

# 1 **Operational Framework: Institutional Controls - The New Deal** 2 **on Data**

3 Daniel "Dazza" Greenwood<sup>1,\*</sup>, Arkadiusz Stopczynski<sup>1,2</sup>, Brian Sweatt<sup>1</sup>, Thomas Hardjono<sup>1</sup>,  
4 Alex Sandy Pentland<sup>1</sup>

5 **1 MIT**

6 **2 DTU**

7 \* **E-mail: dazza@civics.com**

## 8 **Contents**

9	<b>1 The New Realities of Living in a Big Data Society (Arek)</b>	<b>2</b>
10	<b>2 The New Deal on Data (Arek)</b>	<b>4</b>
11	<b>3 Personal Data: Emergence of a New Asset Class (Thomas)</b>	<b>6</b>
12	<b>4 Enforcing the New Deal on Data (Dazza)</b>	<b>10</b>
13	<b>5 Transitioning End-User Assent Practices (Arek)</b>	<b>13</b>
14	<b>6 Business, Legal, and Technical Dimensions of Big Data Systems (Dazza)</b>	<b>14</b>
15	<b>7 Big Data and Personal Data Institutional Controls (Thomas)</b>	<b>16</b>
16	<b>8 Scenarios of Use in Context (Dazza)</b>	<b>20</b>
17	8.1 Example Scenario: Research Systems . . . . .	21
18	8.2 Scenarios of Use Today, Tomorrow and the Day After . . . . .	24
19	<b>9 Future Research (Brian)</b>	<b>26</b>
20	9.1 Research on Design and Deployment of Big Data Systems . . . . .	26
21	9.2 Research on Big Data for Design of Institutions . . . . .	29

## 23 1 The New Realities of Living in a Big Data Society (Arek)

24 To realize the promise and prospects of a Big Data society and avoid its security and confiden-  
25 tiality perils, institutions are updating operational frameworks governing business, legal, and  
26 technical dimensions of their internal organization and interactions with the outside world. In  
27 this chapter we explore the emergence of the Big Data Society, outline ways to support it in the  
28 context of institutional controls, and describe future directions of research and development.

29 The control points traditionally relied upon as part of corporate governance, management  
30 oversight, legal compliance, and enterprise architecture must evolve and expand to match op-  
31 erational frameworks for Big Data. An operational framework used for a Big Data-driven or-  
32 ganization requires a balanced set of institutional controls. These institutional controls must  
33 support and reflect greater user control over personal data, and large scale interoperability for  
34 data sharing between and among institutions. Core capabilities of these controls include re-  
35 sponsive rule-based systems governance and fine-grained authorizations for distributed rights  
36 management.

37 Sustaining a healthy, safe, and efficient society is a scientific and engineering challenge go-  
38 ing back to the 1800s, when the Industrial Revolution spurred rapid urban growth, creating  
39 huge social and environmental problems. The remedy then was to build centralized networks  
40 that delivered clean water and safe food, enabled commerce, removed waste, provided energy,  
41 facilitated transportation, and offered access to centralized healthcare, police, and educational  
42 services. Those networks formed the backbone of the society as we know it today.

43 These century-old solutions are however becoming increasingly obsolete and inefficient. We  
44 have cities jammed with traffic, world-wide outbreaks of disease that are seemingly unstoppable,  
45 and political institutions that are deadlocked and unable to act. We face the challenges of global  
46 warming, uncertain energy, water, and food supplies, and a rising population and urbanization,  
47 that will add 350 million people to the urban population by 2025 in China alone [14].

48 It does not have to be this way. We can have cities that are protected from pandemics, energy  
49 efficient, have secure food and water supplies, and have much better government. To reach these  
50 goals, however, we need to radically rethink our approach. Rather than static fixed systems,  
51 separated by function — water, food, waste, transport, education, energy — we must consider  
52 them as dynamic, data-driven networks. Instead of focusing only on access and distribution,  
53 we need the networked and self-regulating systems, driven by the needs and preferences of the  
54 citizens. Finally, we need to create the channels for the society to agree upon and communicate  
55 those needs.

56 To ensure a sustainable future society, we must use our new technologies to create a *nervous*  
57 *system* maintaining the stability of government, energy, and public health systems around the  
58 globe. Our digital feedback technologies are today capable of creating a level of dynamic re-  
59 sponsiveness that our larger, more complicated modern society requires. We must reinvent the  
60 systems of the societies within a control framework: sensing the situation, combining these obser-  
61 vations with models of demand and dynamic reaction, and finally using the resulting predictions  
62 to tune the system to match the demands.

63 The engine driving this nervous system is Big Data: the newly ubiquitous digital data, now  
64 available about all aspects of human life. We can analyze patterns of human experience and  
65 ideas exchange within the *digital breadcrumbs* that we all leave behind as we move through the  
66 world: call records, credit card transactions, GPS location fixes, among others. By recording  
67 our choices, these data tell the story of our lives. And this may be very different from what  
68 we decide to put on Facebook or Twitter; our postings there are what we choose to tell people,  
69 edited according to the standards of the day and filtered to match the persona we are building.  
70 Mining social networks can give some great insights about human nature [4, 27, 41]; who we  
71 really are is however even more accurately determined by where we spend our time and which  
72 things we buy, rather than just what we say we do [26].

73 The process of analyzing the patterns within these digital breadcrumbs is called reality  
74 mining [13, 31], and through it we can learn an enormous amount about who we are. The

Human Dynamics research group at MIT have found that we can use them to tell if we are likely to get diabetes [32], or whether we are the sort of person who will pay back loans [33]. 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 [29]. For this reason the magazine Technology Review named our development of reality mining as one of the ten technologies that will change the world [16].

## 2 The New Deal on Data (Arek)

The digital breadcrumbs we leave behind provide clues about who we are, what we do and want. This makes these personal data — data about individuals — immensely valuable, both for public good 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” [22]. This new ability to see the details of every interaction can be however used for good or for ill. Therefore, maintaining protection of personal privacy and freedom is critical to our future success as a society. We need to enable even more data sharing for the public good; at the same time, we need to do a much better job in protecting the privacy of the individuals.

