Trusted and GDPR-Compliant Research with the Internet of Things

Jacky Bourgeois
TU Delft
Delft, The Netherlands
J.Bourgeois@tudelft.nl

Gerd Kortuem
TU Delft
Delft, The Netherlands
G.W.Kortuem@tudelft.nl

Fahim Kawsar
Nokia Bell Labs
Cambridge, UK
fahim.kawsar@nokia-bell-labs.com

ABSTRACT

The Internet of Things has become a key enabling technology for data-intensive research across universities and private organisations alike. However, the recent introduction of the General Data Protection Regulation (GDPR) in Europe has raised concerns that the GDPR might hamper data-intensive research. In this paper, we address the question of how to enable ethical and compliant research with personal IoT data in an academic environment. We identify three novel trust principles for GDPR compliant use of personal IoT data in science and research (private-by-default, analytics transparency and Accountable analytics) and propose an architecture for a trusted IoT research infrastructure.

ACM Classification Keywords

H.4.m. Information System Applications: Miscellaneous

Author Keywords

Data-Centric Research; IoT; Personal Databox; GDPR; Compliance

INTRODUCTION

The Internet of Things has become a key enabling technology for data-intensive research across universities and private organisations alike. Scientists now routinely use mobile devices and dedicated sensors to collect data and conduct research in areas such as health, behavioural sciences, education, energy, transportation, safety and security [2, 4, 9]. Companies, on the other hand, use data from connected devices and services to understand product use and optimise product design [23]. Such data provides unprecedented insights into the behaviour and social interactions of people.

However, long-standing concerns about the loss of privacy created by the ability to 'track' behaviours online and increasingly offline has prompted the European Parliament to enact the Global Data Privacy Regulation (GDPR), a legal framework for personal data that impacts organisations all around the globe. The GDPR sets strict rules for the collection and processing of personal data with a focus on control, transparency and accountability (Articles 12–23, GDPR, Chap. 3 Rights of the Data Subject [1]). While the GDPR is broadly welcome, it has raised concerns that it might hamper data-intensive research [8, 18] [14, p.956]

While the GDPR does not specifically address the IoT, IoT data raises particular control, transparency and accountability issues because of the unobtrusive and ubiquitous nature of data collection. As a consequence, some research organisations have enacted strict ethics rules for the collection, use and sharing of personal-identifiable sensor data or have even limited access to opportunistically collected infrastructure data such as data from WI-FI access points. Such data can be used, for example, to investigate people behaviour across a University campus, and theoretically enables researchers to develop individual behaviour profiles. While the GDPR contains exceptions for data collection for research purposes, in practice the GDPR will make it more difficult to collect or obtain personal IoT data for research purposes.

In this paper, we address the question of how to enable ethical and compliant research with personal IoT data in an academic environment. We formulate three trust principles which connect to data privacy by default as well as transparency and accountability of analytics processes. Then, we propose an architecture for a trusted IoT research infrastructure that ensure these principles. We walk through a research case study with Wi-Fi access points' data to demonstrate our approach and we discuss its benefits and potential extension.

RELATED WORK

Data-Centric Research

There is an increasing use of data of from IoT devices and sensors in research, outside of the confines of computer science research. In design research, for example, IoT data has been recognised as a powerful tool for providing insights and perspectives throughout a user-centred research and design process. Bogers and colleagues [3] used data from embedded sensors to understand the context of bottle feeding routines in families, leading to the concept of a connected baby bottle. Similarly, Bourgeois and colleagues [4] explored the emerging relationships between people and self-generated energy to extract insights and drive the design of future home energy systems.

IoT'18, October 15–18, 2018, Santa Barbara, CA, USA

ISBN . . . $15.00
DOI: http://dx.doi.org/

In the healthcare research context, IoT data plays an increasingly important role. This has been recognised by companies such as Apple which offers researchers the opportunity to use the Apple Research kit [2] to conduct large-scale studies that collect personal health data and provide information to data subjects. By sharing their personal data, users contribute to research they connect to and receive direct value through advanced, personalised insights on their health.

As it remains challenging, expensive and time consuming to collect large scale IoT datasets, opening and sharing them with the scientific community is highly valuable. For example, the Extra Sensory dataset [22] provides sensor data from smart phones and smart watches of 60 users, along user activity labelling and states. This provides rich insights into human behaviour and technology usage. In design and research, such data can be used in combination with qualitative data, combining *big* and *thick* data, to create rich pictures of individual and social behaviours [16].

