Operational Framework: Institutional Controls

Daniel "Dazza" Greenwood[1,*], Arek Stopczynski[1,2], Brian Sweatt[1], Thomas Hardjono[1], Alex Sandy Pentland[1]
1 MIT
2 DTU
∗ **E-mail: dazza@civics.com**

Contents

1 Introduction and Overview

To realize the promise and prospects of a big data society and avoid its security and confidentiality perils, institutions are updating operational frameworks governing business, legal and technical dimensions of their organization and it's interactions with the outside world. The control points traditionally relied upon as part of corporate governance, management oversight, legal compliance and enterprise architecture must evolve to match operational frameworks for bid data. The operational framework used for a big data driven organization requires a balanced set of institutional controls are needed to achieve operational frameworks of this nature. These institutional controls must support and reflect greater user control over personal data and also large scale interoperability for data sharing between and among institutions. Core capabilities of these controls include responsive rules-based systems governance and fine-grained authorizations for distributed rights management.

2 The New Realities of Living in a Big Data Society

Sustaining a healthy, safe and efficient society is a scientific and engineering challenge that goes back to the 1800s, when the Industrial Revolution spurred rapid urban growth and created huge social and environmental problems. The remedy then was to build centralized networks that delivered clean water and safe food, enabled commerce, removed waste, provided energy, facilitated transportation and offered access to centralized healthcare, police and educational services.

But these century-old solutions are increasingly obsolete. We have cities jammed with traffic, worldwide outbreaks of disease that are seemingly unstoppable and political institutions that are deadlocked and unable to act. In addition, we face the challenges of global warming, uncertain energy, water, and food supplies, and a rising population that will require building one thousand new cities of a million people each in order just to stay even.

But it does not have to be this way. We can have cities that are protected from pandemics, that are energy efficient, have secure food and water supplies, and have much better government. To reach these goals, however, we need to radically rethink our approach. Rather than static, fixed systems that are separated by function-water, food, waste, transport, education, energy, and so on-we must consider them as dynamic, data driven systems. Instead of focusing only on access and distribution, we need dynamic, networked, self-regulating systems that are driven by the needs and preferences of the citizens.

To ensure a sustainable future society, we must use our new technologies to create a ‘nervous system that maintains the stability of government, energy, and public health systems around the globe. Currently, our digital feedback technologies are capable of creating a level of dynamic responsiveness that our larger, more complicated modern society requires. We must reinvent societies systems within a control framework: sensing the situation, then combining these observations with models of demand and dynamic reaction, and finally using the resulting predictions to tune the system to match the demands being made of it.

The engine that will drive this new nervous system is big data: the newly ubiquitous digital data now available about all aspects of human life. By analyzing patterns of human experience and idea exchange within the ‘digital breadcrumbs that we all leave behind us as we move through the world-call records, credit card transactions, and GPS location fixes, among others. These data tell the story of your life by recording what you have chosen to do. And this is very different than what you put on Facebook; your postings on Facebook are what you choose to tell people, edited according to the standards of the day.

Who you actually are is more accurately determined by where you spend your time and which things you buy not just what you say that you do.

The process of analyzing the patterns within these digital breadcrumbs is called reality mining, and through reality mining we can tell an enormous amount about who you are. The Human Dynamics research group at MIT have found that we can use them to tell if you are likely to get diabetes, or whether you are the sort of person who will pay back loans. And by analyzing these patterns across many people, we are discovering that we can begin to explain many things-crashes, revolutions, bubbles-that previously appeared to be random acts of God. For this reason the magazine Technology Review named our development of reality mining as one of the ten technologies that will change the world.

3 The New Deal on Data

The digital breadcrumbs we leave behind provide clues about who we are and what we want. That makes these personal data immensely valuable, both for public goods and for private companies. As European Consumer Commissioner Meglena Kuneva said recently, Personal data is the new oil of the internet and the new currency of the digital world. This new ability to see the details of every interaction, however, can be used for good or for ill. Therefore maintaining protection of personal privacy and freedom is critical to our future success as a society.

A successful data driven society must be able to guarantee that our data will not be abused; and perhaps especially that government will not abuse the power conferred by access to such fine-grain data. To achieve the positive possibilities of a data driven society, we require what I have called the New Deal on Data workable guarantees that the data needed for public goods are readily available while at the same time protecting the citizenry. We must develop much more powerful and sophisticated tools for privacy and reach a consensus that allows us to use personal data to both build a better society and to protect the rights of the average citizen.

A key insight that motivates the creation of a New Deal on Data is that your data are worth more when shared, because these aggregated data inform improvements in systems such as public health, transportation, and government. For instance, we have demonstrated that data about the way you behave and where you go can be used to minimize the spread of infectious disease. Our research has reported how we were able to use these digital breadcrumbs to track the spread of influenza from person to person on an individual level. And if you can see it, you can stop it. In this instance, the result of sharing your personal data is that we can build a world where the threat of infectious pandemics is greatly diminished.

Similarly, if you're worried about global warming, these shared, aggregated data now show us how patterns of mobility relate to productivity. In turn, this provides us with the ability to design cities that are both more productive and at the same time more energy efficient. But in order to be able to obtain these results and make a greener world, you need to be able to see the people moving around; this depends on many people being willing to contribute their data, even if only anonymously and in aggregate.

While concrete examples such as better health systems and more energy efficient transportation systems motivate a New Deal on Data, there is an even greater public good that can be achieved by efficient and safe data sharing. To enable sharing of personal data and experiences, we also need secure technology and regulation that allow individuals to safely and conveniently share personal information with each other, with corporations, and with government. Consequently, the heart of a New Deal on Data must be to provide both regulatory standards and financial incentives that entice owners to share data while at the same time serving the interests of both individuals and society at large. We must promote greater idea flow among individuals, not just corporations or government departments.

Unfortunately, today most personal data are siloed off in private companies and therefore largely unavailable. Private organizations collect the vast majority of personal data in the form of location patterns, financial transactions, phone and internet communications, and so on. These data must not remain the exclusive domain of private companies, because then the data are less likely to contribute to the common good. Thus, these private organizations must be key players in the New Deal on Datas framework for privacy and data control. Likewise, these data should not become the exclusive domain of the government, because this will not serve the public interest of transparency and we should be suspicious of trusting the government with such power.

4 Personal Data: Emergence of a New Asset Class

It has long been recognized that the first step to promoting liquidity in land and commodity markets is to guarantee ownership rights so that people can safely buy and sell. Similarly, the first step toward creating greater idea and idea flow ('idea liquidity) is to define ownership rights. The only politically viable course is to give individual citizens rights over data that are about them and in fact, in the European Union these rights flow directly from the constitution. We need to recognize personal data as a valuable asset of the individual that is given to companies and government in return for services.

The simplest approach to defining what it means to own your own data is to draw an analogy with the English common law ownership rights of possession, use, and disposal:

You have the right to possess data about you. Regardless of what entity collects the data, the data belong to you, and you can access your data at any time. Data collectors thus play a role akin to a bank, managing the data on behalf of their customers.

You have the right to full control over the use of your data. The terms of use must be opt-in and clearly explained in plain language. If you are not happy with the way a company uses your data, you can remove the data, just as you would close your account with a bank that is not providing satisfactory service.

You have the right to dispose of or distribute your data. You have the option to have data about you destroyed or redeployed elsewhere.

Individual rights to personal data must be balanced with the need of corporations and governments to use certain data-account activity, billing information, and so on-to run their day-to-day operations. This New Deal on Data therefore gives individuals the right to possess, control, and dispose of copies of these required operational data, along with copies of the incidental data collected about you such as location and similar context.

Note that these ownership rights are not exactly the same as literal ownership under modern law, but the practical effect is that disputes are resolved in a different, simpler manner than would be the case for (as an example) land ownership disputes.

In 2007, one author (Pentland) first proposed the New Deal on Data to the World Economic Forum. Since then, this idea has run through various discussions and eventually helped shape the 2012 Consumer Data Bill of Rights in the United States, along with a matching declaration on Personal Data Rights in the EU. These new regulations hope to accomplish the combined trick of breaking data out of the current silos, thus enabling public goods, while at the same time giving individuals greater control over data about them. But, of course this is still a work in progress and the battle for individual control of personal data rages onward.

5 Enforcing the New Deal on Data

How can we enforce this New Deal? The threat of legal action alone is important, but insufficient, because if you cannot see abuses then you cannot prosecute them. Moreover, who wants more lawsuits anyway? Enforcement can be addressed in significant ways without prosecution of public statute or regulation at all. In many fields, companies and governments rely upon multi-party frameworks of agreed rules governing common business, legal and technical practices to create effective self-organization and enforcement. These approaches hold promise as a method for using institutional controls to form a reliable operational framework balancing the needs for big data, privacy and access.

One current best practice is a system of data sharing called trust networks. Trust networks are a combination of networked computers and legal rules defining and governing expectations regarding data. With respect to data belonging to individuals, these networks of technical and legal rules keeps track of user permissions for each piece of personal data, and a legal contract that specifies both what you can and cannot do with the data and what happens if there is a violation of the permissions. For example, in such a system all personal data can have attached labels specifying what the data can, and cannot, be used for. These labels are exactly matched by the network's system rules and terms in legal contracts between all the participants stating penalties for not obeying the permission labels. These rules can, and often do, reference or require audits of relevant systems and data use, demonstrating how traditional internal controls can be leveraged as part of the transition to more novel trust models.

