Decision Support Systems and Education
J. Mantas et al. (Eds.)

doi:10.3233/978-1-61499-921-8-55

Types of Data Clarify Senses of Data Processing Purpose in Health Care

Petteri MUSSALO[a,1], Ulla GAIN[a] and Virpi HOTTI[a]
[a] *University of Eastern Finland, School of Computing, Kuopio, Finland*

Abstract. Data taxonomy facilitates data valuation. The origin-based data taxonomy contains four types of data (provided, observed, derived and inferred) and 10 subcategories. In this paper, we report the results of multivocal literature around the origin-based data taxonomy. The review results are used to refine the definitions of the types of data when to figure out data valuation within health care. Furthermore, we exemplify how the types of data can be recognized in health care (e.g., patient medication, alerting about risk patients, patient logistics, remote monitoring) to realize data valuation based on the proposed data taxonomy around the types of data.

Keywords. Data taxonomy, functional domain, health care.

1. Introduction

Regulations and industry codes or practices have implications on types of data. For example, four types of data (provided, observed, derived and inferred) and their subcategories [1] increase awareness both data sources and processing possibilities. However, sectoral regulation data can be derived and inferred, for example, from browsing histories or credit card data [2].

There is a tension between the repurposing of data and the generation of new types of data [3]. Furthermore, organizations should be able to describe "the purposes for processing and, where necessary, its legitimate interests" [4]. Data valuation begins with the identification and classification of data [5]. In this paper, we report the results of multivocal literature around the origin-based data taxonomy [1]. The review results are used to refine the definitions of the types of data (Section 2) when to figure out data valuation with examples of health care (Section 2.4).

2. Types of data

The origin-based types of data (i.e., data taxonomy or data categories) has been written in the name of the Information Accountability Foundation [1] and used as the background paper of OECD Expert Roundtable Discussion [6]. Our multivocal review (two scientific articles [7,8], one non-scientific paper [4], one book [9], and one presentation [10]) raised examples and definitions of the origin-based data taxonomy:

[1] Petteri Mussalo, School of Computing, University of Eastern Finland, Microkatu 1 F, 70210 Kuopio, Finland; E-mail: mup@iki.fi.

- Derived and inferred data (e.g., suspectibility to a particular disease, life expectancy) affect data collection and creation by ongoing processing [7].
- Ownerships of data are not clear – over provided data ("created by direct actions taken by the user") rights belong to the users, over observed data, derived data ("information the generated mechanically by performing transformations or operations on other data"), and inferred data ("the product of a probability based analytic process") rights might belong to providers [8].
- Data capture categories based on awareness of individuals [9]: when provided (e.g., filled applications, vote and warranty registrations, drive or carry a gun licenses) then individuals are highly aware, when casually observed (e.g., recorded clicks and phones, engine temperatures and tire pressures of cars, traffic camera pictures, recorded movements) then individuals might be partly aware, when derived (e.g., time-on-page, labeled individuals) then individuals are not aware how observed and provided data are manipulated, when inferred (e.g., recommendations, less or more likely to do something) the individuals are not aware because of analytical evaluations.
- Provided data is "consciously given by individuals", observed data is "recorded automatically", derived data is "produced from other data in a relatively simple and straightforward fashion", and inferred data based on probabilities to "find correlations between datasets and using these to categorize or profile people" [4].
- The level of personal awareness is the highest against provided data and the lowest against inferred data. The need for independent oversight is the lowest against provided data and the highest against inferred data [10].

We use entities of their representatives to illustrate either individual or some other identifiable thing (e.g., robot). The entities or their representatives can be in relationships with data controllers that are accountable of recorded consequences. The consequence is an outcome of an event by the entity or its representative. Four types of data – observed (Section 2.1), provided (Section 2.2), inferred (Section 2.3), and derived (Section 2.4) – have subcategories and examples.

2.1. Observed data

When intermediate actions and actors are observed and recorded automatically, then entities or their representatives are in relationship with data controllers. Especially, the Internet of Things (IoT) devices are sources of observed data [4]. There are three subcategories of the observed data [1] – engaged, not anticipated, and passive.

Engaged data. Identifiable observations have been made aware at some points of time. Data describing patient status and based on human measurement (weight) or observations (fever and the sore throat) represent the engaged data on the health care domain.

Not anticipated data. Sensorial data elements might pertain to entities or their representatives. For examples, sensors in cars and mouse movements on the screen [1].