However, there is a growing conflict between the desire to collect more and more data and the need to protect personal data. On the one hand, datasets such as Extra Sensory, have significant benefits for academic research, enabling many researchers with different backgrounds and interests to leverage a dataset in their research. On the other hand, it discloses information to an extent that data subjects probably do not fully understand - partly because data science find ever new ways to extract behavioural insights from human data sets.

Several research initiatives are beginning to address the imbalance in data ownership and control between companies and consumers. Crabtree and colleagues [6] have explored the question of an Accountable Internet of Things as a key to building consumer trust. [5], [6] and Mortier:2016 have defined the IoT databox model as a principle means of enabling accountability and providing individuals with the mechanisms needed to build trust in the IoT. While a databox can be virtual or physical, the core concept is a single place to store personal data. Some approaches include an *app store* empowering data subjects to run services inside their box, getting a data-based service without sharing the data. Data is also considered as a service, a market in which data subjects control their data by deciding which services can consume them within a set of rules. However, there is no guarantee for data subjects to remain in control of their data nor to trust services on how they use their data. In this context, we argue that personal data should remain private by default.

Informed Consent and 'Right to Explanation'

While the GDPR provides a clear motivation to responsibly handle personal IoT data, ethical excellence should remain the primary goal as laws can leave room to interpretations [24]. It should especially be the case for academic research commonly exploring beyond the status quo. In scientific research, the notion of consent and subject information have been around seen a couple of decades [7]. It has become a critical topic with the recent opportunities offered by the IoT. IoT data distinguishes itself from other personal data by its volume, velocity, variety and veracity for *each* individual data subject. These properties make the perception of information and risk related to data sharing very challenging: data subjects cannot take informed decisions for the data itself.

The Apple Research kit has built-in features for researchers build informed consent into their research studies. However, there is no opportunity for the data subject to truly understand the use of data by researchers and ultimately the data subject is left with no choice than to trust the researchers in the use of their personal data.

Privacy as a Service (PRIAAS) [19] provides a trusted entity which outsource the mediation of data transactions between *sources* and *sinks*. This *'operator never stores any generated personal data but acts only as trusted consent manager*. While such construction increases trust via an independent third party, data subjects are still giving away their data, sending a copy of their data out of their control. In addition, one-time consent forms signed at the beginning of a study is not enough [20]. Tolmie and colleagues highlight that single-time consent is not appropriate for IoT data as data subjects discover and understand the value and the risk of sharing their data as the study goes on.

TRUST PRINCIPLES

In order to understand the issues we consider the case of Delft University of Technology (TU Delft), the oldest technical university in the Netherlands with approx. 2000 researchers.

TU Delft has several research groups focused on fundamental aspects of the IoT, but an even larger number of researchers use IoT technologies to conduct science and research in computer science related fields. For example, the Safety and Security Institute beverages sensor data for research in building evacuation strategies, the Green Village – a multidisciplinary living lab – collects and uses a diverse set of sensor data for research in smart energy, smart lighting and workplace well-being, and our own research group at the Design Engineering department uses personal data for smart product and service development. Data from these initiatives is collected and analysed using a variety of back-end systems, most of them custom-built by individual research groups. In addition, [Author's University] ICT service department collects data from the campus-wide building infrastructure including Wi-Fi access point data, building security systems and environmental sensors (we refer to data from these sources as 'opportunistic sensor data'). A recent campus-wide survey identified 14 research-relevant sensors installations across the University[17]. Data from these sensors is of enormous value for researchers conducting in-the-wild and living lab type research. For example, Wi-Fi access point data can be used to understand movements patterns across campus and within office space, and can be correlated with non-sensor quantitative and qualitative data.

As a result of the recent introduction of the GDPR, the University has conducted a survey of personal data collected by the whole organisation, from student records to sensor data and has enacted strict rules on access to this data. As a common practice at the University, researchers need to follow well-prescribed ethics procedure in order to gain permission to collect and use personally identifiable information (PII). While the GDPR provides broad exceptions for data collected for research purposes, it has become clear that researchers will, in the foreseeable future, be much more cognisant about GDPR and privacy implications of the data they collect. Even before the introduction of the GDPR, privacy concerns have let the University prevent researchers from gaining access to opportunistic sensor and Wi-Fi data while at the same time sharing the same data with outside commercial organisations (using legal contracts with specific privacy stipulations).