Complete tracking and regulation of every aspect of a trust network is not the goal or even desirable in order to achieve effective enforcement. Rather, the rules for a trust network align enforcement with the highest priority issues and those upon which trust of participants is premised. The relevant issues arise from the dynamics of data flows, underlying trust models and contextual scenarios within which the networked data and the relationships of parties in the trust network. When a trust network involves use of personal data, then the user permissions and corresponding limits on use are fundamental to the trust model. In this context, the permissions, including the provenance of the data, should require appropriate levels of audit. A well designed trust network, elegantly integrating computer and legal rules, allows automatic auditing of data use and allows individuals to change their permissions and withdraw data.

Having system rules applicable to the networks, applications and data as well as all the services providers other intermediaries, and the users themselves is the mechanism for establishing and operating a trust network. System rules are sometimes called operating regulations in the credit card context, or known as trust frameworks in the identity federations context, or trading parter agreements in a supply value chain context. There are many general examples of multiparty shared architectural and contractual rules that share the generic characteristic of creating binding obligations and enforceable expectations on all participants in scalable networks. Another common characteristic of the system rules design pattern is that the participants in the network can be widely distributed across very heterogeneous business ownership boundaries, legal governance structures and technical security domains. Yet, the parties need not agree to conform all or most aspects of their basic roles, relationships and activities in order to connect to to systems of a trust network. Cross-domain trusted systems must, by their nature, focus mandatory and enforceable rules narrowly upon the critical items that must be commonly agreed in order for that network to achieve it's purpose.

For example, institutions participating in credit card and automated clearinghouse debit transactional networks are subject to profoundly different sets of regulations, business practices, economic conditions and social expectations. The network rules focus upon the topmost agreed items affecting interoperability, reciprocity, risk and revenue allocation. The knowledge that fundamental rules are subject to enforcement actions is one of the foundations of trust as well as a motivation to prevent or address violations before they trigger penalties. A clear example of this approach can be found with the Visa Operating Rules, covering a vast global real-time network of parties that agree to rules governing their roles in the system as merchants, banks, transaction processors, individual or business card holders and other key system roles.

A system like this has made the interbank money transfer system among the safest systems in the world and the daily backbone for exchanges of trillions of dollars, but until recently such systems were only for the ‘big guys. To give individuals a similarly safe method of managing personal data, the Human Dynamics research group here at MIT, in partnership with the Institute for Data Driven Design, co-founded by John Clippinger and one author (Pentland), have helped build openPDS (open Personal Data Store) [fn: See http://openPDS.media.mit.edu for project information and https://github.com/HumanDynamics/openPDS for the open source code]. The openPDS system is a consumer version of a personal cloud trust network and we are now testing it with a variety of industry and government partners. Soon, sharing your personal data could become as safe and secure as transferring money between banks.

[FN: The Human Dynamics Lab has applied the system rules approach to development of integrated business, technical architecture and rules large scale institutional use of personal data stores, available as an example under MIT's creative commons license by MIT, at: github.com/HumanDynamics/SystemRules and the Institute for Data Driven Design has published a guiding vision of digital common law, describing how these concepts can form the basis of next generation legal systems, available at: idcubed.org/??]

The capacity to apply the appropriate methods of enforcement for a trust network depend upon a clear understanding and agreement among parties about the purpose of the trusted system and the respective roles or expectations of those connecting is as participants. Therefor, an anchor is needed to a clear context of a big data operational framework and institutional controls appropriate for access and confidentiality or privacy. The following section posits the trust model and signature traits of such a context, through the lens of the New Deal on Data.

6 Essential Elements of the New Deal of Data

To realize the promise and prospects of Big Data, and to avoid the associated privacy perils, we need a balanced set of institutional controls. Theses controls must support and reflect a greater user control over personal data, as well as large scale interoperability for data sharing between and among institutions.

The core capabilities of theses controls should include responsive rule-based systems governance and fine grained authorizations for distributed rights management.

Our lives are embedded within institutions. We are citizens of countries and cities, receive services from telecom operators, and search for things to buy in online stores. Almost any action we perform generates data, and those recordings of our lives are an important part of the Big Data promise. The data are not curated by us, but are collected 'as is' - and reflect our lives.

Today, all the data people generate in the context of institutions are stored in closed silos. Mobility traces, for example, are owned by the phone providers, while musical preferences are stored and used by music services.

For these data to be useful to society, the silos must be opened, and the data must be used far more often than they are today. If access to data for the purpose of creating value–either for the user or the society–is very limited, it does not matter how big the data is. The value of the data lies not just in the fact that they exist. Rather, it is the knowledge, understanding, and wisdom we gain from them that makes the data valuable. It is an even bigger challenge to open up the data from multiple silos. Accessing the multi-faced data, which exist under multiple jurisdictions, about people may be prohibitively difficult. Silos are hard to crack open. Such data, not just Big but Deep, covering multiple facets of a person's life, may be invaluable for research.

Recently, we have shown how challenging, but also possible, it is to open such institutional Big Data. In the Data For Development (D4D) Challenge [1], the telecom operator Orange opened access to a large dataset of call detail records (CDRs) from the Ivory Coast. Working with the data as part of a challenge, teams of researchers came up with life-changing insights for the country. The privacy of the people was protected not only by the technical means, such as removal of the Personally Identifiable Information (PIIs), but also by legal means, with the researchers signing an agreement they will not use the data for evil. As we have seen in several cases, such as the Netflix Prize privacy disaster [**?**] and other similar privacy breaches [**?**], true anonymization is extremely hard. Some of the weight of privacy protection must rest on the legal framework.

[1] http://www.d4d.orange.com/home

Opening data from the silos by publishing static datasets is important, but it is only the first step. We can do even more important things when the data is available in real time and can become part of a nervous system of a society. Epidemics and traffic congestions can be monitored and prevented in real time, underpferoming students can be helped, and people with health risks can be treated before they get sick. The same data can be used for stalking, burglarizing one's home, and as a reason to charge people more for an insurance policy.

In the Unique in the Crowd project [**?**], we have shown that even though human beings are highly predictable [**?**], we are also very unique. Having access to one dataset, it is easy to uniquely fingerprint someone based on just few datapoints, and use this fingerprint to discover their true identity. The higher the resolution of the data, the better the data, the easier it gets.

The question of privacy in this context effectively becomes a question of control:

Who can release the data of one's movements? To whom? How much and how often? The data are collected by the institution. The data are about people and do not belong to them, they may not even be aware that they exist. People cannot decide upon them, cannot review them. People cannot delete them. Very few parties can use the data, even if people wanted them to. For systems to be truly data driven and capable of transitioning to the networked and highly dynamic assumptions of a big data economy, the key agreements reflected in trust networks must reflect a new deal. The operating frameworks successful institutions are capable of balancing interests in access, confidentiality and every day reliance upon big data including personal and other sensitive information. The institutional controls relevant to achieve, maintain and appropriately adapt these balances support and reflect adherence to the fair information practices.

[Footnote: HEW Report, OECD rendition, EU Directive, DHS/NSTIC version, MGL FIPA and culminating in New Deal on Data adaptation].

Within the existing legal frameworks, it is possible to change the vantage point of the data ownership and put the user, the entity about whom the data are, in control. It may be a copy of the data living in the great silo, which is being given to the user. The user would become the owner of their copy of the data, or whenever possible the original, in the old Common Law sense with the right to use, transfer, and delete the data. An example of such a mechanism in an institutional context is Blue Button initiative [2], where the patients can get a copy of their health records. Once the copy is with the user, they can do with it as they wish: give it to someone, make it public, do research on it, destroy it.

The users can accumulate data about themselves from multiple sources. Information on healthcare records, mobility patterns, favorite movies, etc., all belong to the user and can be accessed based on their authorization. This changes how and what data that can be obtained for the purpose of research and providing services. Rather than gaining access to the movements of millions of people from a telcom operator, one can potentially gain access to a smaller number but of much richer datasets describing the users from the mobility, health, shopping perspectives. New startups do not have to build the user profile from scratch, but can jump in offering competitive services based on the user's collected data. Users can immediately get better services, using their data in new places.

The first, operational challenge of moving towards the end-user data ownership on a large scale, is to create an ecosystem where such user-owned data are noticed and accessed. We are currently embedded in a feudal framework: Facebook owns the data generated by and about their users, and provides the access to them to the 3rd parties that user might or might have not authorized. It is reasonably easy for users to download all their data from Facebook. It is reasonably easy to put it on Dropbox or even create myself-API, becoming a self-hosted API to one's own personal data. The challenge is to have clients to talk to this API and provide services, rather than going to Facebook for one's data. Today, virtually no-one is ready to access user data directly from the user. We have done a slightly better on the Internet scale with identity: one can deploy own OpenID server fairly easily, and many services will allow the user to sign in. We should be heading in the same direction with data.

[2] http://www.healthit.gov/bluebutton

7 Transitioning End-User Assent Practices

The way the user grants authorizations to the data she owns is not a trivial matter. Just answering the very simple question 'Who is authorized to know what city I am in today' may be a challenge. The 'Yes' the user has clicked so many times has given access to the location data to so many services. Every tweet, every geo-tagged picture, and every check-in provides the user's location not only to the primary service, but also to all the applications that have been authorized to access these data. This flow of information was a crucial part of the Web2.0 revolution, with RESTful APIs, mashups, and authorizations. The complexity of the flow became too large for a user to handle and manage.

