, and Abigail C. Durrant. 2015. A Quantified Past: Towards Design for Remembering with Personal Informatics. *HCI* (2015).
3. D A. Epstein, J Kang, L R. Pina, J Fogarty, and S A. Munson. 2016. Reconsidering the Device in the Drawer: Lapses as a Design Opportunity in Personal Informatics. In *Ubicomp*. ACM.
4. J Kay and B Kummerfeld. 2012. Creating personalized systems that people can scrutinize and control: Drivers, principles and experience. *TiiS* (2012).
5. T T. Kuo, H E. Kim, and L Ohno-Machado. 2017. Blockchain distributed ledger technologies for biomedical and health care applications. *Journal of the American Medical Informatics Association* (2017).
6. Y Liu, K P. Gummadi, B Krishnamurthy, and A Mislove. 2011. Analyzing Facebook Privacy Settings: User Expectations vs. Reality. *SIGCOMM* (2011).
7. J Meyer, M Wasmann, W Heuten, A El Ali, and S Boll. 2017. Identification and Classification of Usage Patterns in Long-Term Activity Tracking. *CHI* (2017).
8. Lie Ming Tang and Judy Kay. 2017. Harnessing Long Term Physical Activity Data-How Long-term Trackers Use Data and How an Adherence-based Interface Supports New Insights. *IMWUT* (2017).