Conducting interviews with various stakeholders (researchers, ethics, legal, ICT department), we identified several limitations of the current practice:

- Oversight: The University has limited oversight of IoT data collection and processing by researchers;
- Ethics: The research Ethics Committee has limited understanding of privacy implications of IoT data analysis performed by researchers, especially with respect to the potential for mashing up and analysing data from different sources;
- Insights: Data subjects (staff, students, visitors) have limited insight into personally identifiable information collected about them (despite GDPR regulation);
- Consent: Data subjects are unable to provide *informed* consent to data collection from either University or researchers as they lack an understanding of privacy implications;
- Access: Researchers have limited access to opportunistic sensor data;
- Sharing: Sharing of IoT data between researchers is difficult as each data set comes with unique ethics and consent conditions.

To address these issues we are working with the central ICT department and several research groups to build a *trusted* IoT research infrastructure that enables *scalable and compliant* IoT data collection and use within the University. Based on our research with stakeholders, we formulate three key trust principles for such infrastructure:

P1 Private by default: researchers cannot use personally identifiable data unless data subjects have given voluntary, explicit and informed consent;

P2 Analytics transparency: researchers must disclose their analytics algorithms, making them reviewable and traceable by relevant University stakeholders;

P3 Accountable analytics: data analytics must be performed in a trusted environment that guarantees its analytics processes and the control over the personal IoT data.

Principle 1 states that data subjects should have the opportunity to provide or refuse consent to the use of data by researchers before it can be accessed by researchers. This principle is especially relevant for opportunistic data which is initially collected by the University for purposes other than research.

Principle 2 states that analytics algorithms should be open for scrutiny in a conversational environment that encourages multiple perspectives. This is relevant since privacy aspects can only be understood if it is transparent, which data is analysed and how, especially when mashing up data from multiple sources.

Principle 3 states that analytics algorithms and data requirements should be combined and executed autonomously in a sealed environment. This is relevant to ensure that the reviewed algorithm is the one executed and to prevent data leaks.

In the following sections, we describe conceptual and technical architecture of the *trusted* IoT research infrastructure.

ARCHITECTURE CONCEPTS

We conceive the IoT research infrastructure as a multi-sided platform that enables researchers and data subjects (students, staff, visitors, etc.) to form agreements about the use of personally identifiable information (PII) (such as location or behaviour information), while giving the University effective oversight into data collection and analysis. By analysis we refer to all aspects of using personal data to generate new insights or data for research and science purpose.

The architecture is built around four stakeholder entities and three core system components.

Stakeholder Entities

Data collection and analysis involve for stakeholder entities: the data controller, data requesters, data subjects and analytic reviewers. We represent them as light boxes in the architecture diagram in Figure 1.

The **Data Controller** is the organisation collecting and hosting the data. In our case, the data controller is the university, or more specifically the ICT department, which manages data collection from opportunistic sensors and hosts the server infrastructure on which the trusted platform is implemented. The University is the legally responsible entity for ensuring the security and viability of the IoT research infrastructure and compliance with the GDPR.

The **Data Requester** is a person or a group of identified persons who want to access and analyse personal data. In our case, it refers to researchers and students conducting studies with personal IoT data.

A **Data Subject** is any identified person framed within the data collected by the controller. In our case, it refers to students, staff and visitors.

A **Reviewer** is any identified person who reviews analytics processes.

Core Components

The core system components ensure compliance with the three trust principles, Private by default, Analytics transparency and Accountable analytics.

Personal Data Box (PDB)

The Personal Data Box (PDB) is a virtual environment that aggregates and controls all personal data relating to a Data Subject. Here, we broadly reuse the Databox concept developed by Mortier and colleagues [13], without making any assumptions about concrete architecture implementations. The

PDB is a combination of a data container and a user interface that gives Data Subjects insights into which data is collected about them and allows them to control data access by Data Requesters. The purpose of the PDB is to provide Data Subjects with a single, private, intelligible and actionable place for their personal IoT data. As we will explain, the PDB also functions as a place for Data Subjects to receive and reject or approve data requests from Data Requesters.

The PDB ensures all personal data is private by default (trust principle P1) and offers a portal for each Data Subject to take informed and voluntary decision about the use of their data.

Analytics Registry

The Analytics Registry is an open repository of data analytic implementations, functioning as a cross between a version control system (e.g. Git) and a knowledge network (e.g. Stack Exchange). The purpose of the Analytics Registry is to enable the University and its community of researchers to review analytics implementations from multiple perspectives, and to make them available for reuse.

Researchers publish analytic implementations on the registry, together with a description of their purpose, data inputs and outputs. Like a version control systems, published material can be updated and new version can be created, while the Analytics Registry keeps track of past versions. This allows the University (from ethic committee to researchers) to review analytics implementations, raise questions, and track changes. In this sense, the Analytics Registry ensures compliance with trust principle P2.

