Master Thesis: Open Specification of a user-controlled Web Service for Personal Data

G. Jahn

December 14, 2016

Abstract

Data is the currency of tomorrow. Organizations, whether in the private or public sector, are gathering enormous amounts of personal (big) data. This data is harvested and incorporated by these third parties, but were created by individuals and should, therefore, belong to them. People are depending on their data. Their identity as well as their personality are defined by their personal data. Meanwhile data silo operators are hammering onto these haystacks eagerly trying to find any correlations worth interpreting, thereby almost inevitably discriminating against the rightful owners. To reduce the possibility of discrimination only bare minimum of data required should be handed over to a third party. Thus the individual has to be in charge of the whole process. A personal data service will empower its user to regain full control over her data and facilitates detailed information on every data flow. To be able to trust such a tool, the user should be able to look inside. Therefore a personal data service has to be open source and developed transparently, which would then also encourage self-hosting.

Contents

1

Introduction

1.1 Motivation

Nowadays it is rare to find someone that does not collect data about some kind of thing; particularly humans are the targets of choice for the *Big Data Movement* [1]. Since humans are all individuals, they are - more or less - distinct from each other. However, subsets of individuals might share a minor set of attributes, but the bulk is still very unique to an individual, given that the overall variety of attributes is fairly complex. That small amount of shared attributes might seem to be less important, due to the

nature of inflationary occurrence, but the opposite turns out to be true. These similarities allow to determine the individuals who are part of a subset and the ones who arn't. Stereotypical patterns are applied to these subsets and thus to all relating individuals. Thus enriched information are then used to help predicting outcomes of problems or questions regarding these individuals. In other words, searching for causation where in best the case one might find correlations - or so called *discrimination*, which

> [...] refers to unfair or unequal treatment of people based on membership to a category or a minority, without regard to individual merit. *[2]*

When interacting directly with each other, discrimination of human beings is still a serious issue in our society, but also when humans leverage computers and algorithms to uncover formerly unnoticed information in order to include them in their decision making. For example when qualifying for a loan, hiring employees, investigating crimes or renting flats. Approval or denial, the decision is based on computed data about the individuals in question [3], which is simply discrimination on a much larger scale and with less effort - almost parenthetically. The described phenomenon is originally referred to as *Bias in computer systems* [4]. What at first seems like machines going rouge on humans, is, in fact, the *cognitive bias* [5] of human nature, modeled in machine executable language and made to reveal the patterns their creators were looking for - the *"Inheritance of humanness"* [6] so to say.

In addition to the identity-defining data mentioned above, humans have the habit to create more and more data on a daily basis - pro-actively (e.g by writing a tweet) and passively (e.g by allowing the twitter app accessing

their current location while submitting the tweet). As a result, already tremendous amounts of data keep growing bigger and bigger, waiting to be harvested, collected, aggregated, analyzed and finally interpreted. The crux here is, the more data being made available [7] to *mine*, the higher the chances to isolate data sets, that differ from each other but are coherent in themselves. Then it is just a matter of how to distinguish the data set and thereby the related individuals from each other.

In order to lower potential discrimination we either need to erase responsible parts from the machines, thereby it's crucial raising awareness and teaching people about the issue of discrimination, or we try to prevent our data from falling into these data silos. The latter will be addressed in this work.

1.2 Purpose & Outcome

From an individual's perspective providing data to third parties might not seem harmful at all. Instead eventually one get improved services in return, e.g. more adequate recommendations and fitting advertisement, or more helpful therapies and more secure environments. That said, though it is a matter of perception what's good and bad, what's harmful and what's an advantage. Computing data to leverage decision making is essentially just science and technology and it's up to the humans how such tools are getting utilized and what purposes they are serving. Hence it should be decided by the data creators, how their data get processed and what parts of them are used.

To tackle the described issue the initial idea here is (1) to equip individuals with the ability to control and maintain their entire data distribution and (2)

thus reducing the amount of *potentially discriminatory* [2] attributes leaking into arbitrary calculations. To do so people need a reliable and trustworthy tool, which assists them in managing all their *personal data* and making them accessible for 3rd parties but under their own conditions. After getting permission granted these data consumers might have the most accurate and reliable one-stop resource to an individuals's data at hand, while urged to respect their privacy at the same time. However this also comes with downsides in terms of security and potential data loss. Elaborating on that and discussing different solutions will be part of the [design process][Design].

The way how to solve the described dilemma is not new. Early days of work done in this field can be dated back to the Mid-2000s where studies were made e.g. about recent developments in the industry or user's concerning about privacy, and the term *Vendor Relationship Management (VRM)* were used initially within the context of user-centric personal data management, which also led into the *ProjectVRM* [8] started by the *Berkman Klein Center for Internet & Society at Hardvard University*. Since then a great amount of effort went into this research area until today, while also commercial products and business models trying to solve certain problems. For instance concepts such as the *Personal Data Store (PDS)* [9] or a *MyData* [10] implementation called *Meeco* [11], which will all be covered in a more detailed way within the following chapter.

The work and research done for this thesis will be the foundation for an *Open Specification*, which by itself is a manual to implement a concept called *Personal Data as a Service*. Important topics like how the architecture will look like, where the actual data can be stored, how to obtain data from the ex-

ternal API or what requirements a user interface for data management need to satisfy, will be examined. After the thesis will be finished, the majority of core issues should already be addressed and can then get outlined in the specification document. Only then the task to actual implement certain components can begin. The reason for that is, when sensitive subjects especially like people's privacy is at risk, all aspects in question deserve a careful considerations and then get addressed properly. Thus it is indispensable to put adequate effort primarily into the theoretical work. To be clear though, that doesn't mean writing code to test out theories and ideas can't be done during research and specification development. It might even help to spot some flaws and eventually trigger evolvement.

To ensure a great level of trust to this project and the resulting software, it is vital to make the development process fully transparent and encourage people to get involved. Therefore it is required to open source all related software and documents [12] from day one on.

In summary, this document is meant to be the initial step in a development process fabricating a tool to manage all data that defines her identity controlled and administrated by it's owner, and maybe give her a more precise understanding about where her personal information flow and how this might effect her privacy.

1.3 Scenarios

The following use cases shall depict different situations and possible ways such emerging software might be applicable or useful, while providing it's

user with more control over her personal data. Some of them are more practical and realistic, like ordering and purchasing online a product, others might have no current usage, but showing a certain potential to become more relevant when new technologies and business models emerge, followed by new demands of data.

- order sth online, purchase, and package shipment
- social network accessing arbitrary profile data
- credibility (applying for a loan) validation by a certain financial institution: accessing arbitrary data
- patient/health record
- care (movement) data

1.3.0.1 Ordering a product online

The data owner searches through the web to find a new toaster, since her old one recently broke. After some clicks and reviews, she found her soon-to-become latest member of the household's kitchenware. After putting the model name in a price search engine, hoping to save some money, the first entry, offering a 23% discount, caught her attention. She decides to have a deeper look into the toasters and thus has heading towards the original web shop entry. Finally she came around and put the item onto her card, despite tha fact, that she has never bought something from that online shop before. Then she proceeded to checkout to place her order. The shop-interface is asking her to either insert her credentials, proceed without registration or sign-in, or insert s URI to an endpoint of her *Personal Data as a Service*. TODO: the following description might need some adjustments according

data flow / process description She opens up the management panel of her *PDaaS* and creates a new entry in a list of data consumers, that already have access to characteristics of her personal data. As a result, she receives a URI, which she inserts according, as mentioned before; after she assures herself that the data exchange with the shop through the browser is based on a secure connection (HTTPS). Under this URI, the shop-system can then request data, that is required for a successful transaction. Moving on to the next step after submitting the URI, the data owner is ask to decide how she would like to pay. The choices are: credit card, invoice, paypal or bank transfer. She chooses the last one, submits her selection and thereby completes her order. After a moment, a push notification pops up on her mobile device, which is a permission request from her *PDaaS*, asking for granting the shop-system, she just places the order, access to her full name, address and email. Additionally she can decide between three states of how log the permission wil be valid: *one-time-only*, *expires-on-date* and *until-further-notice*. Since she never ordered at this shop before and might never again, she decided to grant access only for this specific occasion. After the shop-system receives the data, it sends an email to the data owner, containing some information about her order, including the shop's bank details. which then enables her to actually pay the amount due. After the system recognizes the payment has coming in, it triggers the shipment of the toaster. In order to get a full impression of how the whole process might have look like when the data owner had chosen one of the other payment methods, the differences will be describes in the following. If the data owner would have wanted to pay with her credit card, the only difference would have been, that the shop-system had requested also to access the credit card number and it's

belonging secret, and when sending the email the system would have omitted the information about the shop's bank details. Being able to choose paying with invoice where possible only because the *PDaaS* response has indicated, that it's containing *profile data* is certified and therefore trustworthy. Which reduces the shop owner's risk and would have enabled him in case of fraud or misuse to take action. Choosing to involve paypal as a *middleman* to process the payment, requires the data owner to had already granted paypal certain access to her *PDaaS*. If that's the case, then the shop-system would have ask also for her paypal-ID, which then the system will use to request the payment directly from paypal. This on the other side will cause paypal to consult the *PDaaS*, which results in a second notification, asking the data owner for permission to proceed. After the payment transfer was successful, the shipment will gets initiated. And with the package arriving at the data owner's doorstep the whole transaction has finished.