Passive data. Situational data might pertain to entities or their representatives. For example, call logs and facial images from closed-circuit television (CCTV) cameras [1].

2.2. Provided data

Entities or their representatives are highly aware. There are three subcategories of the provided data [1] – initiated, transactional, and posted.

Initiated data. Some actions may have recorded consequences that are publicly available a long time. For example, when entities completed in a profession requiring approval by the authority, then data are available from public records of licenses or rights. Some actions might not affect consequences that are publicly available a long time.

Transactional data. Entities or their representatives are clearly and knowingly identifying themselves. Some actions may have recorded consequences that are publicly available a long time. For example, there are many public records having details of the entities and their representatives as well their efforts. Some actions might not affect consequences that are publicly available a long time. For example, inquiries and surveys responded to [1].

Posted data. Actions are recorded consequences that might be publicly available a long time. For example, social network postings are expressions seen or heard by others [1].

2.3. Inferred data

Characterizations and likelihoods are produced through probability-based analytic process. There are two subcategories of the inferred data [1] – statistical and advanced analytical.

Statistical data. Characterizations are produced emphasizing models and their precision the definition of which is “closeness of agreement between independent test/measurement results obtained under stipulated conditions” [11].

Advanced analytical data. Likelihoods of future outcomes and their accuracy (i.e., the proximity of measurement results to the true value) are produced. In general, the word “likelihood” refers to the chance of something happening [12]. Abrams [1] gave examples as the risk of developing disease based on multi-factor analysis and college success prediction based on big data analysis at age 9.

2.4. Derived data

New data elements are derived mechanical fashion from other data. Entities or their representatives are in relationship with data controllers. There are two subcategories of the derived data – computational and notational [1].

Computational data. New data elements are created “through arithmetic process executed on existing numeric elements” [1].

Notational data. New data elements are created by classifying entities “as being part of a group based on common attributes shown by members of the group” [1]. Types of Data within Functional Domains

Functional domains represent “categorized functions generally used together” [13]. Health care functional domains can be derived, for example, from health care data sources [14] and enterprise architecture (EA) business elements and business domains [15]. Furthermore, health care has own taxonomies [16]. We exemplify data usage statements around health care functional domains as follows:

- *Patient medication and risk patient alerting*: The doctor writes an electronic prescription to the ePrescription center (transactional data) for the patient having fever and sore throat (notational data) as well a positive cultivation sample (engaged data) fitting the tonsillitis (derived data). Before writing the penicillin prescription for bacterial infection (derived data) the doctor gets alert regarding on patient allergy to penicillin (notational data).
- *Patient logistics*: There are problems (derived data) in the operating room (OR) to get the first patients of the day (notational data) in scheduled time to the OR (observed data). The problem settlement highlights that patients coming from the ward have to wait the elevators (transactional data) for a long time (derived data).
- *Remote monitoring:* Patients treated at home have specific aid to improve the quality of life and safety. Home care staff register during the visit (notational data) the observation results (engaged data). A patient weight is measured with the bed scales (observed data), the physical activity with movement sensors (passive data) and the food is provided with the nutrition automat (passive data). The central alerting unit gets the message about patient general condition worsening (inferred data) based on the decreasing weight (derived data), the decreased physical activity (derived data) and the worsening appetite (derived data).

There are transparency requirements, i.e. pipelines from raw data (i.e., observed and provided data) to insights (i.e., derived and inferred data) have to be transparent and even checked by the regulators. Therefore, trueness of the insights have to verify step by step. Furthermore, derived and inferred data can be used to support one another. For example, evaluating measurement data (i.e., observed and provided data) the statistical data (e.g., mean, range, and standard deviation) are adapted to derive calculated values of the indicators (i.e., computational data), or identifying behavioral unknown entities to which the actions should be targeted (i.e., advanced analytical data) the notational data (e.g., labeled datasets) are used and produced when behavioral known entities are classified.

During data lifecycles, the types of data follow each other in a more cyclic than random manner. Observed or provided data are derived or inferred for actionable insights – the calculations or computations of the actionable insights are evaluated, and modified if necessary, based on observed or provided data.

3. Discussion

It is not straightforward to tag data usage statements by the types of data. The substance knowledge is crucial. However, the tagging might be easier if we have categories for usage. In future, we will compare other data taxonomies such as ISO/IEC 19944:2017 [17] with the origin-based data taxonomy. For example, there are six data use categories (advertise, improve, personalize, provide, upsell, and share) [17].