Analytics Box

The Analytics Box is a runtime environment that enables trusted data analytics by combining data from the PDB with analytics implementations from the Analytics Registry. The Analytics Box initialises with a given data analytic implementation, sends a request for data based on a list of data requirements, executes the data analytic with the received data and publish the output. An Analytics Box is instantiated whenever a researcher wants to analyse data and automatically destroyed afterwards.

The Analytics Box can only process data contained in the PDB using analytics implementations from the Analytics Registry. This ensures that all analytics processes are traceable and verifiable. In that sense, the Analytics Box ensures compliance with trust principle P3.

ARCHITECTURE

The PDB, the Analytics Registry and the Analytics Box are components of the University's IT infrastructure. Figure 1 illustrates the architecture which connects these components to enable the data analytic flow. Two components complete the picture to bind them together: a general data store and an analytics hub.

General Data Store

Access to historical data is required for nearly all research. Providing a central repository is more efficient and carries less risk of unauthorised access than each research group harvesting and storing their own sets of data. The general data store is a database containing all personal data collected by the university. Following the principle P1, this data cannot be accessed by anyone but its data subjects through their PDB. Data which is not identifiable can be kept for a limited period of time until identification (e.g. a visitor filling a visitor form to access the Internet). Unidentified data cannot be accessed. The general data store handle secured transport protocols, storage encryption and access policies as a typical cloud storage provider. The general data store provides four APIs.

- **Receive data** from sensors. When received, the data is associated with a data subject when available (i.e. de-anonymised) and stored (P1).
- **Receive call** for data, with the ability to search through the entire data store for valid datasets available. This API enables data requesters to search for data and qualities that fit their needs such as data types, frequency, period, potential amount of data subjects. The requester can either search for data, receiving only statistics, or actually request the data. In the latter, each data subject fitting the call receives a data request to consent (P1).
- **Receive consent** from data subjects via their PDB. The general data store handles the consent of each data subject, tightly managing access rule policies (P1).
- **Serve data** on demand for an Analytic Box. The general data store is in charge of serving the data to a requester. Only consented data is delivered. This can be offline, historical data or a subscription to online, real-time data stream (P3).

When receiving a request for data, data subjects can consult the Analytics Registry to see the summary of the reviews and ratings from the ethics committee, domain experts and general audience. They can vote for useful information and dive into the details of the analytic itself if necessary. If the result of a data analytic generates identifiable personal information, it should be sent to the general data store to be accessible to the data subjects through their PDB. In this case, the consent contains a dedicated section for disclosing results to the researchers.

Analytics Hub

The Analytic Hub is the orchestrator of all personal data analytics. It provides an API for researchers to request a job. This analytic job contains:

- A reference to an analytics implementation on the Analytics Registry. The hub fetches the data requirements of the implementation from the Analytics Registry, including the type and frequency of data;
- A specification of contextual data requirements for a given job such as the time frame, location area, data subject gender or age range;
- The minimal data requirements to execute the requested job such as the amount of data subjects;

The hub composes a call for data, combining all requirements, and sends it to the general data store along the researcher's profile for the data subjects information. It keeps the researchers