1.3.0.2 Interacting with a social network

Entering a social network for the first time, only take the URI to the data owner's *PDaaS* and a password. The data owner receives a notification on her mobile device asking for permission to access certain data about her. If her mobile device is currently not at hand, she can also use the administration panel provided by her *PDaaS* and reachable with a web browser on every internet-enabled device. Within that panel pending permission reviews will be indicated. Whether the data owner has already reviews the request or not, she should be able to login to the social network. After doing so, she should not be able to see any of her information. After granting permissions to

the social network to accessing certain data *until-further-notice* and reloading the session, she then should see all her So every time, someone on that network tries to access her information, whom she has allowed to see that information (which is managed by the user only from within the network), the network pulls the data from the owner's *PDaaS*, if it's still permitted to do so. It's also imaginable, that the social network and a *PDaaS* are establishing a backward channel. This channel could be used to send all the content she would create over time while interacting with the social network and it's participants back to her *PDaaS*. The network itself only stores a reference to all content object, whether it's for example an image, a post or comment on somebody else's post and if it's needed the actual content will be fetched from the owner's *PDaaS*.

1.3.0.3 Applying for a loan and checking creditworthiness

The data owner would like to buy an apartment. In order to finance such a acquisition, she needs a funding, which in her case, will be based on a loan. During a conversation in a credit institute of her choice, an account consultant describes to her what data will be required in order to decide about her creditworthiness. While giving a consensual nod, she takes out her smartphone and brings up the management panel of her *PDaaS*. With a few taps she has just created a new *data consumer*. The panel then shows a QR-Code, that holds a URI to a dedicated endpoint of the data owners *PDaaS*. She shows that code to her consultant, who then scans it. While handling some more formalities and talking about several issues and possible

products she might be interested it, she gets a notification on her phone, informing her about a permission request the institute just made. It lists all the different data points the institute would like to access in order to calculate her scoring, such as address, monthly income, relationship status and family, history of banking or other current loans. After some back and forth and solving some misunderstandings with the help of her consultant, she decided to just partially allow access to the requested data and just for this time and purpose. The consultant kindly pointed out, that these decisions might have an impact on the scoring and thereby on the lending and it's terms. After the consultant got a signal from the computer system, the two then finishing up their meeting and the consultants informed the data data owner about the next steps, which includes a note, that the institute will contact her within the next few days, when they have come to a conclusion. In case of a positive outcome a new appointment need to be made, for doing all the paperwork and signing the contract. From a technical point of view, two different ways of computing the score are imaginable. The first one would be, transferring only the plain data - request, containing the query and response containing the data - including the expire date and information regarding the signature state. But the actual computations and analytics to obtain the score, will happen within the infrastructure of the credit institute. When this process is over, all transferred personal data has to be deleted. An alternative could prevent the data from leaving the *PDaaS*, in which the institute's request won't consists of a data query. Instead it would came along with an chunk of software and some information on how to run it. The *PDaaS* server will provide an isolated runtime in which the software then gets executed. After the process has finished, the result will be send back to the credit institute's

infrastructure.

1.3.0.4 Maintain and provide it's own health/patient record

Some time ago on a hiking trip in a moment of carelessness the data owner has accidently broke her leg. She came into a hospital and went straight into surgery, where the physicians could fix the injury. Time went by and the leg has healed completely. After she woke up today she felt some pain coming from that area where her leg was broken. She decided to call in sick and went straight to a doctor nearby. During her recovery she visited that doctor regularly. At the reception desk, she opens up the *PDaaS*'s management panel on her smartphone and searched through the list of data consumers. After she found the entry for this clinic, she flipped her phone to show the receptionist the corresponding QR-Code, which she started to scans immediately. However the receptionist couldn'd see any data on the screen, because the access has already expired. The data owner only had permitted access for the estimated time of recovery, which was over some time ago. That's why she got a notification, to re-grant some access. Going through the data points the clinic-system has requested, she noticed that her address is incorrect. Last month she moved out and into a bigger apartment just down the street. She must have forgot to change that data, which she corrects immediately right before submitting the access configurations for the clinic-system. She also included the access to all the data originated from that time after her accident. A moment later the receptionist confirms to now being able to see all necessary data. The data owner takes a seat

in the waiting room. While passing some time, she had a deeper look into her list of data consumers; some of them she couldn't even remember and for others she was surprised to what data she has granted access to and started to reduce certain permissions, if it was appropriate in her eyes. She even removed some of the entries. The appointment with her doctor went great. He even had to review the x-ray images in order to make a adequate differential diagnosis. After the visit, she had to make a quick stop at a pharmacy along the way to pickup the drugs her doctor had prescribed for her to reduce the pain. She had to wait in the queue with two other customers being in front of her. She realized, that it's the first time she has been here. So she prepared a new entry in her data consumer list, including all information about her prescriptions. So by the time she get served, she just let the person behind the register scan her code. In the next seconds the data owner gets a quick confirmation notification about the request that just happened. A moment later the pharmacist come back with her drugs, which she then pays in cash and the transaction is done.

1.3.0.5 Vehicle data and mobility

Assuming a car itself has no hardware on board in order to establish a wireless wide area connection to an outside access node. Only from the inside one can connect to the car (wired or wireless). After entering a car, on the data owner's mobile device pops up a notification asking for permission to connect to that device. In addition to the expiration date, the data owner can choose to en- or disable two more options. First, a wifi network with an uplink to the internet can be provided to everyone inside the car. Secondly, connections,

the car might want to establish, in order to emit data via internet - which, regardless, have to go through the currently linked mobile device. Thus the device owner gains full control over any external data transfer that might happen. This again would allow two things: (A) permission management for all outgoing data and (B) funnel all data generated and provided by the car into the *PDaaS* associated with that linked device. It might also be feasible to deny any connection the car is trying to make. Thus the data will only be stored in the *PDaaS*. If somebody is interested in such then have to ask for access permission. That same concept about movement tracking and vehicle data could also be applied to driving (motor) bicycle.

1.4 Terminologies

Web Service: TODO

Open Specification: TODO

Big Data: deep learning, neural networks

profile data: individual's inherent data; TODO

Personal Data: TODO

Personal Information predominantly static data points related to an individual

Personal Data as a Service (PDaaS): a web service controlled, owned and maybe even hosted by an individual, which provides access to the owner's personal data and offers maintainability as well as permission management.

Personal Data Store: TODO

Vendor Relationship Manager: [13]

Personal Information Management Systems (PIMS): [14]

serverless: TODO https://auth0.com/blog/2016/06/09/what-is-serverless/

Digital Footprints: TODO

Owner: person who controls (and probably hosts) the data service containing her personal data (TODO: or maybe *data source/subject* and *data collector*)

(Data) Consumer: Third party, external entity requesting data, authorized by the owner to do so

Data Broker(s): entities with commercial interests, that collect, aggregate and analyze information/data of any kind - in this case about human beings - from different sources in order to enrich the data sets, to finally license the resulting corpora to other organisations. [15]

2

Fundamentals

The following chapter shall provide the foundational knowledge about concepts like *Personal Identity* or *Big Data* and therefore ensures a common understanding on their relation to the problem this work tries to solve. Additionally it is given a brief overview on what existing standards and technologies might be used, and summarizes the research already been made as well as it's current state.