A successful data-driven society must be able to guarantee that our data will not be abused; perhaps especially that government will not abuse the power conferred by access to such fine-grain data. The abuses may be directly targeted at users, for example by offering them higher insurance rates based on their shopping history, or create problems for the entire society in longer run, for example by limiting user choices and closing them into information bubbles [18]. To achieve the positive possibilities of the new society, we require the *New Deal on Data*, workable guarantees that the data needed for public good are readily available while at the same time protecting the citizenry [31].

The key insight that motivates the idea of the New Deal on Data is that our data are worth more when shared, because these aggregated data — averaged, combined across population, and often distilled to high-level features — inform improvements in systems such as public health,

101 transportation, and government. For instance, we have demonstrated that data about the way  
102 we behave and where we go can be used to minimize the spread of infectious disease [25,32]. Our  
103 research has reported how we were able to use these digital breadcrumbs to track the spread of  
104 influenza from person to person on an individual level. And if we can see it, we can stop it.

105 Similarly, if we are worried about global warming, these shared, aggregated data can show us  
106 how patterns of mobility relate to productivity [30]. In turn, this provides us with the ability to  
107 design cities that are more productive and, at the same time, more energy efficient. But in order  
108 to obtain these results and make a greener world, we need to be able to see the people moving  
109 around; this depends on many people willing to contribute their data, even if only anonymously  
110 and in aggregate.

111 To enable sharing of personal data and experiences, we need secure technology and regulation  
112 that allow individuals to safely and conveniently share personal information with each other,  
113 with corporations, and with government. Consequently, the heart of the New Deal on Data  
114 must be to provide both regulatory standards and financial incentives that entice owners to  
115 share data, while at the same time serving the interests of both individuals and society at large.  
116 We must promote greater idea flow among individuals, not just corporations or government  
117 departments.

118 Unfortunately, today most personal data are siloed off in private companies and therefore  
119 largely unavailable. Private organizations collect the vast majority of the personal data in the  
120 form of mobility patterns, financial transactions, phone and Internet communications. These  
121 data must not remain the exclusive domain of private companies, because then they are less  
122 likely to contribute to the common good. These private organizations must be thus the key  
123 players in the New Deal on Data framework for privacy and data control. Likewise, these data  
124 should not become the exclusive domain of the government, as this will not serve the public  
125 interest of transparency; we should be suspicious of trusting the government with such power.  
126 Ultimately, the entities who should be empowered to share and make decisions about their data,  
127 are people themselves: users, participants, citizens.

128       The ultimate goal is to provide the society with tools to analyze and understand what needs  
 129 to be done, and to reach the consensus on how to do it. This goes beyond just creating more  
 130 communication platforms; the assumption that more interactions between users will result in  
 131 better decisions being made, may be very misleading. Although in the recent years we have  
 132 seen some great examples of using social networks for better organization in society, for example  
 133 during political protests [6, 17], we are not even close to the point where we can start reaching  
 134 consensus about the big problems: epidemics, climate change, pollution. The discussions must  
 135 be data driven, involving both experts and wisdom of the crowds – users themselves interested  
 136 in improving the society. The problems we are dealing with as a now global society are not  
 137 easy. We are responsible for many of them, and being able to tackle them on a global scale is  
 138 necessary for our, mankind, survival.

### 139   **3   Personal Data: Emergence of a New Asset Class (Thomas)**

140 It has long been recognized that the first step to promoting liquidity in land and commodity  
 141 markets is to guarantee ownership rights so that people can safely buy and sell. Similarly, the  
 142 first step toward creating greater idea and idea flow (‘idea liquidity’) is to define ownership rights.  
 143 The only politically viable course is to give individual citizens rights over data that are about  
 144 them and in fact, in the European Union these rights flow directly from the constitution **AS:**  
 145 **Citation? There is no ‘EU constitution’ per se. .** We need to recognize personal data  
 146 as a valuable asset of the individual that is given to companies and government in return for  
 147 services.

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

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

- 153 • You have the right to full control over the use of your data. The terms of use must be opt-  
154 in and clearly explained in plain language. If you are not happy with the way a company  
155 uses your data, you can remove the data, just as you would close your account with a bank  
156 that is not providing satisfactory service.
- 157 • You have the right to dispose of or distribute your data. You have the option to have data  
158 about you destroyed or redeployed elsewhere.

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

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

167 In 2007, one author (Pentland) first proposed the New Deal on Data to the World Economic  
168 Forum [42]. Since then, this idea has run through various discussions and eventually helped  
169 shape the 2012 Consumer Data Bill of Rights in the United States, along with a matching  
170 declaration on Personal Data Rights in the EU. These new regulations hope to accomplish the  
171 combined trick of breaking data out of the current silos, thus enabling public goods, while at  
172 the same time giving individuals greater control over data about them. But, of course this is  
173 still a work in progress and the battle for individual control of personal data rages onward.

174 The World Economic Forum (WEF) has dubbed personal data as the “New Oil” or resource  
175 of the 21st century [42]. The discovery of oil and the subsequent development of the oil industry  
176 over the past 100 years has spurred not only the development of the automobile industry but also  
177 the creation of the global transportation infrastructure, including the massive freeway networks  
178 that we see today in the developed nations. The “personal data sector” of the economy today is  
179 still in its infancy, its state akin to the oil industry at the late 1890s prior to the development of

180 the Model-T Ford automobile. The productive collaboration between the Government (building  
 181 the state owned freeways), the private sector (mining and refining oil, building automobiles) and  
 182 the citizen (the user-base of these services) allowed the develop nations to expand its economies  
 183 by creating new markets adjacent to the automobile and oil industries.

184 If personal data, as the new oil, is to reach its global economic potential, there needs to be  
 185 a productive collaboration between all the stakeholders in the establishment of a *personal data*  
 186 *ecosystem*. As mentioned in [42], a number of fundamental questions about privacy, property,  
 187 global governance, human rights — essentially around who should benefit from the products  
 188 and services built upon personal data — are major uncertainties shaping the opportunity. The  
 189 rapid rate of technological change and commercialization in using personal data is undermining  
 190 end user confidence and trust.

191 The current personal data ecosystem is fragmented and inefficient. Too much leverage is  
 192 currently being accorded to service providers that on-board and register end-users. These siloed  
 193 repositories of personal data exemplifies the fragmentation of the ecosystem. These repositories  
 194 contain data of varying qualities. Some are attributes of persons that are unverified, while  
 195 other represent higher quality data that have been cross-correlated with other data points of the  
 196 end-user.