Increasing the amount of data the user controls and increasing the granularity of the control is meaningless if this control cannot be exercised in an informed way. The EULA-catastrophe, where the users may be just as well giving up their soul when signing up without reading, will not bring us closer to the New Deal on Data. In the end it must the be user that makes the informed decision about who will be authorized to access the data and for what purpose. Making the authorization interface too complex is a failure, preventing the user from understanding her decisions. Making it too simple, is also a failure, as it will not convey the complexity of the privacy-related decisions. Writing it in complex legal language makes it very hard to claim that the user expresses informed consent. And if we start asking the users for authorization every 5 minutes, it will only train them to press 'Yes' every time a pop-up is presented.

In addition to the data ownership, we need a better way for the end-user to control what happens with, now their, data. Will users realize that clicking a single 'Yes' gives a service a second-resolution location data? And what can be inferred from such data, regarding alcohol abuse ('we see you a lot in a liquor store'), driving habits, not enough exercise.

This gap between the interface, the single click, and the effect, can render the data ownership meaningless if that click wrenches people and their data into systems and rules that are antithetical to fair information practices, such as is prevalent with todays end user licenses, cloud service and app user agreements. Managing the potentially long term and opposite dynamics fueled by old deal systems operating simultaneously with new deal systems is an important design and migration challenge during the transition to a big data economy Ironically, some approaches to offering personalized data access by individual approval in secure context, would take the power over the data back out of the hands of the approving users. The sames types of system rules that can and should reflect fair information practices in order to build a sustainable big data operational framework can also be abused to expropriate user's personal data. The so-called walled garden comprising the interior of a trust network can protect personal data within it's context as a big data sanctuary or it can keep data owners and others with legitimate interests in data away from knowledge or consent for ever after. Walled-gardens comprised of networks of openPDS friendly operating environments can enable a new data economy or can be used as secret dens to hide and sell personal data. The cost of permitting backward looking industrial era practices better suited to keeping a secret formula in a safe are high. It's time to look forward, and embracing the relevant institutional controls is one actionable method gain deeper benefits of truly networked data today and accelerate emergence of a transformatively better era of big data.

There are several potential paths for institutional controls to eventually work in concert across many systems and enterprises and interoperating networks until a stable general environment for big data can be relied upon. For instance, one or a few regulated sectors of the economy could organize and in effect "get ahead" of prescriptive privacy and other legislation by self-regulating, in effect, by way of contractually based system rules. Or bottom up networks of apps, web services and perhaps diagonal chains of transactions, such as permission based automatic forms filling, could emerge and take hold as a beach head, eventually merging with others. Consumer, privacy or state/local policy maker groups could catalyze a national wave of adoption covering a horizontal type of data (perhaps location or identity data) or specific types of transactions could be first to cross the big data chasm and become broadly adopted safe-zones eventually merging with others and comprising most and then all of the data systems. However, when hubs of centralized systems operating outside a New Deal on Data rather than decentralized systems such as openPDS and other personal cloud services exist simultaneously and can arrogate the key rights to user's data in one fell click, there is a risk to smooth and timely transition.

Now, and during the transition to a big data economy and after the new deal on data is no longer new, personal data must continue to flow in order to be useful. Protecting the data of people outside of a domain of the user is very hard without a combination of cost effective and useful business practices, legal rules and technical solutions. For these reasons, the Human Dynamics Lab has focused upon and collaborated with partners to support the clarification of business, legal and technical short and longer term viable solutions. The World Economic Forum's thought leadership is an excellent example of how all levels of the economy are shining light upon the path to adoption of a New Deal on Data as way of avoiding uncertainty, catalyzing further innovation and a path to dynamic prosperous marketplaces fueled by big data. The Identity Ecosystem Steering Committee has likewise been a beacon on the policy front, demonstrating how fair information practices can form the bedrock of privately adopted operating rules and trust frameworks for identity sharing across all sectors of the economy, all types of business and all individuals of the population.

[Footnote: WEF Relevant papers and blog posts, etc. NSTIC strategy document and IDESG RoA. Not sure best example for the various technical approaches being pursued... perhaps W3C work, or Kerberos/OIDC or something else that is hand-in-glove with openPDS approach?]

8 The Role of Meaningful Individual Consent

Informed consent is a much better than consumer website and app click-wrap as a starting place for the types of practices for institutions to embrace and ensure during the transition period to a big data economy. The basic legal rule governing contractual agreement for consumers operating in commercial environments online today simply requires that the user click something or take an affirmative actions and be on notice of and have an opportunity to review the terms of contracts. There is no general rule requiring users actually be informed of the terms of contracts they are agreeing to (and for the most part will be subject to) when or before they click to install an app or use a web-based service. By contrast, informed consent does require that a person be informed of the key terms and that they actually understand them before their agreement can be valid. This can be an electronic process and the Human Dynamics Lab has several working examples that are currently in use by partner institutions and users of operPDS and related systems. These methods form an important part of how people can and will play a meaningful role in the transition to a big data economy and how institutions can leverage well established practices to ensure agreements regarding use of personal and other sensitive data rests upon firm and enduring foundations.

[Footnote to GitHub Repo and DTU current practices and legal blog posts on innovative IRB practices]

There is a need for Living Informed Consent, where the interface for the user to grant the authorizations is made to give the user understanding of the consequences, benefits and dangers of the granted authorizations. This understanding will never be perfect, but aligning what user understands about their decisions with the reality is the goal of the Living Informed Consent concept.

We envision several ways the Living Informed Consent can improve user's understanding of the authorizations she is granting. The underlying principle is that the status of the authorizations expressed via the interface (website, application) is the contract. By pressing the buttons, the user initiates technical actions (for example creation of OAuth2 tokens), but also changes her business and legal relation with the service. Such single screen, with a time stamped log constitutes a history of the consent. The granularity of offered control may differ, and some actions may or may not be permitted within give institution, agreement, or service. Still, at any point in time, the user is in certain relation with the service, in the Business, Legal, and Technical domains. The consent only makes sense when the user understands what she is consenting to. Why even bother asking otherwise.

Part of the gestalt is to provide concise description of the authorizations written in plain English. The expression of those authorizations may not always be trivial and may sometime turn into paragraphs of text, yet still the goal should be to provide a description comprehensible for the target audience. Additionally, the goal should be not only to ask for the access to data, but also include the purpose of the access. Location is a type of data. Using location to provide personalized music and using location to increase my insurance for careless driving are two very different authorizations. Current authorizations frameworks do not really handle on the purpose part, focusing mainly on the data being access, no the reason for access. This is suboptimal from the user perspective.

Where the authorizations are granted, one possible way to make it easier for the user to understand what is happening with the data, is to reduce the dimensionality of the data already in the user-controlled domain, and only send high-level answers to the service requesting them.

A lot can be inferred from a raw location trace. The moment the raw data leaves the user-controlled domain, it can be used for many things, some of them the user may have never thought about, and could not possibly have expressed informed consent.

Extracting the high-level features of the data on the user side, as described in the openPDS framework, should allow for more informed decisions regarding the data access. All the raw data should not run wildly with every service providing a minuscule service to the user.

It is much easier to control what can happen and thus what are the consequences of disclosing the city you live in versus all your location updates from the last year. It is not a perfect solution; even low-dimensionality data can still be used for evil and can be correlated with other sources. It is however a big step in the right direction, for the user to decide upon disclosing how much liqueur she buys per week versus this information being inferred from the GPS trace provided to a service in exchange for personalized music. When the control over the data access is given to the user together with the tools to reduce the dimensionality, it becomes less important what data are being collected, only hat is disclosed to the external services. The user's personal data store becomes a 'master copy' of the data, containing all the possible data points. The user then controls what answers regarding these data are shown to the world.

In addition, the information about data access and usage needs to be an integral part of Living Informed Consent. How often do services sample user's location? Are they tracking her in real-time, or do they access the data on a weekly basis? Is the user singled out in how much data is queried, or is it the same for all the users? For the user, being able to answer those questions in a simple, even casual way, is crucial for remaining in the state of Living Informed Consent. Authorizations should not be of 'fire and forget' kind, instead they should be re-evaluated in some orderly fashion. How often this should happen depends on multiple factors, including the sensitivity of the data accessed, reputation of the service, user preferences, the balance between control and annoyance.

Giving the data ownership to the end-user makes it easier for the institutions to facilitate the data use. As the users are the fully-empowered parties to make decisions about authorizing access to all their data, multiple silos do not have to be visited and contracts between them made. It is sufficient to talk to the end-user to gain access to all the data about her. This way, the institutions can facilities the data use becoming a matchmaking service for data access rather than data handlers, simplifying their process, while still potentially benefiting from being the 'data directory'.

