Personal Analytics and Privacy
An Individual and Collective Perspective

Riccardo Guidotti[1], Anna Monreale[2]

[1] ISTI-CNR, Via G. Moruzzi, 1, Pisa, {name.surname}@isti.cnr.it
[2] University of Pisa, Largo B. Pontecorvo, 3, Pisa, {name.surname}@di.unipi.it

1 We All Need To Own and Use Our Own Data

Every year, each person leaves behind her more than 5 gigabytes of *digital breadcrumbs*, disseminated by disparate systems that we use for our daily activities: traveling, communicating, paying for goods, banking, searching the web, listening music, reading, playing, posting or tweeting, screening our health. Five gigabytes, without taking into account photos and videos, otherwise numbers would grow considerably. An avalanche of personal information that, in most cases, gets lost. Only each single individual could connect all this personal information into some personal data repository. No Google or Facebook has a similar power today, and we should very carefully avoid this possibility in the future. The fact that in the contemporary initial phase of a measurable society there are few large harvesters, or "latifundists", who store data on masses of people in large inaccessible repositories in an *organization-centric* model, does not mean that *centralization* is the only possible model, nor the most efficient and sustainable.

Nowadays, data and information belong to big organizations (Amazon, Google, Facebook, etc.) which employ *top-down* control over these data. They can create a mosaic of human behaviors used to extract valuable knowledge for marketing purposes: *our personal data is the new gold*. For example, users produce personal data like Facebook posts, or GPS movements using Google Maps, or online shopping through Amazon, and these data are collected and obscurely employed by these companies for marketing or to produce services. On the other hand, individuals do not have the tools and capabilities to extract useful knowledge from their personal data. This is a *Legrand Star* model [11], i.e., a centralized network model, where users can not directly control and exploit their own personal data. Data owning and usage would require not a *bottom-up* system, but a *Baran Web* model, i.e., a *peer distributed approach*, a network of peers, both individual and companies, in which no single node has absolute control of everything but everyone controls thyself, and has only a partial vision of the surrounding peers. The first brick that must be placed to build this *Web* and to start a change of perspective, is the development of *Personal Data Models*, which are sewn on each individual to fit their subjective behaviors.

Data Mining applied to individual data creates an invaluable opportunity for individuals to improve their *self-awareness*, and to enable *personalized services*. However, nowadays users have a limited capability to exploit their personal data, thus we need a change of perspective towards a *user-centric* model for personal data management: a vision compatible with the data protection reform of EU, and promoted by the World Economic Forum [12, 14, 15].

2 Making Sense of Own Personal Big Data

Although some user-centric models like the *Personal Intelligent Management Systems (PIMS)* and the *Personal Data Store (PDS)* are emerging [5, 1], currently there is still a significant lack in terms of algorithms and models specifically designed to capture the knowledge from individual data and to ensure privacy protection in a user-centric scenario.

Personal data analytics and individual privacy protection are the key elements to leverage nowadays services to a new type of systems. The availability of personal analytics tools able to extract hidden knowledge from individual data while protecting the privacy right can help the society to move from organization-centric systems to user-centric systems, where the user is the owner of her personal data and is able to manage, understand, exploit, control and share her own data and the knowledge deliverable from them in a completely safe way.

Recent works are trying to extend the userc-centric models for data management with tools for *Personal Data Analytics* [8]. With *Personal Data Analytics* are indicated the personal data mining processes extracting the user models, and providing self-awareness and personalized services. Personal Data Analytics can be exploited *(i)* to improve the user *self-awareness* thanks to the personal patterns they unveil, and *(ii)* to empower *personalized services* by providing proactive predictions and suggestions on the basis of the user's profile.

In practice, Personal Data Analytics allows a user to make sense of her personal data and to exploit it [9]. Fig. 1-*a)* shows the overall Personal Data Analytics approach. The individual data flow into the PDS and are stored according to one of the possible technique described in the PDS literature [5, 4, 1]. Along the analysis of the continuous digital breadcrumbs, the PDS must consider that it does not exist a unique and constant model describing human behaviors. Indeed, our behaviors will be never "in equilibrium" because we constantly move, we buy new things, we interact with our friends, we listen to music, etc., generating in this way a non-interruptible flow of personal data [2]. Therefore, Personal Data Analytics must be *dynamic* and *adaptable* to continuous changes. The user profile described by Personal Data Analytics can be used to improve the user *self-awareness* and for *personalized services* yet adopting Personal Data Analytics. Self-awareness can be provided for example through a dashboard where the user can navigate and understand her models and patterns. Examples of personalized services can be *recommendation systems* or *predictors* of future actions.

3 The Personal Data Ecosystem and the Privacy Issues

The PDS extended with PDM offers an augmented image of ourselves, our image reflected in a *digital mirror*. However, passive personal data collection and

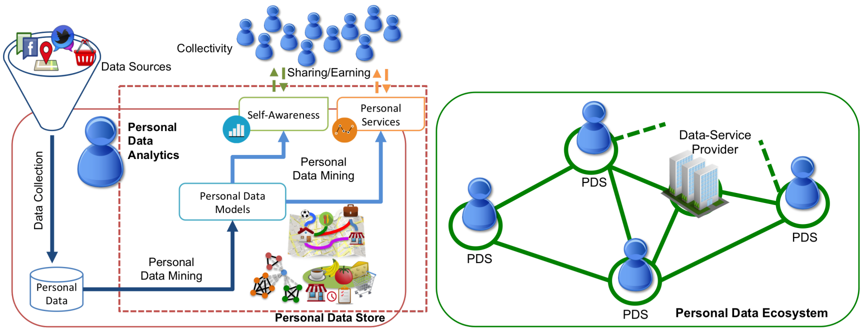

Fig. 1: *a)* Personal Data Analytics (left); *b)* Personal Data Ecosystem (right).

knowledge mining need to be balanced with *participation*, based on a much greater awareness of the value of own personal data for each one of us and the communities that we inhabit, at all scales.