197 For many participants, the risks and liabilities exceed the economic returns. Besides not  
 198 having the infrastructure and tools to manage personal data, many end-users simply do not see  
 199 the benefit of fully participating in the ecosystem. The current focus of many Internet-based  
 200 service providers is to capture as much personal data from the end-user and to sell this data into  
 201 the advertising industry. Personal privacy concerns are thus inadequately addressed at best,  
 202 or simply overlook in the majority of the cases. The current technologies and laws fall short  
 203 of providing the legal and technical infrastructure needed to support a well-functioning digital  
 204 economy.

205 Recently, we have shown how challenging, but also feasible, it is to open such institutional  
 206 Big Data. In the Data For Development (D4D) Challenge <http://www.d4d.orange.com/home>,



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. For example, one team developed a model for how disease spread in the country and demonstrated that information campaigns based on one-to-one phone conversations among members of social groups can be an effective countermeasure [24]. In releasing and analysing this data, the privacy of the people who generated the data 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 re-identification or other nefarious purposes. As we have seen in several cases, such as the Netflix Prize privacy disaster [28] and other similar privacy breaches [36], true anonymization is extremely hard. In the Unique in the Crowd [10], de Montjoye et al. showed that even though human beings are highly predictable [34], we are also very unique. Having access to one dataset, it may be 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 easier it gets to identify a person from this type of data.

The report of the World Economic Forum [42] also suggest a way forward by recommending a number of areas where efforts could be directed:

- Alignment of key stakeholders: Citizens, the private sector and the public sector need to work in support of one another. Efforts such as NSTIC [37] — albeit still in its infancy — represents a promising direction for a global collaboration.
- Viewing “data as money”: There needs to be a new change in mindset where an individual’s personal data items are viewed and treated in the same way as their money. These personal data items would reside in an “account” (like a bank account) where it would be controlled, managed, exchanged and accounted for just like personal banking services operate today.
- End-user centricity: All entities in the ecosystem need to recognize that end-users are vital and independent stakeholders in the co-creation and value exchange of services and

233 experiences. Efforts such as the *User managed Access* (UMA) initiative [2] point in the  
234 right direction by designing systems that are user-centric and managed by the user.

235 Opening data from the silos by publishing static datasets — collected at some point and  
236 unchanging — is important, but it is only the first step. We can do even more substantial things  
237 when the data is available in real time and can become part of a society’s nervous system.  
238 Epidemics can be monitored and prevented in real time [32], underperforming students can be  
239 helped, and people with health risks can be treated before they get sick [9]. The same data can  
240 potentially be used for stalking, burglarizing one’s home, and as justification to charge people  
241 more for an insurance policy.

## 242 4 Enforcing the New Deal on Data (Dazza)

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

251 One current best practice is a system of data sharing called trust networks. Trust networks  
252 are a combination of networked computers and legal rules defining and governing expectations  
253 regarding data. With respect to data belonging to individuals, these networks of technical and  
254 legal rules keeps track of user permissions for each piece of personal data, and a legal contract  
255 that specifies both what you can and cannot do with the data and what happens if there is a  
256 violation of the permissions. For example, in such a system all personal data can have attached  
257 labels specifying what the data can and cannot be used for. These labels are exactly matched

258 by the network's system rules and terms in legal contracts between all the participants, stating  
259 penalties for not obeying the permission labels. These rules can, and often do, reference or  
260 require audits of relevant systems and data use, demonstrating how traditional internal controls  
261 can be leveraged as part of the transition to more novel trust models.

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

273 Having system rules applicable to the networks, applications, and data as well as all the  
274 services providers other intermediaries, and the users themselves is the mechanism for estab-  
275 lishing and operating a trust network. System rules are sometimes called operating regulations  
276 in the credit card context, or known as trust frameworks in the identity federations context, or  
277 trading partner agreements in a supply value chain context. There are many general examples of  
278 multiparty shared architectural and contractual rules that share the generic characteristic of cre-  
279 ating binding obligations and enforceable expectations on all participants in scalable networks.  
280 Another common characteristic of the system rules design pattern is that the participants in  
281 the network can be widely distributed across very heterogeneous business ownership boundaries,  
282 legal governance structures, and technical security domains. Yet, the parties need not agree to  
283 conform all or most aspects of their basic roles, relationships, and activities in order to connect  
284 to to systems of a trust network. Cross-domain trusted systems must, by their nature, focus

285 mandatory and enforceable rules narrowly upon the critical items that must be commonly agreed  
286 in order for that network to achieve it's purpose.

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

296 A system like this has made the interbank money transfer system among the safest systems  
297 in the world and the daily backbone for exchanges of trillions of dollars, but until recently such  
298 systems were only for the 'big guys'. To give individuals a similarly safe method of managing  
299 personal data, the Human Dynamics research group at MIT, in partnership with the Insti-  
300 tute for Data Driven Design, co-founded by John Clippinger and one author (Pentland), have  
301 helped build open Personal Data Store (openPDS) [11]. See <http://openPDS.media.mit.edu>  
302 for project information and <https://github.com/HumanDynamics/openPDS> for the open source  
303 code.

304 The openPDS is a consumer version of a personal cloud trust network that we are now  
305 testing with a variety of industry and government partners. Soon, sharing your personal data  
306 could become as safe and secure as transferring money between banks.

307 The Human Dynamics Lab has applied the system rules approach to development of in-  
308 tegrated business, technical architecture, and rules large scale institutional use of personal  
309 data stores, available as an example under MIT's creative commons license by MIT, at [https:](https://github.com/HumanDynamics/SystemRules)  
310 [//github.com/HumanDynamics/SystemRules](https://github.com/HumanDynamics/SystemRules).

311 The capacity to apply the appropriate methods of enforcement for a trust network depend

312 upon a clear understanding and agreement among parties about the purpose of the trusted  
 313 system and the respective roles or expectations of those connecting as participants. Therefor,  
 314 an anchor is needed to a clear context of a Big Data operational framework and institutional  
 315 controls appropriate for access and confidentiality or privacy. The following section posits the  
 316 trust model and signature traits of such a context, through the lens of the New Deal on Data.