A crucial component for realizing this vision is the identity of the user:

It must be possible for multiple institutions to find the user to give the data to. It must be possible for the user to identify multiple institutions and see where the data is coming from. It must be easier than remembering passwords and logging in separately to every service. All the questions we have asked so far about the data (who owns, who controls, who decides, who accesses) must have the 'who' component addressed. Just as the data of the user should live under single control of this user, the identity should also be brought closer to user control. It does not necessarily mean every user should be their own identity provider, but rather than having hundreds of accounts in multiple services that do not interoperate or know about each other, the identity of the user should be build on the principle of federated identity, where the services allow the user to choose their identity provider. In addition, just like data, certain attributes of identity need to be protected. Service does not need to know the user's email address to be able to log the user in, a pseudonym (random but consistent string) is sufficient. If such service has a valid reason for asking for the users address, it should be based on the users grant of authorization. Authorization can be revoked and monitored.

In the existing system it is often hard to introduce user data ownership. This may be due to technical reasons, such as building the infrastructure to provide the space for the data. It may be for business or legal reasons where the data is considered not suitable for sharing. It may be for the lack of a clear incentive as to why to do it, or how to interact with the users during the process of introduction and afterwards. We feel the first step in introducing more privacy into such system is the notion the user must be entitled to at least know about the existence of the data about her. The right to know about the data existence is hard to deny. It can be realized more smoothly than the transfer of the actual data. It can be the first step towards The New Deal on Data, supported by institutional controls and enforced by a legal framework.

9 Implementing the New Deal on Data with openPDS

The openPDS software is a reference technical implementation of institutional and legal controls envisioned by the New Deal on Data. It provides a repository for storing personal data, as well as a system for user-controlled access to this personal data. Such an implementation can, and indeed should, be based on contemporary Internet standards. To this end, openPDS builds on top of well-defined technologies, such as OpenID for identity management and OAuth 2.0 for resource authorization management. It extends these standards in key ways to facilitate compliance with the New Deal on Data.

9.1 Federated Identity Management

OpenID is an open standard that allows user authentication by co-operating sites via a third party service, known as an OpenID server. While openPDS acts as a personal OpenID server, it allows login with a multitude of disparate identity services. It builds on top of the OpenID standard to allow association of any number of digital identities with a proxy OpenID managed and stored on the user's personal data store. This identity can be used to log into the PDS as well as any Internet service that accepts OpenID credentials.

9.2 Personal Data Access Control

The openPDS software provides an interface for users to authorize access to their personal data for third parties. To achieve this, it builds on top of OAuth 2.0; an open standard for authorization. OAuth allows institutions to specify scopes, which act as identifiers for the type of information that can be access based upon user managed content. For example, a scope might exist that provides access to read a user's email address, or gives the third party institution access to read and write posts on a user's twitter feed. When a user authorizes the third party institution to access their data in the manner associated with the specified scopes, the OAuth server provides the third party with a token; a string of characters the third party must provide whenever accessing data associated with the approved scopes.

OpenPDS extends OAuth to provide additional dimensions for users to control access to their personal data. First, it requires a purpose to be provided in addition to a scope for all approved OAuth tokens. Purpose is a detailed explanation of why the third party desires access to the data associated with the scope, as well as the operations they intend to perform on this data. Second, openPDS provides a rule engine for associating tokens with more fine-grained access control mechanisms, such as geo-fencing. A user may utilize this engine to specify a number of additional conditions that must be met in order for the access to be authorized. For example, these two extensions can be combined to allow controlled access to a user's location data (specified by a scope), for the purpose of identifying whether the user attended a meeting, only while the user is present at their office during business hours.

In order to enforce access to data only for the intended purpose, a compute engine resides within openPDS and is a trusted space for operations to run against raw personal data. In this manner, to get at the data associated with a given scope and purpose, an institution must submit the code detailing the operations to perform on the user's raw data. This code acts as a question from the third party to the user's PDS and runs within the compute engine to construct a privacy-preserving answer based from the raw data contained within the PDS. By disallowing access to the raw data by third parties and enforcing only answers to approved questions about this raw data, openPDS provides a strong data ownership and access control mechanism consistent with the New Deal on Data.

9.3 Systems and Rules

Data literally exists within systems. Specifically, business systems, legal systems and technical systems operating in harmony comprise the type of dynamic, data-driven operational framework enabling institutions to enter and thrive in a big data ecologies. Rule-based flow and use of personal data and big data generally presume interoperability throughout, between and among institutions. Interoperability requires technical protocols, specifications and common configurations of applied standards across many types of domain boundaries. Yet no amount of technical interoperability will result in business interoperability-rather, it is more fundamentally other way around.

10 Business, Legal and Technical Dimensions of Big Data Systems

When it comes to data intended to be accessible over networks-whether big, personal or otherwise-the traditional container of an institution makes less and less sense. Institutional controls apply, by definition by or to some type of institutional entity such as a business, governmental or religious organization. A combined view of the business, legal and technical facts and circumstances surrounding big data is necessary to know what access, confidentiality and other expectations exist. The relevant contextual aspects of big data of one institutional is often profoundly different from that of another. As more and more organizations use and rely upon big data, a single formula for institutional controls will not work for increasingly heterogeneous business, legal and technical environments in play.

Looking at an institution as a business, legal and technical system is one effective approach for dealing with the inherent complexity of managing heterogeneous and distributed networks of actors and interactions. The business models, interface-point operational practices and relevant assumptions must be consistent and frequently carefully agreed at an executive level by and with institutions as part of the value exchange involving data and access to high value, mission critical or sensitive systems and services. The applicable legal frameworks, common assumptions regarding likely allocation of liability and resolution of disputes in the event of losses and expected types of contracting practices need to reflect and support the business goals and purposes for the system and data. When technical standards are selected, configured and applied to systems they too must support and reflect the business and legal dimensions and be supported and reflected by those dimensions.

Once a systems view is adopted, there is a tractable starting point to narrow or broaden the scope of view to see the smaller and larger systems and to make better and more effective use and control of big data. Within a given institution, there may in fact be many different discernable institutions and corresponding systems and any given system of one institution will frequently in fact exist across many different discernable institutions. However, defining as a system the thing to which institutional controls apply provides an achievable and measurable basis for balancing privacy, access and other interests in big data.

11 Big Data and Personal Data Institutional Controls

The phrase "institutional controls" refers to safeguards and protections by use of legal, policy, governance and other non-strictly technical, engineering or mechanical measures. The phrase institutional controls in a big data context can perhaps best be understand by examining how the concept has been applied to other domains. The most prevalent use of institutional controls, per se, has been in the field of environmental regulatory frameworks.

A good example of how this concept supports and reflects the goals and objectives of environmental regulation can be found in the policy documents of the EPA. This following definition is instructive, and is part of the Institutional Control Glossary of Terms, available at: http://www.epa.gov/epawaste/hazard/correctiveaction/re

"Institutional Controls - Non-engineering measures intended to affect human activities in such a way as to prevent or reduce exposure to hazardous substances. They are almost always used in conjunction with, or as a supplement to, other measures such as waste treatment or containment. There are four categories of institutional controls: governmental controls; proprietary controls; enforcement tools; and informational devices."

Going deeper, the article Restrictive Covenants as Institutional Controls for Remediated Sites: Worth the Effort? by Ralph A. DeMeo and Sarah Meyer Doar, available at http://www.floridabar.org/divcom/jn/jnjournal01.nsf/ defines institutional controls thusly:

"Institutional controls are administrative and legal controls that help minimize the potential for human exposure to contamination and/or protect the integrity of the physical remedy. They can include recorded restrictive covenants, but land use laws and regulations, deed restrictions, department consent orders, and conservation easements are all institutional controls."

In domains of information technology, this approach is most commonly reflected as enterprise controls" related to security. See, for example, Secure Data Access Anywhere And Anytime, Current Landscape And Future Outlook Of Enterprise Mobile Security. available at: http://www.juniper.net/us/en/local/pdf/industry-reports/secure-data-access.pdf section "Enterprise Mobility Outlook Mobility Enters A Strategic Planning Stage" at 15, stating: "Enterprise mobility technologies, especially those designed to retrofit enterprise controls on top of consumer mobile devices, are rapidly evolving. This was a message we heard loud and clear in the study." This study and analysis also reveals much about the internal controls needed to accommodate mobile device use by employees. In both capacities as employee, consumer and other roles, the use of mobile devices triggers myriad legal, policy and other implications for institutional controls.

In the legal domain, this concept frequently emerges under the moniker "regulatory compliance or legal compliance anchored in legal and regulatory frameworks such as HIPAA and Sarbanes-Oxley (SOX). These statutory legal frameworks require covered organizations to established integrated sets of governance, legal, transactional, security and other internal controls to avoid violating the rules. The institutional controls are accomplished in tight integration with engineering and other measures in order to ensure compliance and to control legal and security risk. The use of institutional controls of this type are fundamental methods for achieving and maintaining the transition to a digital, networked and big data footing for any private company, government agency or other organization.

Consider again the analogy of institutional controls in the context of environmental law, and how these types of measures can be applied in the big data, privacy and access context to digital environments. Given the relatively mature and stable state of environmental regulation, there is much to be learned by examining this context of institutional controls. Environmental regulatory compliance with waste management cleanup requirements could include institutional controls restricting land use on adjacent property. In these situations, it is possible that the remediation strategy requires significant use of land outside the property boundaries of the cleanup site. In these cases, the regulators and the land owner responsible for the regulated property must find ways to ensure a common approach among multiple owners and across multiple property environments. Use of measures such as a clauses on the relevant deeds, an enforceable consent order or regulations and zoning rules are examples of more severe institutional controls that can be employed to ensure consistent and effective actions are taken across ownership and real property boundaries.