2.1 Digital Identity, Personal Data and Ownership

- *Digital Identity*
  - what is a *DI*? and in comparison to *Personal Data*?
  - what is required to make the PDaaS used or seen as a *DI*?
- *Personal Data* definition
  - general - freely spoken
  - as of EU law (incl citation)
  - as of US law (incl citation)
  - is it just policy/guideline or enforceable too (law/rule)? what relevance/impact have companies *terms and conditions*?
  - EU and USA (since server might be located outside the state or effective range)
- *Ownership* of personal data
  - who is the owner in what situation or under what circumstances?
  - am I the owner when I was the one who was collecting them? Does it depend on whether the resource was public or somewhat private?
  - what will happen with her data service after a person died?

- A **Digital Identity** is a non-physical abstraction of an entity, such as an organisation, an individual, a device or even software, which allows bidirectional association. In the context of this document, it only refers to human beings. Therefore a *digital identity* is the individual's representation in digital systems, consisting of identity-defining data, such as *personal information* and it's history and preferences [16]. *Personal information*, in this case, refers to inherent (date of birth) and

imposed (credit card number) characteristics.

- From a technical perspective a DI is essentially a collection of characteristics, attributes and time series data (e.g. interaction logs or bank account history). A subset of these attributes combined can form unique fingerprint, like certain single data points (e.g. social security number) in their own context might be, too. Thus it might not be necessary to know the values of all attributes in order to identify a person as the rightful owner and physical counterpart. It can also be seen as an avatar in the digital world or as the digital part of a human's identity. Therefore its important to not view the *DI* as a reduction of a living individual to some bits and bytes, but rather as a appropriate representation for certain purposes and contexts.

- It is also possible to provide an additional level of authenticity insurance for data related to an entity. Therefor an unrelated third party, which needs to be approved not only by the related individual, but also by all entities participating in a context, which might be relevant e.g. for some administration purposes.

- But the concept would also impose a new level of attacking vectors to the identity owner, such as identity theft. The attacker is no longer required to be physically present to be able to steal certain unique identifiers from a person. It is sufficient to gain access to area where the sensitive data is stored.

- In the context of this document and all related work, **Personal Data** is specified as a combination of an individual's *Digital Identity* and all of it's ever created intellectual property (e.g. posts, images, tweets or

comments). This includes all sorts of tracking data and interaction monitoring, as well as metadata manually or automated enriching content (e.g.geo-location attached to a tweet as meta information). Data, captured by someone ore something on or about the individual's private living space and property. Simply every data point reflecting the individual's personality - partly or as a whole - is seen as *personal data*.

- The european *Data Protection Regulations* defining *Personal Data* as follows: > 'personal data' means any information relating to an identified or identifiable natural person > ('data subject'); an identifiable natural person is one who can be identified, directly or > indirectly, in particular by reference to an identifier such as a name, an identification > number, location data, an online identifier or to one or more factors specific to the physical, > physiological, genetic, mental, economic, cultural or social identity of that natural person; > *[17]*

- The U.S.A. has little legislation on defining and protecting consumer's privacy. At least they have no explicit bills addressing such area [18]. Though some of the existing sectoral laws consist of partially applicable policies and guidelines [19]; most of them addressing specific types of data. In 2015 the White House made an attempt to fill the gap with the *Consumer Privacy Bill of Rights Act*, but to this date it didn't passes the draft state. According to the critics, it lags of concrete enforceable rules consumers can rely on [20]. The draft contains a general definition of *Personal Data*: > "Personal data" means any data that are under the control of a covered entity, not otherwise > generally available to the public through lawful means, and are linked,

or as a practical matter > linkable by the covered entity, to a specific individual, or linked to a device that is > associated with or routinely used by an individual, including but not limited to [...] > *[21]*

- followed by a list of concrete data points, e.g. email or postal address, name, social security number and alike. Aside from the legislation with bills, a few third-party organisation can also participate by and add new or overwriting existing rules and policies. Namely for example the *Federal Communications Commission* (FCC), recently releasing *Rules to Protect Broadband Consumer Privacy* including a list of categories of sensitive information [22], which wants *Personally Identifiable Information* (alias Personal Data) to be understood as: > [...] any information that is linked or linkable to an individual. [...] information is > "linked" or "linkable" to an individual if it can be used on its own, in context, or in > combination to identify an individual or to logically associate with other information about a > specific individual. > *[23]*

- Despite minor difference in detail, they all have similar ideas of personal data and their belonging. Even though, the version proposed by EU is almost identical with the definition introduced for the context of this work. Although the FCC's statutory authorities might be somewhat debatable regarding certain topics, the *Communications Act* as a U.S. federal law equips the FCC with power to regulate and legislate.

- Having a common opinion on what data points are belonging to person is the foundation to define a set of rules on how deal with *Personal Data* accordingly. Every business, operating within the EU, is required[1]

[1]according to article 12-14 of the *EU General Data Protection Regulation 2016/679*

to provide it's users with a *Privacy Policy*, while e.g. in the U.S. - as mentioned above - only partially and depending on context and data type users must be informed about which and how their data get processed [24].

- A user commonly agrees on the privacy policy, by starting to interact with the author's business, thus every *Privacy Policy* is required to be publicly accessible; e.g. before creating an account. > By clicking Create an account, you agree to our Terms[2] > and that you have read our Data Policy[3], including > our Cookie Use[4]. > *[web_2016_facebooks-landing-page_policy-acknowledgement]*

- It can be seen more likely an information notice, that translates and specifies general given law, rather then a contract.

- With such knowledge at hand, it is up to each individual, if the service's benefits are worth sharing some personal data, while simultaneously acquiescing potential downsides concerning the privacy of such data.

- Every entity who is doing so, muss process Personal data according to the law and their *Privacy Policy*. If they policies are violating existing law or the entity effectively goes against the law with their actual doing, penalties might follow. Depending on the level and impact of their infringement in addition the law itself, aside from revising their wrongdoings the entity might have to compensate the affected individuals, pay a fine or get revoked their license.

[2]https://www.facebook.com/legal/terms
[3]https://www.facebook.com/about/privacy
[4]https://www.facebook.com/policies/cookies/

- Not only privacy laws, but every legal jurisdiction has it's limitations - concerning their territorial nature - which makes legislation not exactly an appropriate tool when it comes to fixing existing issues and strengthen the individual's privacy and rights in a global context like the *world wide web*. If no international agreement is in place [25], only those laws are considered valid and enforcible where the organisation is registered, and maybe the fact where (meaning in which area of jurisdiction) the their servers are located or the data is processed and stored.

Whereas **Ownership** of *Personal Data* has no legal ground foundation so ever. The concepts of intellectual property protection and copyright might intuitively be applicable, because the data, created by its owner, seems to be her *intellectual property*. Such property implies to be a result of a creative process though, but unfortunately there is no *threshold of originality* in facts, like *personal information* is [26].

- Ownership in the sense of having exclusive control over it's personal data and how they get processed at any given point in time; this not only comes with high costs, but is also very inconvenient for both parties - owner and data consumer. It consists of two concepts: (A) the right to do what every is desired with their property and (B) in which rules and mechanisms the ownership can be assigned to someone [27].

- The european DPR[5] contains only one occurrence of the word *ownership*, which is not even related to the context of *personal data* or the

[5]EU Data Protection Regulation

data subject. It only stats, that *"Natural persons should have control of their own personal data."* [28]. Whereas Commissioner J. Rosenworcel of the FCC wants *"consumers [...] to [...] take some ownership of what is done with their personal information."* [29]

- Typically the question of data ownership is addressed in data consumer's *Terms of Service* (ToS), which an individual might have to accept in order to establish a (legal) relationship with it's author. I should be kept in mind, that *ToS* might change over time; not necessarily to the users advantage. All addressed issues (by the ToS) must not violate any applicable or related law, otherwise the *ToS* might not be legally recognized. Taking the following excerpts from different *ToS*:

  > You own all of the content and information you post on Facebook, and you can control how it is shared [...]. *(under "2. Sharing Your Content and Information", by Facebook [30])*

  > You retain your rights to any Content you submit, post or display on or through the Services. What's yours is yours — you own your Content. *(under "3. Content on the Services", by Twitter [31])*

  > Some of our Services allow you to upload, submit, store, send or receive content. You retain ownership of any intellectual property rights that you hold in that content. In short, what belongs to you stays yours. *(under "Your Content in our Services", by Google [32])*

  > Except for material we may license to you, Apple does not claim

ownership of the materials and/or Content you submit or make available on the Service "(under"H. Content Submitted or Made Available by You on the Service", by Apple [33])*

All these statements are followed by the same term, stating that the user grants the author a worldwide license to do almost any imaginable thing with her data. This even applies to Apple, if the user is *"submitting or posting […] Content on areas of the Service that are accessible by the public or other users with whom [the user] consent to share […] Content"* [33].