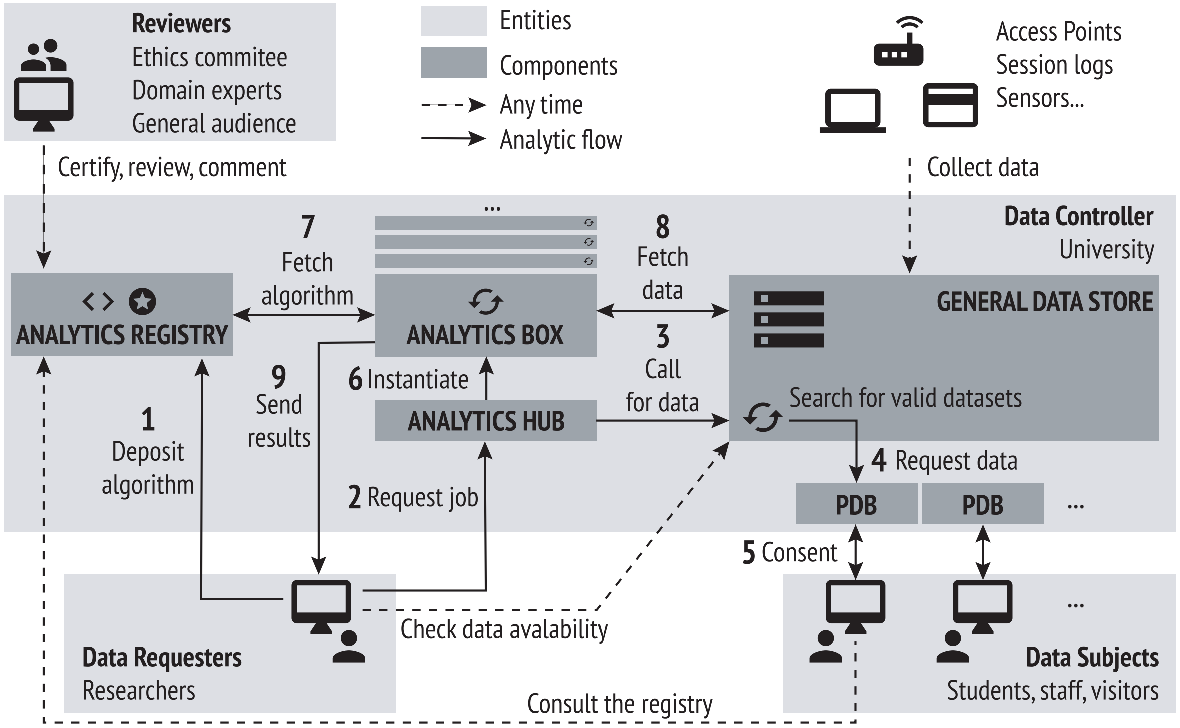

Figure 1. Key components of a GDPR-compliant analytics architecture for personal IoT data. The digits from 1 to 9 identify the ordered steps of the personal data analytic flow.

up-to-date of the state of their requests. Once enough data subjects consented to provide data to meet the requirements, the Analytics Hub creates and handle the life cycle of an Analytics Box.

RESEARCH CASE STUDY

In this section we illustrate our approach by walking through a research case study from the TU Delft Safety and Security Institute. Researchers at this institute investigate and develop effective building evacuation strategies and conduct real-world trials on campus which involve the tracking of people during emergency. In order for researchers to be able to conduct follow-up interviews and interpret evacuation behaviours with respect to demographic characteristics such as gender and age, it is important to identify individuals and their location profiles.

Wi-Fi-based indoor localisation is a well-established method for tracking mobile devices and people in buildings [25]. These methods use access point data as listed in Table 1 to compute location of devices such as mobile phones. If device ownership is known, researchers can identify and track people throughout a building. Collecting Wi-Fi data is technically easy, but making it available to researchers involves a cumbersome administrative process which requires coordination between research ethics committee and the ICT department. In addition, the ICT department is required to extract Wi-Fi data from their logs in a way that ensures that researchers only gain access to data for which they are authorised. This might require filtering out data relating to people who do not participate in studies. Since the GDPR came into effect, the University is increasingly reluctant to share personable identifiable Wi-Fi data, and thus use of this data for research purpose is getting harder.

The proposed architecture enables a trusted, scalable workflow that enables 1) researchers to request access to personally identifiable data from data subjects, 2) data subjects to provide consent 3) researchers to analyse data and 4) the University to oversee the overall process. The workflow has nine distinct steps:

1. Deposit Algorithm – The first step involves the researcher who investigate the potential of using Wi-Fi for indoor emergency location. They sign in to the IoT research platform, query the general data store to get a feeling of which data is available and use the 'data availability' API to formulate more precise requests such as 'How many data subject have Wi-Fi access data in a given building over a specific period'. Along the metadata descriptions, the result contains distributions of data frequency and amount of data subjects over the period, giving the researchers some clues about what they can expect from the data. Then, the researchers develop algorithms that include a specification of the data requirements, and deposit them in the Analytics Registry. At this stage, the data requirements could include the MAC address and signal strength from the Wi-Fi access point at maximum granularity, as well as demographic information associated with each device.

Field	*Description*	*Example*
Client Username	Pseudo Anonymised	zZxY1LCKH5tel5+uxO/c5xmuW/aaQ+v6Pb0kJCIYwkE=
Client MAC Address	Pseudo-Anonymised	C9zrPEQsychXcEUqjjMxd4tyhNCbyQTORWDHT+U4Bps=
Association Time	Date/Time	Fri Sep 22 11:07:09 CEST 2017
AP Name	Access Point name	A-03-0-030
Map Location	Location hierarchy	System Campus > 03-Science Center > Beganegrond
Session Duration	Elapsed time	5 min 4 sec
SNR	Signal to Noise Ratio	30
RSSI	Received Signal Strength Indication	-61

Table 1. Fields of the Wi-Fi Access Point dataset.