See, for example, FDEP, Division of Waste Management, Institutional Controls Procedures Guidance at 22 (Nov. 2010), available at www.dep.state.fl.us/waste/quick_topics/publications/wc/csf/icpg.pdf, ”...RMO III does contemplate contamination beyond the Property boundaries, which would require agreement by the adjacent owners to put an RC on their properties as well.”

The concept of an ”institutional control boundary” is especially clarifying and powerful when applied to the networked and digital boundaries of an institution. In the context of Florida's environmental regulation frameworks, the phrase is applied to describe the various types of combinations risk management levels related to target cleanup standards and extend beyond the area of a physical property boundary. See the Final Technical Report: Development of Cleanup Target Levels (CTLs) for Ch. 62-777, F.A.C., at 134, available at http://www.dep.state.fl.us/waste/quick_topics/publications/wc/FinalGuidanceDocumentsFlowCharts_April2
28-05).pdf ”Risk Management Options Level III, like Level II, allows concentrations above the default groundwater CTLs to remain on site. However, in some rare situations, the institutional control boundary at which default CTLs must be met can extend beyond the site property boundary.”

The EPA provides considerable information on the nature and use of institutional controls, including situations when the situational scope extends to adjacent properties owned by third parties. See, generally, EPA Hazardous Waste Corrective Action Guidance on Institutional Controls, http://www.epa.gov/epawaste/hazard/corre
Also see: Institutional Controls Bibliography: Institutional Control, Remedy Selection, and Post-Construction Completion Guidance and Policy, December 2005, http://www.epa.gov/superfund/policy/ic/guide/biblio.pdf

When institutional controls would apply to ”separately owned neighboring properties” a number of issues arise. Engagement with affected third parties, requiring the party responsible for site cleanup to use ”best efforts” to attain agreement by third parties to institute the relevant institutional controls, use of third party neutrals to resolve disagreements regarding the application with institutional controls or forcing an acquisition of the neighboring land by forcing the party responsible to purchase the property of by purchase of the property directly by the EPA. See, Institutional Controls: A Guide to Planning, Implementing, Maintaining, and Enforcing Institutional Controls at Contaminated Sites, December 2012, at pg 16, Section 4.4 ICs and Landowners, http://www.epa.gov/superfund/policy/ic/guide/Final

In the context of big data, privacy and access, institutional controls are seldom if ever the result of government regulatory frameworks such as are seen in the environmental waste management oversight by the EPA. Rather, institutions applying measures constituting institutional controls in the big data and related information technology and enterprise architecture contexts will typically employ governance safeguards, business practices, legal contracts, technical security, reporting and audit programs and a various risk management measures. Inevitably, institutional controls for big data will have to operate effectively across institutional boundaries just as environmental waste management internal controls must sometimes be applied across real property boundaries and may subject multiple different owner to enforcement actions corresponding to the applicable controls. Short of government regulation, the use of system rules as a general model are one widely understood, accepted and efficient method for defining, agreeing and enforcing institutional and other controls across business, legal and technical domains of ownership, governance and operation.

The use of system rules and integrated participation agreements by developers and end-users is a way to ensure intended operational frameworks conform to applicable institutional controls. The example of ”living consent” described below, demonstrates how institutional controls comprised of legal and definite workflow measures in concert with technical methods can result in a higher level of performance while appropriately balancing legitimate interests of various parties regarding use and access to personal data.

12 Governance of Cross-Boundary System Rules and Institutional Controls

The tight integration of the relevant governing body or bodies responsible for promulgating the system rules is essential not only to their legitimacy and suitability, but also to ensure that governance related institutional controls are adopted and applied as intended. A governance body that approves any updates or exceptions to system rules is needed and can be organized in many different ways. It is not uncommon for such governance bodies to be comprised of a representative mix of the stakeholders or types of stakeholders upon whom the system rules and agreements will be enforceable.

The ID Federation, Inc is an example of an industry-wide body forming as a trade association to govern an identity data and interoperability federation for the insurance industry This example illustrates a well tuned governance, business, legal and technical integrated approach to system rules. The ID Federation, Inc, includes a governing board that is highly representative of the various types of stakeholders and also is organized into three main rule making committees of members: the Business Rules, Legal Rules and Technical Rules committees. These committees have representation from across the industry-both insurance carriers, brokers, agents, vendors and others-but each committee is comprised of individuals that are expert in and authorized to speak for their organizations regarding the relevant business, legal or technical matters involved in developing the minimum rule set needed to have full interoperability across the industry.

In addition, to ensure true alignment vertically, horizontally and diagonally across the various dimensions and contexts of the system rules, the board of directors is the body that ultimately reviews and either requests revisions or adopts and officially publishes each version of the system rules. In order to stream line and accelerate unified changes to the system rules, the chairs of the business, legal and technical committees themselves have a Harmonization Committee which meets every few weeks or so, as needed, to discuss any changes coming from a given committee in advance of proposal, to afford a paced and collegial cadence for feedback across the business, legal and technical domains. Proposals for changes or new versions of the system rules would come from the Harmonization Committee and thereby ensure that, whether the board of directors agrees or seeks further changes, at least each proposal has been tested and circulated across the functional, industry and administrative domains before it is presented for potential final decision.

The reason for these cycles and iterations is partly in return for the trust and risk associated with any given organization agreeing to be subject to a set of rules that will apply as an umbrella across many organizations but with they do not have final say over in the same way they might with a bi-lateral contract. Best practice is also to ensure that no changes to system rules are legally effective for some buffer of time after they have been agreed, perhaps some months. This both affords time to prepare, test and deploy any needed changes and also ensures that no participant will ever be subject to system rules against their agreement and consent. This safeguard and basic system compact can be assured by structuring the system rules to specify the notice period for voluntary withdrawal of any member is shorter than the notice period preceding the effective date of a new or modified system rule. In some instances, a security or other emergency may require exception to these safeguards, which could be accomplished by super majority agreement, permitting waivers and temporary exceptions with stipulations acceptable to members and participants who can not agree to the rules and by other governance methods. To oversimplify, one can think of the governance measures upon which system rules are based as the enterprise, extended-enterprise and inter-enterprise institutional controls for big data. The interdepended business, legal and technical rules and agreements can be thought of not only as a multilateral agreement but as an Interlateral compact that allows the continued use of all existing corporate, jurisdictional, departmental and other boundaries while also enabling deep integration of cross-functional, cross-enterprise, cross-domain teams people to operate under a common umbrella.

13 System Rules Governing Systems of Systems

The same forces that are making it nearly impossible to enumerate, track and control data shared over networks are also blurring the boundaries and very definition of institutions. Large organizations are hardly recognizable as a single entity to start with. The divisions, subsidiaries, departments and other parts of large organizations are often tantamount to independent institutions. Sometimes the divergent nature of parts is the result of mergers or acquisitions and consequent legacy systems not yet or ever fully integrated to rest of the firm. In fact, it is very common for top decision makers to deliberately reject system re-use or compatibility in favor internally non-interoperable systems that will nonetheless provide a cost, quality or other business advantage in a specific competitive marketplace.

Big and complex enterprise systems are themselves part of one or more larger enterprise system. The module for on-boarding a newly hired employee to an organization may be part of a human resources department integrated enterprise system, or may be part of an identity management enterprise system or a financial management system or any number or other systems and in any combination.

Meanwhile, as these systems generate data, they are individually and together the birth events and original dwelling places of big data. And the big data operates in direct sync with the business, legal and technical dimensions of the environments in which is was created and later resides. From a business vantage point, it is the combination of team, resources and processes of a variety of organizations that comprise a given product or service. As more businesses rely upon the use of deeply integrated APIs and orchestrated series of services from many disparate sources, there may soon be no discernable institution upon which to exercise meaningful controls.

From the information technology vantage point, what once was a simple corporate network is now frequently an extended enterprise enabling the integration of many sets of functional capabilities and operations with a wide variety of external organizations. From the business perspective, it is common for important services, core components and even whole networks of an organization to exist completely outside the ownership, control or detailed knowledge of that organization. The same basic trend plays out accordingly from a legal perspective. A composite of partner agreements, exchange network operating rules, supply chain terms, outsourced provider contracts, cloud based service licenses and a wide variety of other legal instruments and methods reflect and support the distributed nature of enterprises today.

13.1 Identifying Relevant Context

Understanding the relevant context within which data exists is the key to deriving value and controlling risk associated with that data. There are a handful of key aspects of the operational framework for any given system from which practical approaches to confidentiality and data access can be selected. The quintessential question to reveal relevant context any system and data is: What interactions do relevant people have with each other and the data? This question is condensed from the following four basic contextual facets and corresponding inquiries:

People: What, when and where are the relevant organizations or individuals and their respective relationships?

Interactions: What, when and where are the relevant exchanges, transactions and other actions?

Technology: What, when and where are the relevant software, platforms and API or other services?

Data: What, when and where is the data and how does it relate to relevant people, interactions and technology?

Establishing these four facets render otherwise imponderable big data quandaries into tractable scenarios.

By describing these relevant contextual facets, it is now possible to derive and describe the applicable business scenarios, legal requirements and technical capabilities necessary to adequately anticipate and manage privacy and access needs and therefore to enable desired outcomes.