- It is worth noticing, that in every *ToS* it is only referred to the data owner's content, not all her personal data. As mentioned above, personal information are no intellectual property, but playing an important role in data analytics though. Which is why *privacy policies* are in place, to ensure at least some user enlightenment, even though it doesn't compensate the lack of control.

- In addition to that, the meaning of *ownership* used in the quoted *ToS* is missing a clear outline and thus causing ambiguity and leaving room for interpretation. Nor the actual definition of *ownership*, as described earlier, is applicable for these kind of cases, since the user losing all its control is by design. Handing over data to the consumer annihilates the exclusive control over the data and revokes the ability of assigning such control. There is no (legislation based) way to establish a feasible concept of *ownership*, if the data consumer has no motivation to promote the user a comprehensive owner of her data.

- Leaving all the legal layer aside for a moment and switching the perspectives a bit; Data consumers might argue, that they had invested

in enabling themselves to collect, process and store personal data, so it belongs to them. But from the data owner's point of view it might only be the case as long as as she would benefit as well somehow, e.g. using products, services or features, offered by consumers, which quality depends on personal data. If the data owner chooses to move to a competitor might what to bring her personal data with her. But then again the former data consumer would object, competitors would benefit from all investments the consumer has made, but without any effort. Though, not entirely wrong, two aspects need to be emphasize. (A) In order to archive a high level of quality for their analytics and therefore in making right decisions to gain improvement, it's vital to huge amount of effort in developing these underlying technologies, not only in acquiring personal data. Which again only constitutes (B) the foundation of various subsequential computations followed by an ongoing collecting, aggregation and analytics of actively and passively created data and metadata (e.g. food deliver history or platform interactions and tracking). Given the initially introduced definition of *personal data* only a fraction of the involved data belongs to its owner. The large part consists of highly valuable metadata [34] [35] and therefore should remain to the data collector and either be deleted or sufficiently anonymized, if the owner cancels the relationship. The data owner should not depend on the collector's willingness when it comes to handing over her personal data (e.g. list of favorites or delivery history). Instead, using her own tool to provide the consumer with required data (e.g. list of favorites) or tap into her data creating interactions (e.g. food deliveries) on her own.

- Whether an individual dies or a user deletes her account, as long as certain data point are shared with / connected to other users, the data will remain. At least when it comes to facebook.
- Generally speaking, all data solely associating with an individual, is in the ownership of the same. But since it doesn't exist any legal concepts on *personal data* ownership, a technical solution could help to regain some control.

2.2 Personal Data in the context of the Big Data Movement

- big data itself initially can be seen as a *huge blob of data* containing more or less structured data sets [36], whose size might have exceeded the capabilities of retrieving certain information almost only by hand. Such high data haystacks usually come along with new challenges in logistic and resource management, when information retrieval needs to get automated on a large scale [37]. Theses practices are commonly referred to *Big Data (Analysis)* including distributed computing and machine learning.
- Big Data, or to be more precise, collecting and analyzing big data, serves the prior purpose to extract useful information, which on the other hand depends on what was the opening question about, but also what data sets the corpus is containing.
- At first, (A) formalizing question(s) that the results have to answer.

Secondly, (B) deciding what data is needed and appropriate and then start collecting. Third, (C) designing data models accordingly and correlate with the data (D) next, analyse and interpret the results. (E) last but not least, make business decisions based und the analyses ([38] Fig. 3).

- since quite a few businesses (in terms of purpose or intention) are based around the concept of customers, which are generally somewhat entities consisting of at least one human being, personal data takes a major part in what *Big Data* can be about. In the context of this thesis, these entities are individuals with a unique identity. And to understand the behaviour, decision making and needs of her customers a vendor, who owns the business, needs to know as much as possible about them, when she wants to know what changes she needs to address in order to move towards the most lucrative business.

- personal data and information are reflecting all this knowledge. It starts with profile data, such as gender, age, residency or income, goes on with time series events like geo-location changes, or web search history and goes all the way up to health data and self-created content like *Tweets*[6] or videos.

- all these classes of personal data hold a major share[7] in the field of data analytics (TODO: find statistics showing shares of data

[6]public massages published by an account on twitter.com, which will be displayed in the timeline of all her subscribers and also might contain additional types of content like images, links or video

[7]it doesn't matter whether an individual or just someone on behalf of an organisation spend money for something. at the end of the day, they are all humans on this planet and in a capitalistic oriented world money needs to flow and profits needs to be maximized. So to know where it will flow or why it will flow in a certain direction it is crucial to know everything about it's decision maker - the humans on this planet.

types/classes/categories, [39] [40])

- but, depending on the specific attributes, they might be not that easy to acquire. in general most businesses obtain data from within their own platforms. some data might be in the user's rang of control (e.g. customer or profile data), but most of the data comes from interacting directly (content creation, inputs) or indirectly (transactions, meta information). the level of sensitivity is mainly based on the purpose of the platform (benefit for the user) and what is the provider's demand from the users commitment (e.g. required inputs or usage requires access to location)

- from a technical perspective collecting passively created data is as simple as integrating logging mechanisms in the program logic. since the industry moved towards the cloud[8] most scenarios utilized server-client architectures. Furthermore the *always-on* philosophy evolved to an imperative state. standalone software is starting to call the author's servers from time to time, just to make sure the user behaves properly. For browsers it was already a common narrative to make here and then requests to the server - still preventable though, but when it comes to native mobile apps it is almost impossible [41] to notice such behaviour and therefore preventing apps from doing so.

- these architectural developments were inducing the gathering of potentially useful information from all over the system on a large scale [42]. Logging events, caused by the user's interactions, on the client, which

[8] side note - one might come to the conclusion, that only the trend towards the *cloud* made it actually possible to collect to such an extent we are all observing these days, because standalone software should not necessarily require internet connection and therefore the vendors had no way to gather information whatsoever

then get forwarded to backend servers. Or keeping track of all kinds of transactions, which is done directly in the backend. Before running together in a designated place, all these collected chucks of data (TODO or "data points") are getting enriched with meta information. Finally get stored and probably never removed again - all for later analyses.

- The mindset in the *Big Data Community* is grounded on the basic assumption of *more data is more helpful*, which already is emphasised by the often-cited concept of the three *Vs* (Volume, Velocity, Variety) [43]. which is not entirely wrong, because it lies in the nature of pattern and correlation discovery, to provide increasing quality results [44], while enriching the overall data with more precise data sets. But when new technologies are emerging, questioning the downsides and possible negative mid- or long-term impacts are typically not very likely to be a high priority. The focus lies on e.g. trying to to reach and eventually breach boundaries while beginning to evolve. So non-technical aspects such as privacy and security awareness doesn't come in naturally, instead a wider range of research needs to be done alongside the evolution process and the increasing adoption rate in order to uncover such issues. Only then they can addressed properly on different levels - technical, political as well as social. So that the *Big Data Community* itself is able to evolve, too. All in all it's a balancing act between respecting the user's privacy and having enough data at hand to satisfy the initial questioning with the computed results. Therefore people working in such contexts need to have advanced domain knowledge, be aware of any downsides or pitfalls and need to be sensible about the ramifications of their approaches and doings. Such improvements are already

happening, not only originating from the field's forward thinkers [45], but also advocated by governments, consumer rights organisations and even leading Tech-Companies start trying to do better [46] [47] [48] - as discussed in the section [TODO see personal data as of the law],

- earlier in the text a difference was made between actively created and passively created data
- based on that one could say *profile/account data* is actively created, because it got into the system by the user's actively made decision to insert these information into a form and submit it - for whatever reason. whereas detecting the user's current location and adding this information to the submitted form is *meta data*
- of cause, it is debatable whether these kind of data belongs, in the sense of being the rightful owner, to the user or to the author or owner of the software containing the code that effectively created the data.
- maybe personal data is every data/information whose creation (or digital existence) is a direct result of user interaction/engagement?
- lets have a look into what the rule book says about that –> next topic (law)

2.3 Personal Data as a Product

- *Big Data Analytics* by itself just comprises a structured and technical-aided procedure, serving the purpose of finding invisible information, that might be helpful to make (right) (business) decisions. Though, if

one would ask data collectors about their motivation, most likely the answer would be something along the lines of PR phrasing like *"We want to have a better understanding of our customers"*. But to do what exactly? To predict what might be the next thing I am supposed to buy Or what things I probably would like to consume but most certainly not yet know of?