We realized that the origin-based types of data (i.e., data taxonomy or data categories) [1] have been even redefined. However, there are need for explicit definitions. Therefore, we partly redefine the types of data.

The data taxonomies increase the data valuation. Furthermore, the types of data will clarify the senses of data processing purposes by data controllers or processors. For example, there are more than half a million registered data controllers by the Information Commissioner's Office [18] - most of them process even sensitive classes of information and they describe purposes for processing information in general level.

References

[1] M. Abrams, The Origins of Personal Data and its Implications for Governance, *OECD Expert Roundtable Discussion, Background Paper*, The Information Accountability Foundation, 2014, http://informationaccountability.org/publications/, or http://dx.doi.org/10.2139/ssrn.2510927, accessed 3.7.2018.

[2] *Article 16, Privacy and Freedom of Expression In the Age of Artificial Intelligence*, 2018, https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf, accessed 3.7.2018.

[3] M. Butterworth, The ICO and artificial intelligence: The role of fairness in the GDPR framework, *Computer Law & Security Review* **34** (2) (2018), 257–268.

[4] Information Commissioner's Office (ICO), *Big data, artificial intelligence, machine learning and data protection, version 2.2*, 2017, https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf, accessed 3.7.2018.

[5] S. Robinson, What's your anonymity worth? Establishing a marketplace for the valuation and control of individuals' anonymity and personal data, *Digital Policy, Regulation and Governance* **19** (5) (2018), 353–366.

[6] Organisation for Economic Co-operation and Development (OECD), *Protecting Privacy in a Data-driven Economy: Taking Stock of Current Thinking*, 2014, http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=dsti/iccp/reg(2014)3&doclanguage=en, accessed 3.7.2018.

[7] J. Sidgman, M. Crompton, Valuing Personal Data to Foster Privacy: A Thought Experiment and Opportunities for Research, *Journal of Information Systems* **30** (2) (2016), 169–181.

[8] C. Bartolini. C. Santos, C. Ullrich, Property and the cloud, *Computer Law & Security Review* **34** (2) (2018), 358–390.

[9] J. Sterne, *Artificial Intelligence for Marketing: Practical Applications*, John Wiley & Sons, 2017.

[10] K. Sproston, *The Challenges of Unlocking Health Data*, Bellberry, 2017, http://slideplayer.com/slide/11969649/, accessed 3.7.2018.

[11] ISO 21748:2017 (en), *Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty evaluation*, https://www.iso.org/obp/ui#iso:std:iso:21748:ed-2:v1:en, accessed 3.7.2018.

[12] ISO 22300:2018 (en*),* *Security and resilience — Vocabulary*, https://www.iso.org/obp/ui#iso:std:iso:22300:ed-2:v1:en, accessed 3.7.2018.

[13] ISO/IEC 26551:2016 (en), *Software and systems engineering — Tools and methods for product line requirements engineering*, https://www.iso.org/obp/ui#iso:std:iso-iec:26551:ed-2:v1:en:, accessed 3.7.2018.

[14] C. Reddy, C. Aggarwal, An Introduction to Healthcare Data Analytics, *In: Reddy C, Aggarwal C (eds) Healthcare Data Analytics*, CRC Press, USA, 2015.

[15] S. Stansfield, N. Orobaton, D. Lubinski, S. Uggowitzer, B. Eng, H. Elec, H. Mwanyika, *The Case for a National Health Information System Architecture; a Missing Link to Guiding National Development and Implementation*, 2008, https://www.researchgate.net/publication/228653432_The_Case_for_a_National_Health_Information_System_Architecture_a_Missing_Link_to_Guiding_National_Development_and_Implementation, accessed 3.7.2018.

[16] G. Acetato, V. Persico, A. Pescapé, The Role of Information and Communication Technologies in healthcare taxonomies, perspectives and challenges, *Journal of Network and Communication Applications* **107** (2018), 125–154.

[17] ISO/IEC 19944:2017 Information technology — Cloud computing — Cloud services and devices: Data flow, data categories and data use, https://www.iso.org/standard/66674.html, accessed https://www.iso.org/standard/66674.html3, accessed 3.7.2018.

[18] Information Commissioner's Office, *Register of data controllers*, 2018, https://ico.org.uk/about-the-ico/what-we-do/register-of-data-controllers/, accessed 3.7.2018.