Mapping “Small Data” to High-Level Activities Reflecting College Student Life

Edison Thomaz
ethomaz@gatech.edu
School of Interactive Computing,
Georgia Institute of Technology
Atlanta, Georgia, USA

As digital technologies continue to infiltrate everyday life, leaving a trace of digital breadcrumbs in their wake, it becomes possible to leverage this “*small data*” [4] in service of applications that promote health and wellbeing. One scenario where this would be particularly desirable is the college campus, where many students are living independently for the first time, often under significant stress. An opportunity exists to build systems that can anticipate, and thus help avoid undesirable or potentially harmful behaviors caused by student life stress.

Although researchers have recognized the value of *small data* sources (e.g., desktop computer activity, location tracking, physical activity information) for modeling and predicting human behaviors and health status [5,6], the task of aggregating, visualizing, and annotating multiple data sources at scale remains largely elusive. A critical step in *model* building is the collection of ground truth labels for training purposes. One approach that has been proposed is to build interfaces that allow people to review their personal *small data* streams and annotate them with a high-level activity (i.e., the label). In the simple case, a *small data* stream is discrete and events can be easily mapped to high-level activities. For example, a background process observing a person using a computer can output events every time a new application is launched or a web page is visited. The individual would select the event in the user interface and indicate the activity (e.g., reading news, communicating over email, social media). Slife is an application that was built for this task (Figure 1).

The complexity of labeling low-level events rises when the high-level activity spans *multiple* streams of *small data*. A trip to the grocery store, a walk around the park, and a visit to a neighbor results in changes to multiple modalities (e.g., location, steps per minute) that might be relevant to uniquely identify the high-level activity. For low-level events that are close together in time, a “drag & drop select” might suffice for annotating in timeline-centric interfaces such as the one employed by Slife. On the other hand, there are classes of *small data* that are continuous and mostly uninterpretable on their own, making the annotation process a tedious and difficult task. For instance, smart watches now possess powerful sensors that can track arm gestures with high-resolution. Reviewing a sensor signal for relevant gestures such as drinking and

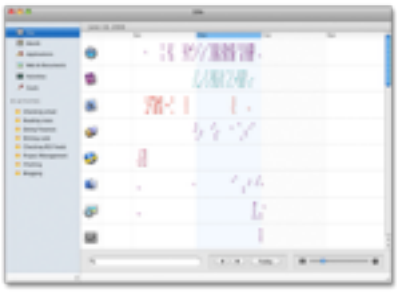

Figure 1. The Slife application interface. When running, it monitors computer activities by application and plots interactions with these applications on a timeline. Users can annotate interaction events by giving them a label.

eating is not feasible unless the sensor stream is accompanied by some photographic evidence of what occurred (Figure 2). And even when coupled with photos or video, it might be difficult to pinpoint exactly which segments of the sensor stream map to activities of interest (or gestures) for annotation purposes. One way this problem is often addressed in practice is by assigning a

weak-label to the entire section of the sensor stream in-between photographic evidence. There are machine learning modeling techniques that have been designed to work with such type of weak labels [2], but it would be preferable if more accuracy could be achieved with direct label assignments.

Figure 2. To facilitate the annotation of a sensor stream, a photo is included to provide context. But the photo might not represent all the activities captured in the sensor data, resulting in mislabeling or "underlabeling".

Annotating *small data* with the goal of building models of healthy or unhealthy student behaviors presents a number of challenges. There are two research directions areas that I consider worth exploring. The first one centers on exploring alternative approaches for annotation; I am interested in studying methods leveraging interactive machine learning in this context. Secondly, a relevant question is whether valuable applications created around personal digital traces can be built without the need for an annotation step. The answer appears to be 'yes'. Researchers such as Rantz et al. have demonstrated that with sensor networks, changes in sensor patterns can be useful in health applications even without modeling specific activities [3]. In the short term, this might represent a fruitful direction for *small data* as well.

References

[1] Slife Labs. http://www.slifelabs.com. (last accessed: 03/25/15).

[2] Hu, B., Chen, Y., Keogh, E. J. (2013). Time Series Classification under More Realistic Assumptions. SDM 2013, 578–586.

[3] Rantz, M. J., Skubic, M., Koopman, R. J., Phillips, L., Alexander, G. L., Miller, S. J., & Guevara, R. D. (2011). Using sensor networks to detect urinary tract infections in older adults. E-Health Networking Applications and Services (Healthcom), 2011 13th IEEE International Conference on, 142–149.

[4] Estrin, D., "small data, where n=me", CACM, Viewpoint Column, Communications of the ACM, Vol. 57 No. 4, Pages 32-34, April 2014.

[5] Epp, C., Lippold, M., & Mandryk, R. L. (2011). Identifying emotional states using keystroke dynamics. ACM CHI 2011, 715–724. http://doi.org/10.1145/1978942.1979046.

[6] Madan, A., Cebrian, M., Lazer, D., & Pentland, A. (2010). Social Sensing for Epidemiological Behavior Change. Ubicomp '10: Proceedings of the 12th ACM international conference on Ubiquitous computing. http://doi.org/10.1145/1864349.1864394.