2. Request Job – Once the researchers deposited the analytic implementation on the registry, they can send multiple analytics jobs to the hub. Each job refers to the implementation deposited on the registry along the data requirements which vary from a job to another. In our case, the researchers could include the access points of a specific building, a given time frame, requiring data from people with reduced mobility (e.g. wheelchair user).

3. Call for data – The analytic hub sends all data requirements to the general data store which search through the data to identify potential data subjects. This process is reported to the analytic hub to keep the researchers up-to-date about the amount of potential data subject found as well as pending and consented requests.

4. Request data – The general data store sends a data request to the data subjects who fit the requirements. This request is personalised based on preferences set in the PDB. For instance, a data subject might decide to reject all requests, to receive them by text message or to automatically consent the ones coming from a specific researcher.

5. Consent – When the data subject consent to provide data for a given analytic job, the general data store create an access rule that links the analytic job ID to the consent. An update is sent to the analytic hub.

6. Instantiate – The Analytics Hub instantiates an Analytics Box, either automatically when all data requirements are met or when the researchers execute the request on the hub's UI.

7. Fetch the algorithm – The Analytics Box fetches the analytics implementation from the registry and extracts the data requirements.

8. Fetch data – The Analytics Box sends the request job ID to the general data store, which sends all the consented data in return. Once ready, the run the implementation with the received data.

9. Send results – In our scenario, the Analytics Box aggregates and sends the results to the researcher. However, the more precise the data requirements, the greater chances of personal analytic results. Without minimum data subject requirements, a job running on the dataset of a single data subject is sent automatically to the general data store. The researchers will need to request this personal data explicitly if needed.

The University (i.e. ethics committee, community of researchers) can oversee the whole workflow by reviewing data requests and the analytics descriptions. This review can take place anytime during the workflow, before, during and after data use. Similarly, data subjects can consult the Analytic Registry throughout the process to inform their decision and update their consent.

DISCUSSION

The GDPR is an important piece of legislation which is having an influence on data practises of companies and universities alike. While it is too early to tell if the GDPR will have a significant positive impact in the long run, it is already impacting the use of personal data in science and research. Universities and other research organisations have a duty to oversee research conduct and ensure that they comply with the GDPR while it is in their interest not to overly restrict how researchers can use data.

The Internet of Things makes the issues more pressing. On the one hand, data from connected sensors and devices has become a hugely important source of insights for researchers and scientists; on the other hand, ubiquitous sensing makes it much more difficult to guard against accidental and unforeseen privacy infractions. Despite the fact that the GDPR contains broad exceptions for research, the uncertainties around the interpretation of the GDPR and the privacy ramifications of sensing technologies are making some universities overly careful in which and how researchers can use IoT data.

In the preceding sections we identified issues in the current use of IoT data (namely oversight, ethics, insights, consent, access and sharing) and presented the proposal of a platform architecture that addresses these issues. The novelty of our approach lies in several aspects:

First, we formulated three novel trust principles (private-by-default, analytics transparency and accountability) for handling of personal IoT data in a research context. These principles bring together three key elements for establishing trust: 1) data ownership, consent and control is in the hands of data subjects 2) researchers are open and transparent about the algorithms they use for analysing data and 3) algorithms are applied to data in a controlled fashion. We view this principle as generic in a sense that they serve as useful underpinnings of any trusted IoT research infrastructure.

Second, we defined three core architecture components to realise these trust principles (Personal Data Box, Analytics Registry and Analytics Box). While the Personal Data Box is inspired by previous work on personal data stores [5], the Analytics Registry and Analytics Box are new concepts. More importantly, the key novelty lies in the combination and interplay between these three components. While each component serves a useful purpose, it is only the combination of all three that establish trust.

The third novel aspect of our approach lies in the fact that we conceive of the IoT research infrastructure as a multi-sided platform. Just like buyers and sellers on eBay have different yet mutually reinforcing interests, researchers, data subjects and the University have diverging yet complementary interests. Researchers are interested in gaining access to personal data. Data subjects (i.e. students, researcher and other University staff) have an interest that their data is protected. At the same time, we can assume that in a University context data subjects also have a broad interest in fostering science and research and are willing to share their data for research purposes – as long as data is used transparently and responsibly. Finally, the University as a research entity has the interest of enabling data-intensive research and, as a legal entity subject to the GDPR, a duty to oversee the use of data in research. Our platform approach takes all these perspectives into account and enables researchers and data subjects to form agreements about the use of personally identifiable information while giving the University effective oversight into data collection and analysis.