- Let's take a look at some examples. An advertising service uses tracking data for targeted advertising. The more information they have about an individual, the more accurate decisions they are able to make about what ads are the ones the individual most likely will click on and disclose with a successful purchase. As a result this makes the placed advertisement more valuable for ad service and therefore more expensive to the advertisers, because of a high precision. Or a streaming provider's content recommendation is also based on heavy user profiling done by looking at her consumption history, tracked platform interactions and probably many more vectors. Another example is *Google Traffic* [49] [50], a service, integrated as a feature in *Google Maps*, which is Google's web mapping service. *Google Traffic* visualises real-time traffic conditions, when using *Maps* as a navigation assistant, to provide the user with a selection of possible paths, but enriched with duration, that takes such conditions into account. The data, required to offer these information, is supplied by mobile devices, constantly sending GPS coordinates with a timestamp into Google's infrastructure. This, however, only is made possible, because Google's services are widely used in addition to the fact that the majority of mobile devices [51] is driven by Android, an mobile operating system

developed by Google, that deeply integrates with it's services. For this case the same assertion can be made - the more constantly streaming geo-location data, the more precise the information are about traffic conditions. Since this information demands the real-time aspect, adding time to the equation, add a other dimension of complexity to problem.

- while the impact on our society of this first example group might be doubtable, a change of perspective opens up a different range of application areas. Such as
  - planing and managing human resources for situations, like e.g. big events or emergency situations where attendees might need some help [52]
  - predicting infrastructure workloads [TODO http://ieeexplore.ieee.org/document/7336
  - making more accurate diagnostics to improve their therapy [53]
  - finding patters in climate changes, which otherwise wouldn't be detected [54].
- Through all these examples, some of them might not necessarily founded on personal data, whereas others primarily depend on them and yet others only implicitly rely on data collected from individuals. As always, it depends on the purpose - also known as *business model* - but it seems to be consensual, that it all comes down to improving and enhancing the collector's product in order to satisfy the customers - and that on the other hand depends on what is meant to be the product and who is seen as customers.
- Putting a top 10 list of industries using utilizing *Big Data* [55] right

next to visualization showing categories of personal data targeted by data collectors

[56], at least 7[9] of these industries can be identified as data collectors, whereas less then a half[10] are taking part of being a *Data Broker*, but almost all of them are using people's personal data, whether collected by themselves or acquired from *Data Broker*.

- At this point it's save to say, that *Personal Data* is either seen directly as a product, especially from a Dater Broker's point of view, or indirectly due to it's essential part in *Big Data* practices. The former generates direct revenue by selling these data and the latter might affect a business's product quality in a positive manner and thereby increasing revenue as well.

- At the end it all comes down to understanding the human being and why she behaves as she does. The challenge is not only to compute certain motives but rather concluding to the right ones. When analyzing computed results with the corresponding data models and trying to conclude, it is important to keep in mind, that correlation is by far no proof of causation.

- individuals then get in role of selling/offering it's own data to those who were previously collecting them

[9]Banking and Securities; Communication, Media & Entertainment; Healthcare Providers; Government; Insurance; Retail & Wholesale Trade; Energy & Utilities

[10]Banking and Securities; Communication, Media & Entertainment; Insurance; Energy & Utilities

2.4 Related Work

The idea of a digital vault, controlled and maintained by it's owner, the individual, isn't that new. Holding her most sensitive and valuable collections of bits and bytes, protected from all these data brokers and authorities, while interacting with the digital and physical world, opening and closing it's door from time to time, to either put something important for her inside or retrieving an information important for someone else. While in the mid and late 2000s the growth of computer performance and capacity were crossing it's zenith (see Moore's Law [57]), at the same time the internet was starting to become a key part in many people's lives and in society as a whole. Facilitated by these circumstances, *cloud computing* has been on the rise, causing the shift towards parallel distributed processing and patterns alike. Thereby making it possible to rethink solutions from the past and trying to go new ways, namely the breakthrough 2007 in *neuronal networks* cutesy of G. Hinton [58]. As a result, fields like *deep machine learning*, *big data analytics* and most recently *data mining*, were gaining a wide range of attention. In almost any industry a greater amount of resources is invested in these areas [59].

The initial research motivation can be seen as a counter-movement away from the *cloud*, starting to focus again on privacy, the individual and it's digital alter ego.

From simple middleware-solutions, via full-fledged software-based platforms, through embedded hardware devices, a great variety of approaches were starting to appear in the mid 2000s until this day. A side effect was, that over time various research teams and projects have invented and coined

different terms, all referring to the same concept. The following list shows some examples *(alphabetical order)*:

- Databox
- Identity Manager
- Personal …
  - Agent
  - Container
  - Data Store/Service/Stream (PDS)
  - Data Vault
  - Information Hub
  - Information Management System (PIMS)
- Vendor Relationship Management (VRM)

2.4.1 Research

One of the first research projects is *ProjectVRM*, which originated from *Berkman Center for Internet & Society* at *Harvard University*. As it's name implies, it was inspired by the idea of turning the concepts of a *Customer Relationship Management* (CRM) upside down. This puts the vendor's customers back in charge of their data priorly managed by the vendors. It also solves the problem of unintended data redundancy. Over time the project has growing to the largest and most influential in this research field. It transformed into an umbrella and hub for all kinds of projects and research related to that topic [60], whether it's frameworks or standards, services offering e.g. privacy protection, reference implementations, applications, software or hardware components. *VRM* became more and more a synonym for a set

of principles [61], including for example *"Customers must have control of data they generate and gather. [They] must be able to assert their own terms of engagement."* These principles can be found in various ways across a lot of research done within this area.

Another research that is worth mentioning, because of the foundational work it has been done, is the european funded project called *Trusted Architecture for Securely Shared Service* (TAS3). The project led to a open source reference implementation called *ZXID*.[11] The major goal was, to develop an architecture, that takes all involved parties into account, whether it's commercial businesses (vendors) or it's users (customers), in order to fit into more sophisticated and dynamic processes, but at the same time demanding a high level of user-centric security facilitate i.a. by a developed policy framework. Due to these requirements the architecture ended up being rather complex [62]. *ZXID* as it's implementation incorporates several standards like SAML 2.0[12] and XACML,[13] has only three third-party dependencies which are *OpenSSL*, *cURL (libcurl)* and *zlib* and as of now it supports Java, PHP and Perl. The project lasted for a period of 4 years, but after it ended in 2011, the research work has pursued i.a. by the *Liberty Alliance Project*, which is now part of the *Kantara Initiative* [63], including all documents and results. These results were taken up occasionally, recently from the IEEE [64].

A research project, which is probably the closest to what this document aims to create, bears the name *openPDS* [65] and is done by *Humans Dynamics Lab* [66], which is part of *MIT Media Laboratories*. Despite the usual con-

[11] more information on the project, the code and the author, Sampo Kellomäki, can be found under *zxid.org*

[12] Security Assertion Markup Language 2.0

[13] eXtensible Access Control Markup Language

cepts of a *PDS*, it introduces multi-platform components and user interfaces including a mobile devices and separating the persistence layer physically at the same time. This facilitates administrative tasks regardless of the data owner's position and time. Moreover, with their idea of *SafeAnswers* [67], the team even goes a step further. The concept behind that, is based around *remote code execution*, briefly described in one of the user stories during the first chapter. It abstracts the concept of a data request to a more human-understandable level, a simple question. This question consists of two representation: (A) a short explanation of what the data consumer wants to know and which data might be involved and thus what information a data consumer actually will receive, instead of raw data the consumer could then use for all kinds of purposes e.g. data aggregation or mining. Aside from that, the request payload also includes (B) a code-based representation, which gets executed in a sandbox on the data owner's *PDS* system with the necessary data as arguments. The resulting output is answer and response all in once.

Aside from all the research projects done within the scientific context, applications with a commercial interest were starting to occur in a variety of sectors, too. Microsoft's HealthVault [68], for example, which aims to replace all the paper-based patient file and combine them in one digital version. This results in a patient-centered medical data and documents archive, helping doctors to make the most accurate decisions on medical treatment.

Meeco [69] [70], based on the MyData-Project [whitepaper_2014_mydata-a-nordic-model-for-human-centered-personal-data-management-and-processing], which essentially just cuts out the advertisement service

provider as a middle man inherits that role by itself. The platform does provide the data owners with more control over what information they reveal, but it doesn't go the so eagerly demanded next step, which would means real decoupling from the advertisement market and finding a suitable business model that focuses on the data owner, instead of surrounding them with just another walled garden.