## 317 5 Transitioning End-User Assent Practices (Arek)

318 The way users grant authorizations to their data is not a trivial matter. The flow of personal  
 319 information, such as location data, purchases, health records can be very complex. Every tweet,  
 320 geo-tagged picture, phone call, or purchase with credit card, provide the user's location not only  
 321 to the primary service, but also to all the applications and services that have been authorized  
 322 to access and re-use these data. The authorizations may come from the end-user or be granted  
 323 by the collecting service, based on an umbrella terms of service, allowing the re-use of the data.  
 324 Implementation of such flows was a crucial part of the Web 2.0 revolution, realized with RESTful  
 325 APIs, mashups, and authorization-based access. The way the personal data travel between the  
 326 services has however become arguably too complex for a user to handle and manage.

327 Increasing the amount of data controlled by the user and granularity of this control is mean-  
 328 ingless if it cannot be exercised in an informed way. For many years, the End User License  
 329 Agreements (EULAs), long incomprehensible texts have been accepted blindly by the user,  
 330 trusting they have not agreed to anything that could harm them. The process of granting the  
 331 authorizations cannot be too complex, as it would prevent the user from understanding her deci-  
 332 sions. At the same time, it cannot be too simplistic, as it may not sufficiently convey the weight  
 333 of the privacy-related decisions. It is a challenge in itself, to build the end-user assent systems  
 334 that allow the user to understand and adjust their privacy settings. Complex EULAs do not  
 335 promote the privacy of the users, effectively pushing them to press *I Agree* in every presented  
 336 window.

337 This gap between the interface — single click — and the effect, can render the data owner-

338 ship meaningless; the click may wrench people and their data into systems and rules that are  
 339 antithetical to fair information practices, such as is prevalent with today's end-user licenses in  
 340 cloud services or applications. Managing the potentially long term and opposite dynamics fueled  
 341 by old deal systems operating simultaneously with the new deal systems is an important design  
 342 and migration challenge during the transition to a Big Data economy. During this transition  
 343 and after the New Deal on Data is no longer new, personal data must continue to flow in order  
 344 to be useful. Protecting the data of people outside of the user-controlled domain is very hard  
 345 without a combination of cost effective and useful business practices, legal rules, and technical  
 346 solutions.

347 We envision Living Informed Consent, where the user is entitled to know what data is being  
 348 collected about her by which entities, empowered to understand the implications of data sharing,  
 349 and finally put in charge of the sharing authorizations. We suggest the readers ask themselves a  
 350 question: *Which services know which city I am in today?*. Google? Apple? Twitter? Amazon?  
 351 Facebook? Flickr? This small application we have authorized a few years ago to access our  
 352 Facebook check-ins and forgot since then? This is an example of a fundamental question related  
 353 to user privacy and assent, and yet finding the answer to it may be surprisingly difficult in today's  
 354 ecosystem. We can hope that most of the services treat the data responsibly and according to  
 355 user authorizations. In the complex network of data flows however, it is relatively easy for the  
 356 data to leak to services careless with it or simply malicious [7]. We need to build the solutions  
 357 to help the user to make well thought-through decisions about data sharing.

## 358 **6 Business, Legal, and Technical Dimensions of Big Data Sys-** 359 **tems (Dazza)**

360 When it comes to data intended to be accessible over networks — whether big, personal, or  
 361 otherwise — the traditional container of an institution makes less and less sense. Institutional  
 362 controls apply, by definition by or to some type of institutional entity such as a business, gov-

ernmental, 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 upon 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. **AS: The paragraph above is hard to understand I think.**

Many organizations are structured with clear leadership on business, legal, and technical issues functionally assigned to top level executive roles. Business issues are typically allocated

390 to roles such as CEO, COO or CFO, while leadership on legal issues is commonly assigned to  
 391 roles like general counsel and regulatory compliance and technical leads are often the roles of  
 392 CIO, CTO or CSO. Having top level leadership for each of the business, legal, and technical  
 393 aspects of a trust network is a critical success factor.

## 394 **7 Big Data and Personal Data Institutional Controls (Thomas)**

395 The phrase “institutional controls” refers to safeguards and protections by use of legal, policy,  
 396 governance, and other non-strictly technical, engineering, or mechanical measures. The phrase  
 397 institutional controls in a Big Data context can perhaps best be understood by examining how  
 398 the concept has been applied to other domains. The most prevalent use of institutional controls  
 399 has been in the field of environmental regulatory frameworks.

400 A good example of how this concept supports and reflects the goals and objectives of en-  
 401 vironmental regulation can be found in the policy documents of the Environmental Protection  
 402 Agency (EPA). This following definition is instructive, and is part of the Institutional Control  
 403 Glossary of Terms [39]:

404 “Institutional Controls - Non-engineering measures intended to affect human activi-  
 405 ties in such a way as to prevent or reduce exposure to hazardous substances. They  
 406 are almost always used in conjunction with, or as a supplement to, other measures  
 407 such as waste treatment or containment. There are four categories of institutional  
 408 controls: governmental controls; proprietary controls; enforcement tools; and infor-  
 409 mational devices.”

410 Going deeper, the article by DeMeo and Doar [12] defines institutional controls thusly:

411 “Institutional controls are administrative and legal controls that help minimize the  
 412 potential for human exposure to contamination and/or protect the integrity of the  
 413 physical remedy. They can include recorded restrictive covenants, but land use



414 laws and regulations, deed restrictions, department consent orders, and conservation  
 415 easements are all institutional controls.”

416 In domains of information technology, this approach is most commonly reflected as “enter-  
 417 prise controls” related to security. See, for example, the report [21] stating: “Enterprise mobility  
 418 technologies, especially those designed to retrofit enterprise controls on top of consumer mobile  
 419 devices, are rapidly evolving. This was a message we heard loud and clear in the study.” This  
 420 study and analysis also reveals much about the internal controls needed to accommodate mobile  
 421 device use by employees. In both capacities as employee, consumer, and other roles, the use of  
 422 mobile devices triggers myriad legal, policy, and other implications for institutional controls.

423 In the legal domain, this concept frequently emerges under the moniker “regulatory compli-  
 424 ance” or “legal compliance” anchored in legal and regulatory frameworks such as Health Insur-  
 425 ance Portability and Accountability Act (HIPAA) and Sarbanes-Oxley (SOX). These statutory  
 426 legal frameworks require covered organizations to established integrated sets of governance,  
 427 legal, transactional, security, and other internal controls to avoid violating the rules. The in-  
 428 stitutional controls are accomplished in tight integration with engineering and other measures  
 429 in order to ensure compliance and to control legal and security risk. The use of institutional  
 430 controls of this type are fundamental methods for achieving and maintaining the transition to a  
 431 digital, networked, and Big Data footing for any private company, government agency, or other  
 432 organization.