13.2 People and Context

All four facets together comprise a sufficiently complete view upon which to base important decisions regarding big data, but the first inquiry regarding people and their relationships is the most important. For example, as is noted below, big data often depends for its value on personal information, therefore requiring a new deal on data affording individuals deeper visibility and control over their data. Noting which people are relevant will, in many situations, include reference to all individuals with personal data existing within the big data sets. The act of identifying and enumerating relevant people is a precursor to measurably assess whether institutional controls have achieved their purposes.

Describing which people are relevant includes both who they are (i.e. identity information such as names and other identity attributes such as unique identifiers) and also their respective relationships. For any given big data system, a definite number of people and organizations can be identified and therefore actual parties can be established. The relationships with, among and between people in a given system at a given time can likewise be definitely established. Establishing peoples relevant relationships necessarily reveals their respective roles, such as merchant, consumer and advertiser in one context or perhaps insurance claims adjuster, patient and doctor in another context. Consequently, relevant expectations, rights and obligations related to or arising from each role can be defined. A basis to describe the applicable roles, relationships, rights and responsibilities of relevant people goes a long way toward forming a solid foundation for design of systems for big data and the institutional or other controls required.

Inquiring what people are relevant provides a means to determine the exact parties and therefore whether, for instance, some parties may have provided consent to share data or may have a right to access data. Knowing some of the relevant people were, say, acting in their roles as lawyers, treasurers or regulators, can lead to very different expectations regarding certain data based upon professional or fiduciary duties of care that may apply and could require or prohibit the sharing of a particular set of data.

13.3 Interrelated Contexts of Systems and Data

To some extent, any one of a systems contextual aspects enumerated above will interrelate with, interdepend upon and even blur into other aspects. For example, when the transaction being conducted is a credit card authorization then the data-sets can be deduced from the relevant transaction type (which specifies that particular data be transmitted in a particular format and method, including such things as the identity of the card holder, the merchant, the acquiring bang, the issuing bank, the transaction processor, the dollar amount requested for authorization, and so on). When such data is accessed, a specific range of institutional and other controls are triggered and in some contexts, the role and transactions being conducted by the party accessing the data may trigger still other rules.

To illustrate how interrelated contexts can lead to complex multifaceted sets of purposes for access and use of data, consider that a party who is a member in good standing of an exchange network may have a contingent right to access otherwise prohibited data-sets to reconcile transactional records, perform system diagnostics or respond to a wide ranging judicial or regulatory demand. Some of the data subject to, for example, a legitimate accounting reconciliation review of records can be expected to have had simplified metadata associated with very limited summaries of the intended primary purpose for collection and use of the data but can not be expected to in effect repeat all conditions, contingencies and potential situations that might arise under a flexible partner agreement.

To be even more clear, consider that an individual data owner might have provided personal data to get music recommendations or for participate in a social science research study. In this case, the relevant transactions, parties and corresponding systems and data can be clearly noted and a basic purpose for collection and intended future access can be highlighted in a few words. Later, when partners operating under the system rules exercise right or obligations to, for example, trouble shoot a system access issue or responding to an internal departments request for deeper business analytics on their customer base and likely future buying patterns, the access for such later purposes will be differed from the purposes as originally noted. This is, in reality, a significant feature of big date, despite the genuine privacy, confidentiality and other legal issues that are raised with respect to ownership of data. Meanwhile, this dynamically evolving context, intentions and purposes for data access and use strongly suggest a need to ensure system rules for big data can quickly evolve over time, ideally sometimes based upon automated processes or trigger events responded to be software agents or robots.

The proper rules and controls can be determined by describing how key combinations of factors such as the parties and transactions relevant to when it was loaded into a system and the new parties and transactions relevant when that data is later used or shared.

Often, any one piece of relevant information such as which technologies were used will shed light on other important facets of context. For instance, if the big data arises from an air traffic control system then much can be deduced or inferred about the relevant and their respective roles and relationships (e.g. airlines, pilots, ground crew, municipal regulators, approved auditors, etc). Identifying not only what technology exists but also where provides a means for establishing which airport or perhaps region is relevant as well as the applicable national or local laws and other space-based factors. Explicitly identifying when the relevant technology exists provides a means for establishing, for instance, if data created or stored in systems of that period may be subject to other rules or whether the technology of that period processed, structured, formatted or transmitted data in a materially different ways than assumed for current systems. In some situations, the relevant periods might be partly in the past and partly in the future, for example, based on predictive analytics. Knowing that an air traffic control system will be replaced within the next year by a system with very different capabilities or processes may well impact many other factors, including what people, interactions and data will be relevant at that point.

The sub-parts of these components will frequently correspond to different sets of expectations for data access, confidentiality and other outcomes, precisely because various subparts have different parties, interactions, rules, and other knowable context enumerated above. Consider that the part of a system comprising Ground Movement Control applies to vehicles, interactions, parties, etc that differ materially from Terminal Radar Approach Control and so on for many parts of the system, and therefore respective expectations for data access and confidentiality will correspond accordingly. And yet, the applicable rules, relevant parties, specific data, particular interactions and other key aspects can be definitely established for any part or for the entire system at any point in time.

The controls applied by or to any given governmental, corporate or other institution may be profoundly inadequate to guess how confidentiality and data access do apply or could be structured to apply to big data running through that organization. Because each of the enumerated contextual system aspects can be known for a system at any given time, therefore a rational and reliable basis exists to establish confidentiality, data access and other important questions related to the big data of that system. Seeing the surrounding context of big data through the lens of systems affords a realistic and actionable premise to develop and deploy workable controls and reasonable expectations with respect to big data.

13.4 Context and Purpose

The facets of context provide a coherent and complete premise for establishing the actual purpose of data sharing. Determining and recording the purpose for which data was shared with others is quintessential for maintaining confidentiality, privacy and fair information practices in general. There are many examples of laws and contracts stipulating that the purpose of third party access to an individuals personal information must be noted and in some cases must also be disclosed to the individual themselves. How is purpose described and what amount or type of information is needed for this rule to be effective Phrases like purpose of access was for quality control or for business intelligence may prevent disclosure of proprietary or sensitive analytics techniques but ambiguous terms will not suitably describe the relevant context giving rise to the access.

A very good example of forward looking and comprehensive personal data access and protection statue is the Massachusetts Fair Information Practices Act (FIPA). The statute applies to personal data held in any system of the state government, which is a very large organization. The state government comprises the largest single employer, real property owner, operator of the largest vehicle fleet within Massachusetts and would be a fortune 100 company, if it were private needless to say, the organization holds a vast amount and variety of personal data. The FIPA statute requires the purpose for each instance of third party access to an individuals personal data be recorded and disclosed upon request of that individual. Specifically, the Act mandates, at Chapter 66, Sec 2(f)&(g) that:

Every holder maintaining personal data shallmaintain a complete and accurate record of every access to and every use of any personal data by persons or organizations outside of or other than the holder of the data, including the identity of all such persons and organizations which have gained access to the personal data and their intended use of such data [and shall] make available to a data subject upon his request in a form comprehensible to him, a list of the uses made of his personal data, including the identity of all persons and organizations which have gained access to the data https://malegislature.gov/Laws/GeneralLaws/PartI/TitleX/Chapter66A/Section2

While these sections of the statute are best-of-class reflection of international standards for fair information practices, nonetheless, it is ambiguous what constitutes a complete and accurate record or each use and intended use of personal information. The federal rendition of fair information practices is most recently reflected by the National Strategy for Trusted Identities in Cyberspace, which provides:

Organizations should be transparent and notify individuals regarding collection, use, dissemination, and maintenance of personally identifiable information[and] Organizations should specifically articulate the authority that permits the collection of PII and specifically articulate the purpose or purposes for which the PII is intended to be used. http://www.nist.gov/nstic/NSTIC-FIPPs.pdf

Again, what is sufficient to constitutes notice of a specific articulation of purposes for which personal information is collected and intended to be used is not clear. Having a re-usable metadata tagging or classification system referring to the relevant purposes would be an important capability. However, such a system would require first establishing a stable set of widely understood and agreed statements of purpose. In order to understand a given purpose or intent to access information, one must understand the relevant context surrounding that access. To the extent that a third party individual or organization discloses their purpose for accessing personal information, that information should include or be easily linked to a log of all the relevant parties, transactions, systems and relevant data-sets associated with the access. The specific purpose for the third party access is squarely part of the transactional facet of context and should be recorded in a transaction log of third party access. Taken in context, a fair and full understanding of how and why personal information was accessed and used is possible.

14 System Rules and Legal Agreements

The rules governing systems are perhaps the more important lever of institutional control for big data. Especially when large, complex systems comprised of smaller or overlapping systems crossing business boundaries are in play, it is the system rules agreed by al parties that actually define, bound and govern the expectations about data and its uses.

The process of deliberately aligning, harmonizing and integrating business, legal and technical systems architecture makes it possible for software like openPDS to be used by institutions as part of a safe transition to a big data footing and to realize the new deal on data. These results do not happen by architecture alone, in the enterprise architecture sense. Rather, the results are most powerful and effective when expressed as a common set of rules agreed to apply to each system interoperating under a given API-based or other interoperability umbrella. Examples of inter-domain and cross-context system rules can be found at http://civics.com/trust-frameworks generally and a model reference implementation of system rules designed for multi-enterprise use of openPDS can be found at: https://github.com/HumanDynamics/SystemRules.