A recently announced project, sponsored by Germany's *Federal Ministry of Education and Research*, but developed and maintained primarily by *Fraunhofer-Gesellschaft* in cooperation with several private companies like *PricewaterhouseCoopers AG*, *Volkswagen AG*, *thyssenkrupp AG* or *REWE Systems GmbH*, is the so called *Industrial Data Space* [71]. The project unifies both, research and commercial interests and runs over time period of three years until the third quarter of 2018. It aims to *"[…] to facilitate the secure exchange and easy linkage of data in business ecosystems"*, where at the same time *"[…] ensuring digital sovereignty of data owners"* [72]. It will be interesting to see how these two, yet rather distinct objectives, will come together in the future. Based on the white paper, the project's focus mainly seems to lie in enabling and standardizing the way companies collect, exchange and aggregate data with each other across process chains to ensure high interoperability and accessibility.

Hereafter a selective list can be found of further research projects, work and commercial products regarding the issue around *personal data*:

2.4.1.1 Research

- Higgins [https://www.eclipse.org/higgins/]
- Hub-of-All-Things [http://hubofallthings.com/what-is-the-hat/]
- ownyourinfo [http://www.ownyourinfo.com]
- PAGORA [http://www.paoga.com]
- PRIME/PrimeLife [https://www.prime-project.eu, http://primelife.ercim.eu/]
- databox.me (reference implementation of the "solid" framework)
- Polis (greek research project from 2008) [http://polis.ee.duth.gr/Polis/index.php]

2.4.1.2 Organisations

- Open Identity Exchange [http://openidentityexchange.org/resources/white-papers/]
- Qiy Foundation [https://www.qiyfoundation.org/]

2.4.1.3 Commercial Products

- MyData [https://mydatafi.wordpress.com/]
- RESPECT network [https://www.respectnetwork.com/]
- aWise AEGIS [http://www.ewise.com/aegis]

2.5 Standards and Specifications

The overall attempt is to involve as much standards as possible, because it increases the chances of interoperability and thereby it lowers the effort, that

needs to be made,in order to integrate with third parties or other APIs.

- http(s)
- all the *Semantic Web* stuff
- Container/App spec
- JWT
- oAuth (?)
- JSON
- REST
- GraphQL

3

Core Principles

In the following chapter the certain core principles of the system will be examined

NOTE: here we discuss a variety of possibilities –> conceptual work

3.1 Data Ownership

- user-centric, full control

3.2 Identity Verification

- maybe go with a Signing/verifying Authority (aka CA)
  - do I trust the gov or certain companies more? Which interests do these Role/Stakeholder have?
  - revoking the cert which provides the authenticity of the individual's digital identity should only be possible with a two-factor secret. One part of this secret is owned by the CA and the other half has the individual behind the personal API
- TODO: look into
  - PKI
  - ePerso
  - E-Post/de-mail
- Authentication

3.3 Authorisation (remove in favour of data access?)

- NOTE: does not mean this tool authenticates it's owner against third party platforms like OpenID does. but it could play the role of the 2n factor in a multi-factor authentication process (if the mobile-device-architecture was chosen)
- refers primarily to the process of a data consumer (third party, which needs the data for whatever reason) verifies her admission to request

3.4 Authentic Data

- is this data (in this case identity) certified or not (results in higher value)

3.5 Supervised Data Access

- pure/plain data request/resonse
- remote computation/execution (assuming there is no client for the consumer) like https://webtask.io/

3.6 Encapsulation

- containerization (coreos, rkt, mirageos aka unikernal)

3.7 Open Development

- which and why open standards
- why open source
- collaborative transparent development
- Hosting & Administration
  - DYI
  - Usability

4

Requirements

4.1 User Interaction

- as effortless as possible

4.2 Management & Organisation

4.3 Administration

5

Design Discussion

major design decisions will be discussed and views from several perspectives

5.1 Architecture

- showing possible directions, e.g.:
  - cloud or local storage
  - which components can go where
  - remote execution, to prevent data from leaving the system

5.1.1 Overview

- distributed architecture (e.g. notification/queue server + mobile device for persistence and administration)

5.1.2 Components

5.1.3 Plugins

- but for what? and not harm security at the same time? maybe just for data types and admin UI to display analytical (time based) data in other ways
- what about extensions (see iOS 10) to let other apps consume data; only on a mobile device and only if the data is stored there

5.2 Data

- keep in mind to make it all somehow extendible, e.g. by using and storing corresponding schemas
- NOTE: step numbers marked with a * are somehow tasks which are happening in the background and don't require any user interaction

5.2.1 Modelling

5.2.2 Categories (or Classes)

5.2.3 Types

5.2.4 Persistence

- database requirements

5.2.5 Access & Permission

- data needs to have an expiration date

IF01 - Authorizing a consumer to request certain data

1) owner creates a new endpoint URI (like *pdaas.ownersdomain.tld/e/consumer-name*) within the *management user interface*
2) owner passes this URI on to the *consumer*, e.g. through submitting a form or using any arbitrary, eventually insecure channel 3*) consumer need to call this URI for the fist time to verify it's authenticity
3) owner then get's a notification which asks her for permissions to access certain data under the listed conditions 5*) consumer will be informed about the outcome of the owner's decision (NOTE: alongside with some details? how do they look like? XXX need to be in the spec)

5.2.6 Consumption (data inflow)

- how data will get into the system
- hwo is the user able to do that, and how does it works

5.2.6.1 Manually

5.2.6.2 Automatically

5.2.7 Emission (data outflow)

- depending on what category of data, they might need to get anonymized somehow before they leave the system
- oAuth (1.0a and 2) requires consumers to register upfront. Since the current flow indicates that the initial step is done by the owner, that would cause an overhead in user interactions. Although the owner already *authorized* the consumer simply by submitting a unique URI (`pdaas-server.tld/register?crt=CONSUMER_REGISTER_TOKEN`), of which the `crt` is considered private. Even though the registration provides the consumer with mandatory information such as a consumer identifier (v1: `oauth_consumer_key`, v2: `client_id`) and, depending on the client type, a secret (see https://tools.ietf.org/html/rfc6749#section-2), this process it is not part the specification (https://oauth.net/core/1.0a/#rfc.section.4.2, https://tools.ietf.org/html/rfc6749#section-2). This enables the possibility of integrating oAuth into the consumer registration flow

by using the `CONSUMER_REGISTRATION_TOKEN` as oAuth's *client identifier*. The lack of credentials (v1: `auth_consumer_secret`, v2: `client_secret`) would require transferring the consumer identifier done over a secure channel (e.g. TLS). That would leave *oAuth2* as the version of choice, since it relies on *HTTPS* adn therefore makes the *secret* optional. Where on the other side oAuth 1.0a requires a *secret* to create a signature in order to support insecure connections..

- A general and URI for 3rd parties to register (aka requesting authentication) would raise the issue of dealing with spam request and how to distinct these from the actual ones.

5.2.8 History

- data versioning
- access logs

5.3 Interfaces

5.3.1 Internal

- UI for Management & Administration

5.3.2 External

- should there be a way to somehow request information about what data is available/queryable, or would this be result in spam/crawler and security issues (also a question for the topic of permissions/sensibility level of certain data)
- certain types of requests, depending on expire date:
  - "ask me any time"
  - "allowed until further notice"
  - one-time permission (but respecting certain http error codes and possible timeout - that might not count)

6

Specification

- what does *open* in Open Specification even mean?

6.1 Processes (TODO: find another word; “Protocol flows”?)

6.2 Application Programming Interfaces

6.3 Graphical User Interfaces

6.4 Security

- the downside of having not just parts of the personal data in different places (which is currently the common way to store), is in case of security breach, it would increase the possible damage by an exponential rate Thereby all data is exposed at once, instead of not just the parts which a single service has stored
- does it matter from what origin the data request was made? how to check that? is the requester's server domain in the http header? eventually there is no way to check that, so me might need to go with request logging and trying to detect abnormal behaviour
- is the consumer able to call the access request URI repeatedly and any time? (meaning will this be stateless or stateful?)
- initial consumer registration would be done on a common and valid https:443 CA-certified connection. after transferring their cert to them as a response, all subsequent calls

need to go to their own endpoint, defined as subdomains like `consumer-name.owners-notification-server.tld`

6.4.1 Environment

6.4.2 Transport

- communication between internal components *must* be done in https only, but which ciphers? eventually even http/2?