433 Consider again the analogy of institutional controls in the context of environmental law, and  
 434 how these types of measures can be applied in the Big Data, privacy, and access context to digital  
 435 environments. Given the relatively mature and stable state of environmental regulation, there is  
 436 much to be learned by examining this context of institutional controls. Environmental regulatory  
 437 compliance with waste management cleanup requirements could include institutional controls  
 438 restricting land use on adjacent property. In these situations, it is possible that the remediation  
 439 strategy requires significant use of land outside the property boundaries of the cleanup site.  
 440 In these cases, the regulators and the land owner responsible for the regulated property must

441 find ways to ensure a common approach among multiple owners and across multiple property  
 442 environments. Use of measures such as a clauses on the relevant deeds, an enforceable consent  
 443 order, or regulations and zoning rules are examples of more severe institutional controls that  
 444 can be employed to ensure consistent and effective actions are taken across ownership and real  
 445 property boundaries.

446 See, for example, Florida Department of Environmental Protection (FDEP), Division of  
 447 Waste Management [15] which states that “...RMO III does contemplate contamination beyond  
 448 the Property boundaries, which would require agreement by the adjacent owners to put an RC  
 449 on their properties as well.”

450 The concept of an “institutional control boundary” is especially clarifying and powerful when  
 451 applied to the networked and digital boundaries of an institution. In the context of Florida’s  
 452 environmental regulation frameworks, the phrase is applied to describe the various types of  
 453 combinations risk management levels related to target cleanup standards and extend beyond  
 454 the area of a physical property boundary. Also see a recent University of Florida report on  
 455 Development of Cleanup Target Levels (CTLs) [8] stating “Risk Management Options Level  
 456 III, like Level II, allows concentrations above the default groundwater CTLs to remain on site.  
 457 However, in some rare situations, the institutional control boundary at which default CTLs must  
 458 be met can extend beyond the site property boundary.”

459 The EPA provides considerable information on the nature and use of institutional controls,  
 460 including situations when the situational scope extends to adjacent properties owned by third  
 461 parties. See, generally, *EPA Hazardous Waste Corrective Action Guidance on Institutional Con-*  
 462 *trols* [39]. Also see: *Institutional Controls Bibliography: Institutional Control, Remedy Selection,*  
 463 *and Post-Construction Completion Guidance and Policy, December 2005* [38].

464 When institutional controls would apply to “separately owned neighboring properties” a  
 465 number of issues arise. Engagement with affected third parties, requiring the party responsible  
 466 for site cleanup to use “best efforts” to attain agreement by third parties to institute the relevant  
 467 institutional controls, use of third party neutrals to resolve disagreements regarding the appli-

468 cation with institutional control,s or forcing an acquisition of the neighboring land by forcing  
 469 the party responsible to purchase the property of by purchase of the property directly by the  
 470 EPA [40].

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

484 The use of system rules and integrated participation agreements by developers and end-  
 485 users is a way to ensure intended operational frameworks conform to applicable institutional  
 486 controls. The example of Living Informed Consent described in this chapter, demonstrates how  
 487 institutional controls comprised of legal and definite workflow measures, in concert with technical  
 488 methods, can result in a higher level of performance, while appropriately balancing legitimate  
 489 interests of various parties regarding use and access to personal data.

490 Following the World Economic Forum recommendations of treating personal data stores in  
 491 the manner of bank accounts [42], there are a number of infrastructure improvements that need to  
 492 be realized, if the personal data ecosystem is to flourish and deliver new economic opportunities.  
 493 We believe the following infrastructure improvements are necessary for the coming personal data  
 494 ecosystem: **AS: We should remove the bullets, turn them into continuous text.**

- 495 • *New global data provenance network:* In order for personal data to be treated like bank  
 496 accounts, the origin information regarding data items coming into the data store must be  
 497 maintained [20]. In other words, the provenance of all data items must be accounted for  
 498 by the IT infrastructure upon which the personal data store operates. The heterogeneous  
 499 provenance databases must then be interconnected in order to provide a resilient and  
 500 scalable platform for audit and accounting systems to track and reconcile the movement  
 501 of personal data from the respective data stores.
  
- 502 • *Trust network for computational law:* In order for trust to be established between parties  
 503 who wish to exchange personal data, we foresee that some degree of “computational law”  
 504 technologies may have to be integrated into the design of personal data systems. Such  
 505 technologies should not only verify terms of contracts (e.g. terms of data use) against user-  
 506 defined policies but also have mechanisms built-in to ensure non-repudiation of entities  
 507 who have accepted these digital contracts. Efforts such as [1, 2] are beginning to bring  
 508 non-repudiation and enforceability of contracts into the technical protocol flows.
  
- 509 • *Development of institutional controls for digital institutions:* Currently there are a number  
 510 of proposal for the creation of virtual currencies (e.g. BitCoin [5], Ven [35]) in which the  
 511 systems have the potential to evolve into self-governing “digital institutions” [19]. Such  
 512 systems and institutions that operate on them will necessitate the development of a new  
 513 paradigm to understand the aspects of institutional control within their context.

## 514 8 Scenarios of Use in Context (Dazza)

515 Supporting the effective development of institutional controls for big data requires an under-  
 516 standing of how to define and work with the applicable context surrounding the scenarios within  
 517 which the Big Data exists. In particular, the New Deal on Data will require a set of Institu-  
 518 tional Controls involving governance, business, legal, and technical aspects that are knowable  
 519 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.

## 8.1 Example Scenario: Research Systems

**AS: This entire section requires significant write-through.**

Computational Social Science (CSS) studies are based on data collected often with an extremely high resolution and scale [23]. 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**AS: 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.