From a legal vantage point, asking what actions with big data are being conducted by which parties can enable a lawyer to determine the applicable laws or other rules. When properly defined, designed, developed and deployed, system rules provide a definite set of rules applicable based upon the relevant context. By ensuring at each phase of a system the relevant context is noted in a common method and accessible as needed, it is possible for any participant to establish for themselves precisely what statutes, regulations, judicial holdings, certifications, licenses, contracts and other more specific rules will apply.

Lawyers need to act specific questions to determine exactly by what time, place and manner the relevant parties in fact conducted specific transactions or other interactions. Legal analysis and judgment about what rules apply, by whom and how the rule will be applied, and the likely result is fundamentally about applying law to facts. This process is one simple way to define the practice of law. Knowing this can be a valuable asset because it is then possible to consciously create and preserve the very types of information that will be needed for proper, context-anchored and fact-specific legal analysis and advice. Further, it allows the design of system rules and agreements that fit hand-and-glove with the legal dimension of systems.

15 Heterogeneity and Commonality for Cross-Domain Trust Infrastructure

Some controls will have to be applied by and to many different institutions in order to prevent leakage and to achieve predictable results. To address privacy and data access, a variety of institutional controls will be needed, and in varied combinations with engineering and other controls. Establishing an appropriate set of institutional controls will depend upon such factors as the type of institution using big data, what big data being used, how it is used, who is using it, and when it is used.

The personal data ecosystem of the future will most likely be consisting of heterogeneous systems that are connected by the Internet as the common medium of data transport. In order to achieve scalability of data access in this heterogeneous construction, the standardization of protocols and application programming interface (APIs) and the deployment of these standards among personal data ecosystem entities will be a crucial factor. The use of standards leads to better integration with existing infrastructures that implement identity and authorization services. Many of these identity, authorization and attribute services exist today to differing degrees, and one of the major challenges of the Internet industry today is to continue the development of these standards.

A second aspect related to standardization and integrated services is the need for open source implementations of these standards. The history of computing and the Internet for the last two decades have shown that open source implementations allows the widespread acceptance of new technologies at the grassroots level. The availability of publicly readable and usable open source code promotes testing and experimentation at the grassroots level, leading to the use of the same open source code as the basis for commercial products and services.

For many institutions with an Enterprise outlook and mindset, the emergence of a personal data ecosystem and personal data market offers promising sources of revenue. However, a new outlook on IT governance coupled with new change management approaches are need in order for many of these institutions to meet the opportunities and benefits of the new personal data ecosystem. The MIT Kerberos and Internet Trust (MIT-KIT) Consortium seeks to provide a venue and forum for these institutions to collaborate on open architecture initiatives that answer the specific business goals of these institutions in the personal data space. Such open architectures initiatives should as a key priority promote the development of new standards for the personal data ecosystem, as such standards always benefits these institutions both in the short term (e.g. standing-up new services) and in the long-term (e.g. retaining investments in infrastructure). The MIT-KIT also seeks to deliver open-source implementations of these standards as a low-cost means to seed the ecosystem with new services that expose new business opportunities based on personal data. As such, a high level coordination can be established where standards, open source implementations, new personal data services and the private sector can collaborate to establish the emerging personal data market.

16 Scenarios of Use in Context

Supporting the effective development of institutional controls for big data requires an understanding of how to define and work with the applicable context surrounding the scenarios within which the big data exists. In particular, the New Deal on Data will require a set of Institutional Controls involving governance, business, legal and technical aspects that are knowable only with reference to the relevant context of a factually based scenario of use. The following scenarios demonstrate signature features of the New Deal on Data in various contexts and serve as an anchor to evaluate what Institutional Controls are well aligned.

16.1 Example Scenario: Research Systems

Computational Social Science (CSS) studies are based on data collected often with an extremely high resolution and scale. Using computational power combined with mathematical models, such data can be used to provide insights into human nature. Much of the data collected, for example mobility traces are sensitive and private; most individuals would feel uncomfortable sharing them publicly. The need for solutions to ensure the privacy of the individuals has grown alongside the data collection efforts.

The data collection in the CSS context is based on the informed consent of the participants. Countries have different bodies regulating such studies, for example Institutional Research Boards (IRBs) in the US. Although certain minimal requirements for implementing informed consent exist[TODO: reference], they are often not very well suited for the large-scale studies, where the amount and sensitivity of the data calls for sophisticated privacy controls. As the scale of the studies grows, in terms of the number of participants, collected bits per user, and duration, the EULA-style informed consent is no longer sufficient and makes it hard to claim that participants in fact expressed informed consent.

This year we have deployed a 1,000 phones study at Technical University of Denmark, where we handed out mobile phones to freshmen students in order to study their networks and social behavior in the important change moment of their lives, when they join the university. The study, called SensibleDTU, uses not only data collected from the mobile phones (location, Bluetooth-based proximity, call and sms logs etc.) but also data collected from social networks, questionnaires filled out by participants, behavior in economic games and so on. As the data is collected in the context of the university, there is potentially a big issues of students feeling obliged to participate in the study, feeling that their grades may depend on it, or that the data may influence their grades. In this context, we see the implementation of Living Informed Consent not only as a technical mean to put participants in control of the data we collect, but also to convey the message about the opt-in nature of the study, the boundaries of the data usage, and parties accessing the data.

It is not feasible to explain the terms and answer all the questions to all 1,000 students personally. The controls must be self-explanatory as much as possible, and guide the user from the first opening of the link to the study to the grant of the authorizations. At the same time, every click made by the user, should be an expression of an informed decision, so the user journey must be a balance of guidance and understanding. For this reason we have created a set of web applications, allowing the users to enroll into the study, express informed consent, and interact with their data.

As the study will last for several years, hopefully allowing us to see the life of a student from the very first friendships made until the graduation party, the consent must remain alive. It is again a matter of balance: we do not want the participants to feel under constant surveillance (as they are not, the data is used mostly in aggregated form), at the same time to remember that in fact, the data is being collected and used. We are still trying to understand how to achieve this equilibrium: how often should we remind the users about the collection effort? should they re-authorize applications from time to time? We see a great hope in the applications we create for the users to provide certain services, simple such as life-logging where they can see how active they are, what are their top places etc. and more advanced, such as artistic visualizations of their social networks. Making the user aware of the data by transforming them into value, can greatly benefit the privacy, making users constantly aware what is being collected, but also what kind of value they can get out of it.

When a study of such scale is deployed, the particular experiments and sub-studies may not be exactly defined from the very beginning. The initial deployment is a creation of a testbed, where shorter or longer experiments can take place; for example part of the population may participate in the experiment of quantifying the impact of feedback application on their activity levels. Being able to create such experiments in an efficient way is a huge value for the researchers. To do that in the most frictionless way, we give the users the choice to opt-in to those additional experiments, providing some financial or other benefits. This is only possible if there is a notion of identity of the participants, stronger and more useful than a piece of paper with a signature. This identity allows us to reach out to people, offer them additional experiments, and let them agree or disagree to them.

This touches upon the re-usability of data, as the new experiments may require additional data to be collected, but also have access to all the existing data, based on user authorization. We can imagine going even further, where entirely different studies can re-use participants data from a previous study based on their authorization. When the data are owned buy the users, they are free to authorize access to them to any party that requests it. We can see a New Deal on Data pattern here: rather than services (studies) talking to each other about the user data, they talk directly to the users, seeking their authorization. This can address a very important problem in the research context, the data re-use in a privacy-aware manner. Rather than publishing a static dataset, where the users have lost control over their data, live and fresh data can be continuously accessed by any study that the user agrees to be a part of.

Many studies will be willing to offer money or other value for the access to the data. Other will provide the user the opportunity to have new data collected. This way, the data collection becomes an opportunity for the user to enrich their personal dataset, and to benefit from it in the future. Join our study and we will provide you with a smartphone and collect your movement patterns for a year; we will do science and you will gain new data that can get you better value or deals in different services. You may now be eligible for a different study. Or your music recommendation may get better, because your music service can make a use of this extra data. Your data.

16.2 Scenarios of Use Today, Tomorrow and the Day After

By inquiring into and noting the four facets of relevant context described above, it is possible to describe the basic material contours of any scenario within which big data exists such that the operational framework and adequate approaches to access, use, confidentiality and other key interests can be sustainably balanced. In a commercial scenario the relevant people might be a consumer, merchants, banks, products manufacturers, third party app developers and individual members of that consumers bowling team. The relevant transactions might be a purchase of goods by the consumer from the merchant and the corresponding app that was embedded in the goods and the downstream transaction of involving the consumer now transacting with the merchant bowling alley and interacting with a bowling team, with whom activity and sports performance data are shared and aggregated and further mashed up. The rest of the context can be described for any given scenario and this all could be expressed specifically rather than by role simply by running a report from the system to indicate it was in fact John Doe, of openpds.org/owner/571 purchasing a smart bowling ball from Bowl-a-Tronic of bowlappgood.com/store/221 and so on for each party that played a role in the relevant scenario. The same techniques, used for scenarios in other economic sectors and social endeavors shed light on the fundamental nature and implications of big data and options for the use of operational frameworks acting across domains to balance privacy and access, among other intersts.