6.4.3 Storage

- documents based DB instead of Relational DBS, because of structure/model flexibility

6.4.4 Authentication

- how should consumer authenticate?

6.5 Recommendations

6.5.1 Software Dependencies

6.5.2 Host Environment

7

Conclusion

7.1 Ethical & Social Impact (TODO: or "Relevance")

7.2 Business Models & Monetisation

- possible resulting direct or indirect business models
- owner might want to sell her data, only under her conditions. therefore some kind of infrastructure and process is required (such as payment transfer, data anonymization, market place to offer data)

7.3 Challenges

- adoption rate of such technology

7.4 Solutions

7.5 Attack Scenarios

7.6 Future Work

- maybe enable the tool to play the role of an own OpenID provider?
- going one step further and train machine (predictor) by our self with our own data (https://www.technologyreview.com/s/514356/stephen-wolfram-on-personal-analytics/)

7.7 Summary

- main focus
- unique features
- technology stack & standards
- resources
- the tool might be not a bulletproof vest, but

[1] "Big data privacy international." [Online]. Available: https://www.

privacyinternational.org/node/8. [Accessed: 15-Nov-2016]

[2] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 560–568 [Online]. Available: http://dl.acm.org/citation.cfm?id=1401959. [Accessed: 03-Nov-2016]

[3] S. Spiekermann, *Ethical IT Innovation: A Value-Based System Design Approach*. CRC Press; Taylor & Francis Group, LLC, 2015, pp. 66–72 [Online]. Available: https://www.crcpress.com/Ethical-IT-Innovation-A-Value-Based-System-Design-Approach/Spiekermann/p/book/9781482226355

[4] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Transactions on Information Systems (TOIS)*, vol. 14, no. 3, pp. 330–347, 1996 [Online]. Available: http://dl.acm.org/citation.cfm?id=230561. [Accessed: 07-Nov-2016]

[5] "Cognitive bias," *Wikipedia*, Oct-2016. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Cognitive_bias&oldid=742803386. [Accessed: 08-Nov-2016]

[6] R. Lemov, "Why big data is actually small, personal and very human. Aeon essays," 16-Jun-2016. [Online]. Available: https://aeon.co/essays/why-big-data-is-actually-small-personal-and-very-human. [Accessed: 17-Nov-2016]

[7] A. Dewes, "C3TV - Say hi to your new boss: How algorithms might soon control our lives." 29-Dec-2015. [Online]. Available: https://media.ccc.de/v/32c3-7482-say_hi_to_your_new_boss_how_

algorithms_might_soon_control_our_lives#video&t=1538. [Accessed: 03-Nov-2016]

[8] "ProjectVRM - about. ProjectVRM," 25-Feb-2010. [Online]. Available: https://blogs.harvard.edu/vrm/about/. [Accessed: 09-Nov-2016]

[9] Tom Kirkham, Sandra Winfield, Serge Ravet, and S. Kellomaki, "The personal data store approach to personal data security," *IEEE Security & Privacy*, vol. 11, no. 5, pp. 12–19, 2013.

[10] A. Poikola, K. Kuikkaniemi, and H. Honko, "MyData – a nordic model for human-centered personal data management and processing," pp. 1–12, Jun. 2015 [Online]. Available: https://www.lvm.fi/documents/20181/859937/MyData-nordic-model/2e9b4eb0-68d7-463b-9460-821493449a63. [Accessed: 10-Nov-2016]

[11] "Meeco how it works." [Online]. Available: https://meeco.me/how-it-works.html. [Accessed: 09-Nov-2016]

[12] "Open specification of the concept called personal data as a service (pdaas). GitHub." [Online]. Available: https://github.com/lucendio/pdaas_spec. [Accessed: 11-Nov-2016]

[13] "ProjectVRM wiki - about VRM." [Online]. Available: https://cyber.harvard.edu/projectvrm/Main_Page#About_VRM. [Accessed: 11-Nov-2016]

[14] "ProjectVRM wiki - list of personal information management systems." [Online]. Available: https://cyber.harvard.edu/projectvrm/VRM_Development_Work#Personal_Information_Management_Systems_

.28PIMS.29. [Accessed: 11-Nov-2016]

[15] F. T. C. USA, "Data brokers," May 2014 [Online]. Available: https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-re 140527databrokerreport.pdf. [Accessed: 17-Nov-2016]

[16] J. Rose, O. Rehse, and B. Röber, "The value of our digital identity," *Boston Cons. Gr*, 2012 [Online]. Available: https://www.libertyglobal.com/PDF/public-policy/The-Value-of-Our-Digital-Identity.pdf

[17] *General data protection regulation.* 2016, p. L 119/33 [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

[18] Wikipedia, "Information privacy law," 13-Nov-2016. [Online]. Available: https://en.wikipedia.org/wiki/Information_privacy_law#United_States. [Accessed: 20-Nov-2016]

[19] I. J. (Loeb & Loeb), "PLC - data protection in the united states: Overview," 01-Jul-2013. [Online]. Available: http://us.practicallaw.com/6-502-0467. [Accessed: 20-Nov-2016]

[20] A. Wilhelm, "White house drops 'consumer privacy bill of rights act' draft. TechCrunch," 27-Feb-2015. [Online]. Available: http://social.techcrunch.com/2015/02/27/white-house-drops-consumer-privacy-bill-of-rights-act-draft/. [Accessed: 20-Nov-2016]

[21] *Administration discussion draft: Consumer privacy bill of rights act of 2015.* 2015 [Online]. Available: https://www.whitehouse.gov/sites/default/

files/omb/legislative/letters/cpbr-act-of-2015-discussion-draft.pdf

[22] *Report and order.* 2016 [Online]. Available: https://transition.fcc.gov/Daily_Releases/Daily_Business/2016/db1103/FCC-16-148A1.pdf. [Accessed: 20-Nov-2016]

[23] *Notice of proposed rulemaking.* 2016 [Online]. Available: https://apps.fcc.gov/edocs_public/attachmatch/FCC-16-39A1.pdf. [Accessed: 20-Nov-2016]

[24] "Privacy policies are mandatory by law," 23-Oct-2016. [Online]. Available: https://termsfeed.com/blog/privacy-policy-mandatory-law/. [Accessed: 20-Nov-2016]

[25] "International privacy standards," 29-Sep-2016. [Online]. Available: https://www.eff.org/issues/international-privacy-standards. [Accessed: 20-Nov-2016]

[26] G. Rosner, "Who owns your data?" presented at the UbiComp '14, september 13 - 17 2014, seattle, wa, usa, 2014, pp. 623–628 [Online]. Available: http://dl.acm.org/citation.cfm?doid=2638728.2641679. [Accessed: 01-Dec-2016]

[27] J. Grunebaum, *Private ownership.* Routledge & Kegan Paul, 1987, p. 213.

[28] *General data protection regulation.* 2016, p. L 119/12 [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

[29] *Report and order.* 2016 [Online]. Available: https://transition.fcc.

gov/Daily_Releases/Daily_Business/2016/db1103/FCC-16-148A1.pdf. [Accessed: 20-Nov-2016]

[30] Facebook, "Facebooks's terms of service. Statement of rights and responsibilities," 30-Jan-2015. [Online]. Available: https://www.facebook.com/legal/terms. [Accessed: 01-Dec-2016]

[31] Twitter, "Twitters's terms of service. Twitter terms of service," 30-Sep-2016. [Online]. Available: https://twitter.com/tos#intlTerms. [Accessed: 01-Dec-2016]

[32] Google, "Google's terms of service. Google terms of service," 30-Apr-2014. [Online]. Available: https://www.google.com/intl/en/policies/terms/regional.html. [Accessed: 01-Dec-2016]

[33] Apple, "Apple's iClound terms and conditions. V. content and your conduct," 25-Sep-2016. [Online]. Available: https://www.apple.com/legal/internet-services/icloud/en/terms.html. [Accessed: 01-Dec-2016]

[34] "Why metadata matters," 07-Jun-2013. [Online]. Available: https://www.eff.org/deeplinks/2013/06/why-metadata-matters. [Accessed: 24-Nov-2016]

[35] J. P. Stevens, "Why you need metadata for big data success," 06-Apr-2016. [Online]. Available: http://www.datasciencecentral.com/profiles/blogs/why-you-need-metadata-for-big-data-success. [Accessed: 24-Nov-2016]

[36] "Big data n." [Online]. Available: http://www.oed.com/view/Entry/

18833#eid301162177. [Accessed: 11-Nov-2016]