Personal Data Analytics enables the comparison of our individual patterns with the collective ones of the communities we belong to, provided we have a way to interact and collaborate with other peers, individuals and institutions that are, in turn, equipped with their PDS's and connected to each other in a network. Fig. 1-*a)* (top right) shows how, to provide and obtain improved *self-awareness* and *personalized services*, a user can share information and, at the same time, earn knowledge, by communicating with the *collectivity*.

This enables a *Personal Data Ecosystem (PDE)* (Fig. 1-*b)*) that can be modeled as distributed network of peers, both individual users and public or private institutions and companies, each one with their own PDS and PDM [8]. The PDE can generate an innovative form of *collective awareness*, characterized by a self-reinforcing loop where *(i)* superior individual knowledge is created by comparison with collective knowledge, enhancing individual ability to better align own goals with common interest, and *(ii)* superior collective knowledge is created through the active participation of individuals in a decentralized system, i.e., without centralization of unnecessarily large amounts of information. The PDE is in line with the *peer progressive* idea of a decentralized network where news, money, and knowledge come from the periphery instead of from the center [11]. In [10] it is compared the concept of "wise king" (centralized system) against the Adam Smith's "invisible hand" regulating a decentralized self-organizing system. Furthermore, the PDE idea outlines the *Nervousnet* project: a globally distributed, self-organizing, techno-social system for answering analytical questions about the status of world-wide society, based on social sensing, mining and the idea of trust networks and privacy-aware social mining [7].

Although the PDE setting enables new opportunities for individuals who may exploit their own data to improve the daily life with more and more self-awareness, it is impossible to ignore the possible privacy risks that may derive from the sharing of personal data and the knowledge extractable from them. Indeed, the worrying aspect of this story is that often, individual data provide a very fine detail of the individual activities and thus, in case of *sensitive* activities the opportunities of discovering knowledge increase with the risks of *privacy violation*. The threat goes as far to recognize personal or even sensitive aspects of their lives, such as home location, habits and religious or political convictions. Managing this kind of data is a very complex task, and often we cannot solely rely on de-identification (i.e., removing the direct identifiers contained in the data) to preserve the privacy of the people involved. In fact, many examples of re-identification from supposedly anonymous data have been reported in the scientific literature and in the media, from health records to GPS trajectories and, even, from movie ratings of on-demands services. Several techniques have been proposed to develop technological frameworks for countering privacy violations, without losing the benefits of big data analytics technology [6]. Unfortunately, most of the research work done in the context of privacy-preserving data mining and data analytics focuses on an organization-centric model for the personal data management, where the individuals have a very limited possibility to control their own data and to take advantage of them according to their needs and wills. PDE instead provides to individuals the chance to have a central and active role in the control of the lifecycle of her own personal data introducing also a layer of transparency. In particular, it enables individuals to control a copy of their data and/or the knowledge extracted from them. In practices, the individuals acquire the right to dispose or distribute their own individual information with the desired privacy level in order to receive services or other benefits or in order to increase their knowledge about themselves or about the society they live in. In this setting to guarantee the information sharing with the level of desired privacy level, *Privacy-by-Design* data transformations [3, 13] must be applied before data leave the user. This guarantees the prevention of privacy attacks to the PDE addressing possible privacy issues with a pro-active approach. This encourages the voluntary participation of users, limiting the fear and skepticism that often leads people to not access the benefits of extracting knowledge from their own data, both at personal and collective level.

Acknowledgments. This work and workshop has been partially supported by the European Community's H2020 Program under the funding scheme "INFRAIA-1-2014-2015: Research Infrastructures" grant agreement 654024, "SoBigData: Social Mining & Big Data Ecosystem"(http://www.sobigdata.eu).

References

[1] S. Abiteboul, B. André, and D. Kaplan. Managing your digital life. *Communications of the ACM*, 58(5):32–35, 2015.

[2] Ball. *Why society is a complex matter*. Springer, 2012.

[3] Cavoukian. Privacy design principles for an integrated justice system. TR, 2000.

[4] V. d et al. My data store: toward user awareness and control on personal data. In *Ubicomp*, pages 179–182, 2014.

[5] Y.-A. de Montjoye, E. Shmueli, S. S. Wang, and A. S. Pentland. openpds: Protecting the privacy of metadata through safeanswers. *PloS one*, 9(7):e98790, 2014.

[6] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53, 2010.

[7] Giannotti et al. A planetary nervous system for social mining and collective awareness. *The European Physical Journal Special Topics*, 214(1):49–75, 2012.

[8] R. Guidotti. Personal data analytics: Capturing human behavior to improve self-awareness and personal services through individual and collective knowledge. 2017.

[9] R. Guidotti et al. Towards user-centric data management: individual mobility analytics for collective services. In *ACM SIGSPATIAL*, pages 80–83, 2015.

[10] Helbing. The automation of society is next: How to survive the digital revolution. *Available at SSRN 2694312*, 2015.

[11] Johnson. *Future perfect:The case for progress in a networked age*. Penguin,2012.

[12] Kalapesi. Unlocking the value of personal data:from collection to usage.WEF,2013.

[13] A. Monreale, S. Rinzivillo, F. Pratesi, F. Giannotti, and D. Pedreschi. Privacy-by-design in big data analytics and social mining. *EPJ Data Science*, 3(1):10, Sep 2014.

[14] A. Pentland et al. Personal data: The emergence of a new asset class.WEF,2011.

[15] Rose et al. Rethinking personal data: Strengthening trust. In *WEF*, 2012.