This book represents a high value opportunity to take stalk of the current state and dominant trends related to big data and help to illuminate important choices at a moment of early adoption, dynamic innovation and wide open possibilities. By contemplating the relevant contexts of todays scenarios of use in, say, the fields of education, entertainment, government, manufacturing, transportation and many other core anchors of human activity, we have traction to postulate how todays prevailing trends are likely to result and what changes perhaps quite small but of profound long term impact could lead to materially different better outcomes. Consider that if the essence of the New Deal on Data were accepted today, or soon, the nature, tenor, capabilities and experience of living by future generations could be unrecognizably better. Simply extrapolate from the current anomalous practices regarding personal data and individual identity and push forward the timeline by 5, 10, 20 years and beyond. The current trajectory ends up with dystopian scenarios that effectively reverse hard fought but easily lost constitutional deal of the United States and social compact of common law and compatible societies.

By contrast, by adopting the New Deal on Data now it is entirely possible to enjoy a period of unprecedented prosperity and invention even before the new deal on data frameworks are formally launched. This is because the uncertainly and confusion about the basic premises and expectations around personal data and identity will be resolved and so investment and risk taking on a firm foundation can be unleashed. Many policy techniques can be used to make adoption of the new deal on data all but impossible to oppose on grounds of cost, disruption or over regulation. For instance, generous easy to implement support such as phase-in periods of many years, opportunities for delays and flexible approaches to adoption, etc.

17 Future Research

Our traditional methods of testing and improving government, organizations, and so on are of limited use in building a data driven society. Even the scientific method as we normally use it no longer works, because there are so many potential connections that our standard statistical tools generate nonsense results.

The reason is that with such rich data, you can easily uncover misleading correlations. For instance, lets imagine we discover that people who are unusually active are more likely to get the flu. This is a real example: when we examined the minute-by-minute behavior of a small university community a real-time flow of gigabytes per day for an entire year we noticed that an unusual level of running around often predicted onset of the flu. But if we can only analyze the data using traditional statistical methods, we have the problem of why is it true? Is it because flu virus makes us more active in order to spread itself more quickly? Or did interacting with many more people than usual make you more likely to catch the flu? Or is it something else? From the real-time stream of data by itself you just cant know.

The point here is that normal analysis methods don't suffice to answer these sorts of questions, because we dont know all the possible alternatives and so we cant form a limited, testable number of clear hypotheses. Instead, we need to devise new ways to test the causality of connections in the real world. We can no longer rely on laboratory experiments; we need to actually do the experiments in the real world, and usually on massive, real-time streams of data.

18 Research on Design and Deployment of Big Data Systems

The highest value, lowest risks and overall best outcomes can be achieved most efficiently by applying top current research to design and deployment of the coming global wave of big data systems. To understand and address the unique problems and prospects affiliated with big data, the relevant context must be identified and corresponding rules-driven capabilities must be designed into the underlying systems.

People and/or rules engines can determine the right rules to apply to data when the right information is reliably attached to or logically associated with that data in a standard manner. Any system that can make, use, receive or share big data must be capable of associating provenance and purpose for all data in a common and actionable manner. Requiring a lot of narrative documentation and background about the nuances and circumstances surrounding every data set is both impractical and counterproductive. By contrast, a small amount metadata listing or reliably linking to the parties, transactions, systems and provenance of the data would suffice. This relevant context together

It is important for science and research to develop further solutions and options ensuring contextually appropriate rules can be applied by big data systems. For rules to be effectively applied, systems must not only be able to establish which rules apply but also support the right functional capabilities and have appropriate information structure, format and meta-data.

Today, computational social science can provide unprecedented insights into the business, legal and technical dimensions of big data driven systems. Harnessing these insights it will be possible to conduct research enabling common design patterns and reference implementations for responsive enterprise architectures that can orchestrate services and adapt rules based on dynamic real-time big data analytics. Advanced analytics reveals the reality of situations, and can be a powerful guide to the further optimization of financial management, user experience and control, conditions catalyzing innovation and other key inputs to overall economic impact.

Some capabilities will likely be essential to all big data systems, such as highly scalable active storage, standard methods for integration with other big data systems and a processing architecture enabling high speed statistical analytics. But there are and will continue to emerge multiple types of big data systems. Some functions or controls will likely be important - or even feasible - only for certain types of future systems. For instance, it is reasonable to expect some systems will specialize in enormous volumes of entirely non-personal data from many real-time sources (e.g. for soil science, materials engineering, astronomy, etc) while other big data systems will hinge upon mass quantities of highly sensitive personal information (e.g. for clinical medicine, education and life-long learning, social entertainment, etc).

While some capabilities, such as ingesting and processing astronomical data-sets, will be unique to only a subset of big data systems it is reasonable to anticipate that data will be increasingly cross-tabulated, merged and otherwise shared with other systems and data. It can be nearly impossible to conclusively predict for the entire life of a system what data will be received by, created in or transmitted from that system at the design phase. This prediction is all the harder to make when the systems are intended for big data.

The four contextual facets of people, interactions, technology and data were initially developed to provide a sound underpinning for the design of new big data and web 2.0 systems. The existing systems design and development processes of establishing business cases, use cases, agile stories, functional requirements, etc. do not reliably identify the factors most relevant to use of big data, especially in a web 2.0 massively distributed environment. The four facets can also be used to analyze appropriate, required or prohibited uses for existing big data systems. However, it can be difficult to extract the relevant information from or apply any effective control on systems used for big data but designed to achieve limited purposes in hierarchical closed environments.

Big data, by its nature, represents a new set of business, legal and technical capabilities and requirements. Most of the worlds systems today are not capable of ingesting, storing, using or dynamically flowing big data with other systems. Considering that a) big data is of high value immediately and higher value in the short and long terms, and b) the young but competitive marketplace of big data system components, platforms, applications and other solutions is a hotbed of innovation it can be predicted that a transition to big data systems will continue. The key observation is that virtually all big data systems have yet to be designed, implemented, customized or deployed. Institutions that are the current early adopters of todays big data system will soon replace those systems and the rest of the world will adopt big data systems in phases over time. Based upon this observation,

19 Research on Big Data for Design of Institutions

Using massive, live data to design institutions and policies is outside of our normal way of managing things. We live in an era that builds on centuries of science and engineering, and the standard choices for improving systems, governments, organizations, and so on are fairly well understood. Therefore our scientific experiments normally need only consider a few clear alternatives (i.e., plausible hypotheses).

But with the coming of big data, we are going to be operating very much out of our old, familiar ballpark. These data are often indirect and noisy, and so interpretation of the data requires greater care than is usual. Even more importantly, a great deal of the data is about human behavior, and the questions are ones that seek to connect physical conditions to social outcomes. Until we have a solid, well-proven and quantitative theory of social physics, we wont be able to formulate and test hypotheses in the way we can when we design bridges or develop new drugs.

Therefore, we must move beyond the closed, laboratory-based question-and-answering process that we currently use and begin to manage our society in a new way. We have to begin to test connections in the real world far earlier and more frequently than we have ever had to do before, using the methods my research group and I have developed for the Friends and Family study or the Social Evolution study. We need to construct Living Laboratories communities willing to try a new way of doing things or, to put it bluntly, to be guinea pigs in order to test and prove our ideas. This is new territory and so it is important for us to constantly try out new ideas in the real world in order to see what works and what doesnt.

An example of such a Living Lab is the ‘open data city just launched by one author (Pentland) with the city of Trento in Italy, along with Telecom Italia, Telefonica, the research university Fondazione Bruno Kessler, the Institute for Data Driven Design, and local companies. Importantly, this Living Lab has the approval and informed consent of all its participants they know that they are part of a gigantic experiment whose goal is to invent a better way of living. More detail on this Living Lab can be found at http://www.mobileterritoriallab.eu/

The goal of this Living Lab is to develop new ways of sharing data to promote greater civic engagement and exploration. One specific goal is to build upon and test trust-network software such as our openPDS (Personal Data Store) system . Tools such as openPDS make it safe for individuals to share personal data (e.g., health data, facts about your children) by controlling where your data go and what is done with them.

The specific research questions we are exploring depend upon a set of personal data services designed to enable users to collect, store, manage, disclose, share and use data about themselves. These data can be used for the personal self-empowerment of each member, or (when aggregated) for the improvement of the community through data commons that enable social network incentives. The ability to share data safely should enable better idea flow among individuals, companies, and government, and we want to see if these tools can in fact increase productivity and creative output at the scale of an entire city.

An example of an application enabled by the openPDS trust frame work is sharing of best practices among families with young children. How do other families spend their money? How much do they get out and socialize? Which preschools or doctors do people stay with for the longest time? Once the individual gives permission, our openPDS system allows such personal data to be collected, anonymized and shared with other young families safely and automatically.

The openPDS system lets the community of young families learn from each other without the work of entering data by hand or the risk of sharing through current social media. While the Trento experiment is still in its early days, the initial reaction from participating families is that these sorts of data sharing capabilities are valuable, and they feel safe sharing their data using the openPDS system.

The Trento Living Lab will let us investigate how to deal with the sensitivities of collecting and using deeply personal data in real-world situations. In particular, the Lab will be used as a pilot for the New Deal on Data and for new ways to give users control of the use of their personal data. For example, we will explore different techniques and methodologies to protect the users privacy while at the same time being able to use these personal data to generate a useful data commons. We will also explore different user interfaces for privacy settings, for configuring the data collected, for the data disclosed to applications and for those shared with other users, all in the context of a trust framework.