A key advantage of our approach is that it achieves *scalability* in terms of the number of researchers and data subjects involved and the number of data sets and analytics algorithms. What before used to be a largely manual and often spotty process of ethics reviews and compliance monitoring can now become a semi-automated process with a well-defined workflow. The Analytics Registry provides reviewable records of algorithms while the Analytics Box provides reviewable traces of analytics processes involving personal data. These records and traces are automatically generated and could - theoretically - automatically be verified. However, how this could be done is at this point an open question that warrants further research.

The work we present in this paper has several limitations. First and foremost, the proposed architecture is a high-level architecture with many details missing. For example, we have not yet specified what form the entries of the Analytics Registry should take. We assume a combination of algorithm specification, algorithms implementation and algorithm metadata. There is relevant work on algorithms indexing, searching, discovery, and analysis [21]and semantic annotation of data processing pipelines [12] but it is unclear if these approaches are suitable in our context.

Another important aspect which we have not yet explored in depths is the matter of informed consent in an IoT research context. It has long been recognised in the medical literature that informed consent as practised for example in medical trials is not working well [10]. Similar concerns have been raised in social network analysis [11] and with respect to mobile imaging, pervasive sensing, and location tracking[15]. However, it is unclear how informed consent within a research organisation, where data subjects have a heightened understanding of research and potentially an interest in supporting research, differs from informed consent in a context where there is no established sense of common purpose between researcher and data subject.

CONCLUSION

The Internet of Things has become a key enabling technology for data-intensive research across universities and private organisations. However, the recent introduction of the GDPR has raised concerns that the GDPR might hamper data-intensive research [8, 18] [14, p. 956] and there is clear evidence that universities are becoming more aware of the privacy risks involved in the use of IoT data for research purposes, with some universities limiting access to such data. Researchers have a vital interest in taking these concerns seriously and working towards solutions that address the valid perspectives of data subjects and University administration. In this paper, we identified trust principles for an IoT research infrastructure and proposed an architecture to realise trusted IoT data use at scale in a research context.

REFERENCES

1. 2018. General Data Protection Regulation (GDPR). (2018). `https://gdpr-info.eu`

2. Apple. 2018. ResearchKit and CareKit. (2018). `https://www.apple.com/lae/researchkit`

3. Sander Bogers, Joep Frens, Janne van Kollenburg, Eva Deckers, and Caroline Hummels. 2016. Connected Baby Bottle: A Design Case Study Towards a Framework for Data-Enabled Design. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems (DIS '16)*. ACM, New York, NY, USA, 301–311. DOI: `http://dx.doi.org/10.1145/2901790.2901855`

4. Jacky Bourgeois, Janet van der Linden, Gerd Kortuem, Blaine A. Price, and Christopher Rimmer. 2014. Conversations with My Washing Machine: An In-the-wild Study of Demand Shifting with Self-generated Energy. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. ACM, New York, NY, USA, 459–470. DOI: `http://dx.doi.org/10.1145/2632048.2632106`

5. Perera Charith, Wakenshaw Susan Y. L., Baarslag Tim, Haddadi Hamed, Bandara Arosha K., Mortier Richard, Crabtree Andy, Ng Irene C. L., McAuley Derek, and Crowcroft Jon. Valorising the IoT Databox: creating value for everyone. *Transactions on Emerging Telecommunications Technologies* 28, 1 (????), e3125. DOI:`http://dx.doi.org/10.1002/ett.3125` e3125 ett.3125.

6. A. Crabtree, T. Lodge, J. Colley, C. Greenghalgh, and R. Mortier. 2017. Accountable Internet of Things? Outline of the IoT databox model. In *2017 IEEE 18th*

International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM). 1–6. DOI: http://dx.doi.org/10.1109/WoWMoM.2017.7974335

7. David Friedman. 2000. Privacy and Technology. *Social Philosophy and Policy* 17, 2 (2000), 186.

8. Valerio Grossi, Beatrice Rapisarda, Fosca Giannotti, and Dino Pedreschi. 2018. Data science at SoBigData: the European research infrastructure for social mining and big data analytics. *International Journal of Data Science and Analytics* (15 May 2018). DOI: http://dx.doi.org/10.1007/s41060-018-0126-x

9. M. Handte, S. Foell, S. Wagner, G. Kortuem, and P. J. MarrÃşn. 2016. An Internet-of-Things Enabled Connected Navigation System for Urban Bus Riders. *IEEE Internet of Things Journal* 3, 5 (Oct 2016), 735–744. DOI: http://dx.doi.org/10.1109/JIOT.2016.2554146

10. Gail E Henderson. 2011. Is informed consent broken? *The American journal of the medical sciences* 342, 4 (2011), 267–272.