One author (Stopczynski) deployed this year a 1,000 phones study at Technical University of Denmark, freshmen students received mobile phones in order to study their networks and social behavior in the important change moment of their lives, when joining the university. The study, called SensibleDTU (<https://www.sensible.dtu.dk/?lang=en>), 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 issue of students feeling obliged to participate in the study, feeling that their

546 grades may depend on it, or that the data may influence their grades. In this context, we see  
547 the implementation of Living Informed Consent not only as a technical mean to put participants  
548 in control of the data we collect, but also to convey the message about the opt-in nature of the  
549 study, the boundaries of the data usage, and parties accessing the data.

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

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

569 When a study of such scale is deployed, the particular experiments and sub-studies may  
570 not be exactly defined from the very beginning. The initial deployment is a creation of a  
571 testbed, where shorter or longer experiments can take place; for example part of the population  
572 may participate in the experiment of quantifying the impact of feedback application on their

573 activity levels. Being able to create such experiments in an efficient way is a huge value for the  
574 researchers. To do that in the most frictionless way, we give the users the choice to opt-in to  
575 those additional experiments, providing some financial or other benefits. This is only possible  
576 if there is a notion of identity of the participants, stronger and more useful than a piece of  
577 paper with a signature. This identity allows us to reach out to people, offer them additional  
578 experiments, and let them agree or disagree to them.

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

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

## 597 8.2 Scenarios of Use Today, Tomorrow and the Day After

598 **AS: This paragraph is impossible to follow for someone without deep background**  
 599 **knowledge of what is the message. Too many random made up scenarios, entities,**  
 600 **all mashed together.**

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

618 **AS: Bold claims here, not sure if we have sufficient support for them in the**  
 619 **chapter.**

620 This book represents a high value opportunity to take stock of the current state and dom-  
 621 inant trends related to Big Data and help to illuminate important choices at a moment of  
 622 early adoption, dynamic innovation, and wide open possibilities. By contemplating the relevant  
 623 contexts of todays scenarios of use in, say, the fields of education, entertainment, government,



624 manufacturing, transportation, and many other core anchors of human activity, we have traction  
625 to postulate how today's prevailing trends are likely to result and what changes - perhaps quite  
626 small but of profound long term impact - could lead to materially different better outcomes.  
627 Consider that if the essence of the New Deal on Data was accepted today, or soon, the na-  
628 ture, tenor, capabilities, and experience of living by future generations could be unrecognizingly  
629 better. Simply extrapolate from the current anomalous practices regarding personal data and  
630 individual identity and push forward the timeline by 5, 10, 20 years and beyond. The current  
631 trajectory ends up with dystopian scenarios that effectively reverse hard fought, but easily lost  
632 constitutional deal of the United States and social compact of common law societies.

633 By contrast, by adopting the New Deal on Data now it is possible to set conditions that  
634 promote prosperity and invention even before the New Deal on Data frameworks are formally  
635 launched. This is because the uncertainty and confusion about the basic premises and expecta-  
636 tions around personal data and identity will be resolved and so investment and risk taking on  
637 a firm foundation can be unleashed. The value of Big Data can be accessed at less direct cost  
638 and lower risk when uncertainties about privacy liability are addressed and significant the new  
639 value is created by enabling wide scale permission based access to personal data and compu-  
640 tations about such data. Adopting use of personal data services in phases, such one economic  
641 sector, transaction type or data type at a time enables access to the lower costs and new value  
642 in a reasonable manner that allows for time to prepare for and stage each phase of adoption.  
643 By staging and phasing the New Deal on Data typical objections to change based on grounds  
644 of cost, disruption or over regulation can be addressed. Policy incentives can further address  
645 these objections, such as allowing safe harbor protections for conduct of organizations operating  
646 under the rules of a trust network. Policy makers can resolve other difficulties by combina-  
647 tions of strategic transition management methods like allowing safe harbor compliance delays,  
648 or approving alternative adoption paths and granting other non-substantive waivers to ease any  
649 burdens of migrating to new business methods. The key point is change management can be  
650 designed to achieve enough value at every phase for every key stakeholder group such that self

651 interests and the broader interests are all aligned with the public good.

## 652 9 Future Research (Brian)

653 Our traditional methods of testing and improving government, organizations, and so on are of  
654 limited use in building a data-driven society. Even the scientific method that we normally use  
655 do not work as well as we might expect, because there are so many potential connections that  
656 our standard statistical tools generate less than useful results.

657 The reason is that with such rich data, you can easily uncover misleading or unactionable  
658 correlations. For instance, let us imagine we discover that people who are unusually active are  
659 more likely to get the flu. This is a real example: when we examined the minute-by-minute  
660 behavior of a small university community - a real-time flow of gigabytes per day for an entire  
661 year - we noticed that an unusual level of running around often predicted onset of the flu [25].  
662 But if we can only analyze the data using traditional statistical methods, we have the problem  
663 of discerning why this is true. Is it because the flu virus makes us more active in order to spread  
664 itself more quickly? While it is more likely that interacting with many more people than usual  
665 makes you more likely to catch the flu, you can't be sure that this is the true cause based on  
666 the real-time stream of data alone.

667 Normal analysis methods do not suffice to answer this type questions, because we do not  
668 know all the possible alternatives, and so we cannot form a limited, testable number of clear  
669 hypotheses. Instead, we need to devise new ways to test the causality of connections in the real  
670 world. We can no longer rely on laboratory experiments; we need to do the experiments in the  
671 real world, typically on massive, real-time streams of data.

### 672 9.1 Research on Design and Deployment of Big Data Systems

673 **AS: I do not understand this paragraph? What is top current research? Where is it**  
674 **applied?** In order to achieve low risk, high value outcomes efficiently, design and deployment  
675 of the coming global wave of Big Data systems should apply top current research. To understand

and address the unique problems and prospects associated with big personal data, the relevant context must be identified and corresponding rules-driven capabilities must be designed into the underlying systems.

People or systems 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 **AS: I think I understand this previous sentences but I' m not sure. What is 'a standard manner' here? What is the right information? It seems it is described in the next sentences, maybe remove this one then?** . 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 of metadata listing or reliably linking the parties, transactions, systems and provenance of the data would suffice. This relevant context together with the data forms the basis for accountable analysis on big personal data.

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.

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) 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).