[37] Wikipedia, "Big data," 11-Nov-2016. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Big_data&oldid=748964100. [Accessed: 11-Nov-2016]

[38] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, "Big data analytics: A survey," *Journal of Big Data*, vol. 2, no. 1, p. 21, Oct. 2015 [Online]. Available: http://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0030-3. [Accessed: 13-Nov-2016]

[39] O. R. Zaïane, *Principles of knowledge discovery in databases*. 1999, pp. 1–2 [Online]. Available: https://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/. [Accessed: 13-Nov-2016]

[40] "Big data collection collides with privacy concerns, analysts say. PCWorld," 10-Feb-2013. [Online]. Available: http://www.pcworld.com/article/2027789/big-data-collection-collides-with-privacy-concerns-analysts-say.html. [Accessed: 15-Nov-2016]

[41] "Answers.io. Answers." [Online]. Available: https://answers.io/answers. [Accessed: 14-Nov-2016]

[42] A. L. Burgelman, N. L. Burgelman, and NGDATA, "Attention, big data enthusiasts: Here's what you shouldn't ignore. WIRED." [Online]. Available: https://www.wired.com/insights/2013/02/attention-big-data-enthusiasts-heres-what-you-shou
[Accessed: 15-Nov-2016]

[43] D. Laney, "3D data management: Controlling data volume, velocity, and variety," META Group, February 2001 [Online]. Available: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-

pdf

[44] M. Hilbert, "Big data for development: A review of promises and challenges," *Development Policy Review*, vol. 34, no. 1, pp. 135–174, December 2015 [Online]. Available: http://dx.doi.org/10.1111/dpr.12142

[45] N. Davis Kho, "The state of big data," 24-Feb-2016. [Online]. Available: http://www.econtentmag.com/Articles/Editorial/Feature/The-State-of-Big-Data-108666.htm. [Accessed: 18-Nov-2016]

[46] T. C. (Apple's CEO), "A message to our customers. Customer letter," 16-Feb-2016. [Online]. Available: http://www.apple.com/customer-letter/. [Accessed: 18-Nov-2016]

[47] M. Green, "What is differential privacy? A few thoughts on cryptographic engineering," 15-Jun-2016. [Online]. Available: https://blog.cryptographyengineering.com/2016/06/15/what-is-differential-privacy/. [Accessed: 18-Nov-2016]

[48] B. Budington, "WhatsApp rolls out end-to-end encryption to its over one billion users," 07-Apr-2016. [Online]. Available: https://www.eff.org/deeplinks/2016/04/whatsapp-rolls-out-end-end-encryption-its-1bn-users. [Accessed: 18-Nov-2016]

[49] "Stuck in traffic? Insights from googlers into our products, technology, and the google culture," 28-Feb-2007. [Online]. Available: https://googleblog.blogspot.com/2007/02/stuck-in-traffic.html. [Accessed: 18-Nov-2016]

[50] Wikipedia, "Google traffic," 25-Oct-2016. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Google_Traffic&oldid=746200591.

[Accessed: 18-Nov-2016]

[51] "Global mobile OS market share." [Online]. Available: https://www.statista.com/statistics/266136/global-market-share-held-by-smartphone-operating-systems/. [Accessed: 18-Nov-2016]

[52] J. Ao, P. Zhang, and Y. Cao, "Estimating the Locations of Emergency Events from Twitter Streams," *Procedia Computer Science*, vol. 31, pp. 731–739, 2014 [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1877050914004980. [Accessed: 05-Nov-2016]

[53] G. Palem, "The Practice of Predictive Analytics in Healthcare," *ResearchGate*, Apr. 2013 [Online]. Available: https://www.researchgate.net/publication/236336250_The_Practice_of_Predictive_Analytics_in_Healthcare. [Accessed: 05-Nov-2016]

[54] N. Burger, B. Ghosh-Dastidar, A. Grant, G. Joseph, T. Ruder, O. Tchakeva, and Q. Wodon, "Data Collection for the Study on Climate Change and Migration in the MENA Region," 2014 [Online]. Available: https://mpra.ub.uni-muenchen.de/56929/. [Accessed: 04-Nov-2016]

[55] M. Gaitho, "Applications of big data in 10 industry verticals," 20-Oct-2015. [Online]. Available: https://www.simplilearn.com/big-data-applications-in-industries-article. [Accessed: 19-Nov-2016]

[56] F. T. C. USA, "Personal data ecosystem," *Protecting Consumer Privacy in an Era of Rapid Change - Recommendations for Business and Policymakers - FTC Report*, March-2012. [Online]. Available: https://www.ftc.gov/sites/default/files/documents/public_events/exploring-privacy-roundtable-series/personaldataecosystem.pdf. [Accessed:

17-Nov-2016]

[57] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, p. 4, Apr. 1965 [Online]. Available: https://drive.google.com/file/d/0By83v5TWkGjvQkpBcXJKT1I1TTA/. [Accessed: 07-Dec-2016]

[58] T. Pritlove and U. Schöneberg, *Neuronale netze*. 2015 [Online]. Available: https://cre.fm/cre208-neuronale-netze. [Accessed: 06-Dec-2016]

[59] L. Columbus, "51% of enterprises intend to invest more in big data," 22-May-2016. [Online]. Available: http://www.forbes.com/sites/louiscolumbus/2016/05/22/51-of-enterprises-intend-to-invest-more-in-big-data/. [Accessed: 07-Dec-2016]

[60] "ProjectVRM - cDevelopment work. ProjectVRM," 28-Nov-2016. [Online]. Available: https://cyber.harvard.edu/projectvrm/VRM_Development_Work. [Accessed: 09-Dec-2016]

[61] "ProjectVRM - principles. ProjectVRM," 28-Nov-2016. [Online]. Available: https://cyber.harvard.edu/projectvrm/Main_Page#VRM_Principles. [Accessed: 09-Dec-2016]

[62] The TAS3 Consortium, "TAS3 architecture - figure 2.2: Major components of organization domain." Jul. 2011 [Online]. Available: http://homes.esat.kuleuven.ac.be/~decockd/tas3/final.deliverables/pm42/TAS3_D02p1_TAS3.Architecture_final.pdf

[63] "Kantara initiative – join. innovate. trust." [Online]. Available: https:

//kantarainitiative.org/. [Accessed: 14-Dec-2016]

[64] T. Kirkham, S. Winfield, S. Ravet, and S. Kellomaki, "The personal data store approach to personal data security," *IEEE Security & Privacy*, vol. 11, no. 5, pp. 12–19, 2013.

[65] Y.-A. de Montjoye, S. S. Wang, A. Pentland, D. T. T. Anh, A. Datta, and others, "On the trusted use of large-scale personal data." *IEEE Data Eng. Bull.*, vol. 35, no. 4, pp. 5–8, 2012 [Online]. Available: http://sites.computer.org/debull/a12dec/a12dec-cd.pdf#page=7. [Accessed: 30-Oct-2016]

[66] "openPDS/SafeAnswers - the privacy-preserving personal data store." [Online]. Available: http://openpds.media.mit.edu/. [Accessed: 14-Dec-2016]

[67] Y.-A. de Montjoye, E. Shmueli, S. S. Wang, and A. S. Pentland, "openPDS: Protecting the privacy of metadata through SafeAnswers," *PLoS ONE*, vol. 9, no. 7, p. e98790, Jul. 2014 [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0098790. [Accessed: 30-Oct-2016]

[68] "Microsoft HealthVault. Overview." [Online]. Available: https://www.healthvault.com/de/en/overview. [Accessed: 14-Dec-2016]

[69] "How it works meeco." [Online]. Available: https://meeco.me/how-it-works.html. [Accessed: 14-Dec-2016]

[70] M. Page, "Online adver>sing – booming or broken?" Sep-2015 [Online]. Available: https://meeco.me/assets/pdf/Meeco_Case_Study_Online_

Advertising-Booming_or_Broken_Sept_2015.pdf

[71] "The principles. Industrial data space e.V." [Online]. Available: http://www.industrialdataspace.org/en/the-principles/. [Accessed: 14-Dec-2016]

[72] B. Prof. Dr.-Ing. Otto, S. Prof. Dr. Auer, J. Cirullies, J. Prof. Dr. Jürjens, N. Menz, J. Schon, and S. Dr. Wenzel, "Industrial data space - digital sovereignity over data." Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V., 17-Aug-2016 [Online]. Available: http://www.industrialdataspace.org/wp-content/uploads/2016/09/whitepaper-industrial-data-space-eng.pdf