11. Luke Hutton and Tristan Henderson. 2015. " I didn't sign up for this!": Informed consent in social network research. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM)*.

12. Sepideh Mesbah, Kyriakos Fragkeskos, Christoph Lofi, Alessandro Bozzon, and Geert-Jan Houben. 2017. Semantic Annotation of Data Processing Pipelines in Scientific Publications. In *European Semantic Web Conference*. Springer, 321–336.

13. Richard Mortier, Jianxin Zhao, Jon Crowcroft, Liang Wang, Qi Li, Hamed Haddadi, Yousef Amar, Andy Crabtree, James Colley, Tom Lodge, Tosh Brown, Derek McAuley, and Chris Greenhalgh. 2016. Personal Data Management with the Databox: What's Inside the Box?. In *Proceedings of the 2016 ACM Workshop on Cloud-Assisted Networking (CAN '16)*. ACM, New York, NY, USA, 49–54. DOI: http://dx.doi.org/10.1145/3010079.3010082

14. Menno Mostert, Annelien L Bredenoord, Monique C I H Biesaart, and Johannes J M van Delden. 2016. Big Data in medical research and EU data protection law: challenges to the consent or anonymise approach. *European journal of human genetics : EJHG* 24, 7 (July 2016), 956âĂŤ960. DOI: http://dx.doi.org/10.1038/ejhg.2015.239

15. Camille Nebeker, Tiffany Lagare, Michelle Takemoto, Brittany Lewars, Katie Crist, Cinnamon S Bloss, and Jacqueline Kerr. 2016. Engaging research participants to inform the ethical conduct of mobile imaging, pervasive sensing, and location tracking research. *Translational behavioral medicine* 6, 4 (2016), 577–586.

16. Pamela Pavliscak. 2015. *Data-Informed Product Design*. O'Reilly.

17. Brian Reinders. 2017. *Inventarisatie sensoren*. Technical Report. Delft Safety and Security Institute.

18. Michael John Mark Rumbold and Barbara Pierscionek. 2017. The Effect of the General Data Protection Regulation on Medical Research. *J Med Internet Res* 19, 2 (24 Feb 2017), e47. DOI: http://dx.doi.org/10.2196/jmir.7108

19. X. Su, J. Hyysalo, M. Rautiainen, J. Riekki, J. Sauvola, A. I. Maarala, H. Hirvonsalo, P. Li, and H. Honko. 2016. Privacy as a Service: Protecting the Individual in Healthcare Data Processing. *Computer* 49, 11 (Nov 2016), 49–59. DOI: http://dx.doi.org/10.1109/MC.2016.337

20. Peter Tolmie, Andy Crabtree, Tom Rodden, James Colley, and Ewa Luger. 2016. "This Has to Be the Cats": Personal Data Legibility in Networked Sensing Systems. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 491–502. DOI: http://dx.doi.org/10.1145/2818048.2819992

21. Suppawong Tuarob, Sumit Bhatia, Prasenjit Mitra, and C Lee Giles. 2016. AlgorithmSeer: A system for extracting and searching for algorithms in scholarly big data. *IEEE Transactions on Big Data* 2, 1 (2016), 3–17.

22. Y. Vaizman, K. Ellis, and G. Lanckriet. 2017. Recognizing Detailed Human Context in the Wild from Smartphones and Smartwatches. *IEEE Pervasive Computing* 16, 4 (October 2017), 62–74. DOI: http://dx.doi.org/10.1109/MPRV.2017.3971131

23. Janne van Kollenburg, Sander Bogers, Eva Deckers, Joep Frens, and Caroline Hummels. 2017. How Design-inclusive UXR Influenced the Integration of Project Activities: Three Design Cases from Industry. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1408–1418. DOI: http://dx.doi.org/10.1145/3025453.3025541

24. Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law* 7, 2 (2017), 76–99. DOI: http://dx.doi.org/10.1093/idpl/ipx005

25. Jiang Xiao, Zimu Zhou, Youwen Yi, and Lionel M. Ni. 2016. A Survey on Wireless Indoor Localization from the Device Perspective. *ACM Comput. Surv.* 49, 2, Article 25 (June 2016), 31 pages. DOI: http://dx.doi.org/10.1145/2933232