703 **AS: I feel Big Data term is abused in this section...**

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

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

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

## 728 9.2 Research on Big Data for Design of Institutions

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

734 With the coming of Big Data, we are going to be operating very much out of our old,  
735 familiar ballpark. These data are often indirect and noisy, and so interpretation of the data  
736 requires greater care than usual. Even more importantly, a great deal of the data is about  
737 human behavior, and the questions are ones that seek to connect physical conditions to social  
738 outcomes. Until we have a solid, well-proven, and quantitative theory of social physics, we will  
739 not be able to formulate and test hypotheses in the way we can when we design bridges or  
740 develop new drugs.

741 Therefore, we must move beyond the closed, laboratory-based question-and-answering pro-  
742 cess that we currently use, and begin to manage our society in a new way. We must begin to test  
743 connections in the real world far earlier and more frequently than we have ever had to do before,  
744 using the methods the Human Dynamics research group have developed with our collaborators  
745 for the Friends and Family [3] or the SensibleDTU (<https://www.sensible.dtu.dk>) study. We  
746 need to construct Living Laboratories — communities willing to try a new way of doing things  
747 or, to put it bluntly, to be guinea pigs — in order to test and prove our ideas. This is new  
748 territory and so it is important for us to constantly try out new ideas in the real world in order  
749 to see what works and what does not.

750 An example of such a Living Lab is the ‘open data city just launched by one author (Pentland)  
751 with the city of Trento in Italy, along with Telecom Italia, Telefonica, the research university  
752 Fondazione Bruno Kessler, the Institute for Data Driven Design, and local companies. Import-  
753 tantly, this Living Lab has the approval and informed consent of all its participants they know  
754 that they are part of a gigantic experiment whose goal is to invent a better way of living. More

755 detail on this Living Lab can be found at <http://www.mobileterritoriallab.eu/>.

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

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

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

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

778     The Trento Living Lab will let us investigate how to deal with the sensitivities of collecting  
779 and using deeply personal data in real-world situations. In particular, the Lab will be used as a  
780 pilot for the New Deal on Data and for new ways to give users control of the use of their personal  
781 data. For example, we will explore different techniques and methodologies to protect the users

782 privacy while at the same time being able to use these personal data to generate a useful data  
783 commons. We will also explore different user interfaces for privacy settings, for configuring the  
784 data collected, for the data disclosed to applications and for those shared with other users, all  
785 in the context of a trust framework.

## 786 10 Conclusions

787 Our societies today face unprecedented challenges. Solving those problems will require access  
788 to the personal data, so we can understand how the society works, how we move around, what  
789 makes us productive, how the ideas and diseases spread. The insights must be actionable,  
790 available in real-time, and engaging the population, creating the nervous system of the society.  
791 In this chapter we have reviewed how Big Data collected in institutional context can be used for  
792 the public good. In many cases, the data needed for creating better society is already collected  
793 and exists closed in silos of companies and governments. Using well designed and implemented  
794 set of institutional controls, covering business, legal, and technical dimensions, we described how  
795 the silos can be opened. The framework for doing this — the New Deal on Data — postulates  
796 that the primary driver of the change must be the ownership of the personal data, given to  
797 people about whom the data is. This ownership, the right to use, transfer, and remove the data  
798 ensures that the data is available for public good, while at the same time protecting the privacy  
799 of the citizens.

800 The New Deal on Data is still new. Here we described our efforts in understanding the  
801 technical means of how it can be implemented, the legal framework around it, business ramifi-  
802 cations, and the direct value that can be derived from researchers, companies, governments, and  
803 users having more access to the data. It is clear that companies must play the major role in the  
804 implementation of the New Deal, incentivized by business opportunities and pressured by the  
805 legislation and demand of the users. Only with such orchestration it will be possible to change  
806 the current feudal system of the data ownership and finally put the immense quantities of the  
807 collected personal data to good use.

## References

1. Binding obligations on User-Managed Access (UMA) participants. Technical Specifications draft-maler-oauth-umatrust-01, Kantara Initiative, July 2013.
2. User-Managed Access (UMA) profile of OAuth2.0. Technical Specifications draft-hardjono-oauth-umacore-08, Kantara Initiative, December 2013.
3. Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6):643–659, 2011.
4. Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.
5. Simon Barber, Xavier Boyen, Elaine Shi, and Ersin Uzun. Bitter to Better – how to make Bitcoin a better currency. In *Proceedings Financial Cryptography and Data Security Conference (Lecture Notes in Computer Science Volume 7397)*, pages 399–414, April 2012.
6. Ellen Barry. Protests in moldova explode, with help of twitter. *New York Times*, 8, 2009.
7. Nick Bilton. Girls around me: An app takes creepy to a new level. *The New York Times*.
8. Center for Environmental & Human Toxicology University of Florida. Development of Cleanup Target Levels (CTLs) For Chapter 62-777, F.A.C. Technical report, Division of Waste Management Florida Department of Environmental Protection, February 2005.
9. Paul Lukowicz Bert Arnrich Cornelia Setz Gerhard Troster David Tacconi, Oscar Mayora and Christian Haring. Activity and emotion recognition to support early diagnosis of psychiatric diseases. pages 100–102. IEEE, 2008.
10. Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.



- 831 11. Yves-Alexandre de Montjoye, Samuel S Wang, Alex Pentland, Dinh Tien Tuan Anh, An-  
 832 witaman Datta, Kevin W Hamlen, Lalana Kagal, Murat Kantarcioglu, Vaibhav Khadilkar,  
 833 Kerim Yasin Oktay, et al. On the trusted use of large-scale personal data. *IEEE Data*  
 834 *Eng. Bull.*, 35(4):5–8, 2012.
- 835 12. Ralph A. DeMeo and Sarah Meyer Doar. Restrictive covenants as institutional controls  
 836 for remediated sites: Worth the effort? *The Florida Bar Journal*, 85(2), 2011.
- 837 13. Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Per-*  
 838 *sonal and ubiquitous computing*, 10(4):255–268, 2006.
- 839 14. Jonathan Woetzel et al. Preparing for china’s urban billion. 2009.
- 840 15. Florida Department of Environmental Protection - Division of Waste Management. Insti-  
 841 tutional Controls Procedures Guidance. [http://www.dep.state.fl.us/waste/quick\](http://www.dep.state.fl.us/waste/quick\_topics/publications/wc/csf/icpg.pdf)  
 842 [\\_topics/publications/wc/csf/icpg.pdf](http://www.dep.state.fl.us/waste/quick\_topics/publications/wc/csf/icpg.pdf), June 2012.
- 843 16. Kate Greene. Reality mining. *Technology Review*, 2008.
- 844 17. Lev Grossman. Iran protests: Twitter, the medium of the movement. *Time Magazine*,  
 845 17, 2009.
- 846 18. Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy,  
 847 David Lazer, Alan Mislove, and Christo Wilson. Measuring personalization of web search.  
 848 In *Proceedings of the 22nd international conference on World Wide Web*, pages 527–538.  
 849 International World Wide Web Conferences Steering Committee, 2013.
- 850 19. Thomas Hardjono, Patrick Deegan, and John Clippinger. On the Design of Trustworthy  
 851 Compute Frameworks for Self-Organizing Digital Institutions. In *Proceedings of the 16th*  
 852 *International Conference on Human-Computer Interaction*, 2014.

- 853 20. Thomas Hardjono, Daniel Greenwood, and Alex Pentland. Towards a trustworthy digital  
854 infrastructure for core identities and personal data stores. In *Proceedings of the ID360*  
855 *Conference on Identity*. University of Texas, April 2013.
- 856 21. Juniper Networks. Secure Data Access Anywhere and Anytime: Current Landscape and  
857 Future Outlook of Enterprise Mobile Security. A forrester consulting thought leadership  
858 paper commissioned by att and juniper networks, Forrester Research, October 2012.
- 859 22. Meglena Kuneva. Roundtable on Online Data Collection, Targeting and Profiling . [http:](http://europa.eu/rapid/press-release_SPEECH-09-156_en.htm)  
860 [//europa.eu/rapid/press-release\\_SPEECH-09-156\\_en.htm](http://europa.eu/rapid/press-release_SPEECH-09-156_en.htm), 2009.
- 861 23. David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi,  
862 Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann,  
863 et al. Life in the network: the coming age of computational social science. *Science (New*  
864 *York, NY)*, 323(5915):721, 2009.
- 865 24. Antonio Lima, Manlio De Domenico, Veljko Pejovic, and Mirco Musolesi. Exploiting  
866 cellular data for disease containment and information campaigns strategies in country-  
867 wide epidemics. School of computer science university of birmingham technical report  
868 csr-13-01, University of Birmingham, May 2013.
- 869 25. Anmol Madan, Manuel Cebrian, David Lazer, and Alex Pentland. Social sensing for  
870 epidemiological behavior change. In *Proceedings of the 12th ACM international conference*  
871 *on Ubiquitous computing*, pages 291–300. ACM, 2010.
- 872 26. AC Madrigal. Dark social: We have the whole history of the web wrong. *The Atlantic*,  
873 2013.
- 874 27. Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosen-  
875 quist. Pulse of the nation: Us mood throughout the day inferred from twitter. *Accessed*  
876 *November, 22(2011):2011*, 2010.

- 877 28. Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse  
878 datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125.  
879 IEEE, 2008.
- 880 29. Wei Pan, Yaniv Altshuler, and Alex Sandy Pentland. Decoding social influence and  
881 the wisdom of the crowd in financial trading network. In *Privacy, Security, Risk and*  
882 *Trust (PASSAT), 2012 International Conference on and 2012 International Confernece*  
883 *on Social Computing (SocialCom)*, pages 203–209. IEEE, 2012.
- 884 30. Wei Pan, Gourab Ghoshal, Coco Krumme, Manuel Cebrian, and Alex Pentland. Urban  
885 characteristics attributable to density-driven tie formation. *Nature communications*, 4,  
886 2013.
- 887 31. ALEX PENTLAND. Reality mining of mobile communications: Toward a new deal on  
888 data. *The Global Information Technology Report 2008–2009*, page 1981, 2009.
- 889 32. Alex Pentland, David Lazer, Devon Brewer, and Tracy Heibeck. Using reality mining to  
890 improve public health and medicine. *Stud Health Technol Inform*, 149:93–102, 2009.
- 891 33. Vivek K Singh, Laura Freeman, Bruno Lepri, and Alex Sandy Pentland. Classifying  
892 spending behavior using socio-mobile data. *HUMAN*, 2(2):pp–99, 2013.
- 893 34. Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of  
894 predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
- 895 35. Stan Stalnaker. The Ven currency, 2013. <http://www.ven.vc>.
- 896 36. Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Fran-*  
897 *cisco)*, pages 1–34, 2000.
- 898 37. The White House. National Strategy for Trusted Identities in Cyberspace: Enhancing On-  
899 line Choice, Efficiency, Security, and Privacy. The White House, April 2011. Available on  
900 [http://www.whitehouse.gov/sites/default/files/rss\\_viewer/NSTICstrategy\\_041511.pdf](http://www.whitehouse.gov/sites/default/files/rss_viewer/NSTICstrategy_041511.pdf).

- 901 38. United States Environmental Protection Agency. Institutional Controls Bibliography.  
902 <http://www.epa.gov/superfund/policy/ic/guide/biblio.pdf>, December 2005.
- 903 39. United States Environmental Protection Agency. RCRA Corrective Action Institu-  
904 tional Controls - glossary. [http://www.epa.gov/epawaste/hazard/correctiveaction/](http://www.epa.gov/epawaste/hazard/correctiveaction/resources/guidance/ics/glossary1.pdf)  
905 [resources/guidance/ics/glossary1.pdf](http://www.epa.gov/epawaste/hazard/correctiveaction/resources/guidance/ics/glossary1.pdf), 2007.
- 906 40. United States Environmental Protection Agency. Institutional Controls: A Guide to Plan-  
907 ning, Implementing, Maintaining, and Enforcing Institutional Controls at Contaminated  
908 Sites. Technical Report OSWER 9355.0-89 EPA-540-R-09-001, EPA, December 2012.
- 909 41. Jessica Vitak, Paul Zube, Andrew Smock, Caleb T Carr, Nicole Ellison, and Cliff Lampe.  
910 It's complicated: Facebook users' political participation in the 2008 election. *CyberPsy-*  
911 *chology, behavior, and social networking*, 14(3):107–114, 2011.
- 912 42. World Economic Forum. Personal Data: The Emergence of a New  
913 Asset Class, 2011. Available on [http://www.weforum.org/reports/](http://www.weforum.org/reports/personal-data-emergence-new-asset-class)  
914 [personal-data-emergence-new-asset-class](http://www.weforum.org/reports/personal-data-emergence-new-asset-class).