

# 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</b>	<b>2</b>
10	<b>2 The New Deal on Data</b>	<b>4</b>
11	<b>3 Personal Data: Emergence of a New Asset Class</b>	<b>6</b>
12	<b>4 Enforcing the New Deal on Data</b>	<b>10</b>
13	<b>5 Transitioning End-User Assent Practices</b>	<b>13</b>
14	<b>6 Business, Legal, and Technical Dimensions of Big Data Systems</b>	<b>15</b>
15	<b>7 Big Data and Personal Data Institutional Controls</b>	<b>16</b>
16	<b>8 Scenarios of Use in Context</b>	<b>21</b>
17	8.1 Example Scenario: Research System for Computational Social Science . . . . .	25
18	8.2 Scenarios of Use Today, Tomorrow and the Day After . . . . .	28
19	<b>9 Future Research</b>	<b>29</b>
20	9.1 Research on Design and Deployment of Big Data Systems . . . . .	30
21	9.2 Research on Big Data for Design of Institutions . . . . .	32

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

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 within the framework of the New Deal on Data, and describe  
29 future directions for research and development.

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

37 Sustaining a healthy, safe, and efficient society is a scientific and engineering challenge going  
38 back to the 1800s when the Industrial Revolution spurred rapid urban growth, thereby 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 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 [15].

48 It does not have to be this way. We can have cities that are energy efficient, have secure food  
 49 and water supplies, are protected from pandemics, and enjoy much better governance. To reach  
 50 these 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.

55 Sustainable future society depends on our new technologies being used to create a *nervous*  
 56 *system* maintaining the stability of government, energy, and public health systems around the  
 57 globe. The digital feedback technologies are today capable of creating a level of dynamic re-  
 58 sponsiveness that our larger, more complicated modern society requires. We must reinvent the  
 59 systems of societies within a control framework: sensing the situation, combining these observa-  
 60 tions with models of demand and dynamic reaction, and finally using the resulting predictions  
 61 to tune the system to match the demands.

62 The engine driving this nervous system is Big Data: the newly ubiquitous digital data, now  
 63 available about all aspects of human life. We can analyze patterns of human experience and  
 64 ideas exchange within the *digital breadcrumbs* that we all leave behind as we move through  
 65 the world: call records, credit card transactions, GPS location fixes, among others [25]. By  
 66 recording our choices, these data tell the story of our lives. And this may be very different from  
 67 what we decide to put on Facebook or Twitter; our postings there are what we choose to tell  
 68 people, edited according to the standards of the day and filtered to match the persona we are  
 69 building. Mining social networks can give some great insights about human nature [4, 29, 43];  
 70 who we really are, however, is even more accurately determined by where we spend our time  
 71 and which things we buy, rather than just what we say we do [28].

72 The process of analyzing the patterns within these digital breadcrumbs is called reality  
 73 mining [14, 33], and through it we can learn an enormous amount about who we are. The  
 74 Human Dynamics research group at MIT found that we can use them to tell if we are likely

75 to get diabetes [34], or whether we are the sort of person who will pay back loans [35]. By  
 76 analyzing these patterns across many people, we are discovering that we can begin to explain  
 77 many things — crashes, revolutions, bubbles — that previously appeared to be random acts of  
 78 God [31]. For this reason, the magazine *Technology Review* named our development of reality  
 79 mining as one of the ten technologies that will change the world [18].

## 80 2 The New Deal on Data

81 The digital breadcrumbs we leave behind provide clues about who we are, what we do and what  
 82 we want. This makes personal data — data about individuals — immensely valuable, both for  
 83 public good and for private companies. As European Consumer Commissioner, Meglena Kuneva  
 84 said recently, “Personal data is the new oil of the Internet and the new currency of the digital  
 85 world” [24]. This new ability to see the details of every interaction can be used for good or for  
 86 ill. Therefore, maintaining protection of personal privacy and freedom is critical to our future  
 87 success as a society. We need to enable even more data sharing for the public good; at the same  
 88 time, we need to do a much better job in protecting the privacy of the individuals.

89 A successful data-driven society must be able to guarantee that our data will not be abused;  
 90 perhaps especially that government will not abuse the power conferred by access to such fine-  
 91 grain data. The abuses may be directly targeted at users, for example by offering them higher  
 92 insurance rates based on their shopping history [17], or create problems for the entire society in  
 93 the long run, for example by limiting user choices and closing them into information bubbles [20].  
 94 To achieve the positive possibilities of the new society, we require the *New Deal on Data*, workable  
 95 guarantees that the data needed for public good are readily available while at the same time  
 96 protecting the citizenry [33].

97 The key insight that motivates the idea of the New Deal on Data is that our data are worth  
 98 more when shared, because these aggregated data — averaged, combined across population, and  
 99 often distilled to high-level features — inform improvements in systems such as public health,  
 100 transportation, and government. For instance, we have demonstrated that data about the way

101 we behave and where we go can be used to minimize the spread of infectious disease [27,34]. Our  
102 research has reported how we were able to use these digital breadcrumbs to track the spread of  
103 influenza from person to person on an individual level. And if we can see it, we can stop it.

104 Similarly, if we are worried about global warming, these shared, aggregated data can show us  
105 how patterns of mobility relate to productivity [32]. In turn, this provides us with the ability to  
106 design cities that are more productive and, at the same time, more energy efficient. But in order  
107 to obtain these results and make a greener world, we need to be able to see the people moving  
108 around; this depends on many people willing to contribute their data, even if only anonymously  
109 and in aggregate.

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

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

127 Through the years, the great goal of human societies was to find the efficient ways of gov-

ernance. The Big Data transformation can contribute to this ultimate goal of providing the society with tools to analyze and understand what needs to be done, and to reach the consensus on how to do it. This goes beyond simple creation of more communication platforms; the assumption that more interactions between users will result in better decisions being made, may be very misleading. Although in the recent years we have seen some great examples of using social networks for better organization in society, for example during political protests [6,19], we are not even close to the point where we can start reaching consensus about the big problems: epidemics, climate change, pollution. We can improve the discussions by making them data driven, involving both experts and wisdom of the crowds – users themselves interested in improving the society. The problems we are dealing with as a now global society are more difficult than ever. We are responsible for many of them, and being able to tackle them on a global scale is necessary for our survival as a people.

### 3 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 more new ideas and greater flow ideas (idea liquidity) is to define ownership rights. The only politically viable course is to give individual citizens key rights over data that are about them and in fact, these types of rights have undergirded the European Union’s Privacy Directive since 1995 [13].

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 on 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

153        play a role akin to a bank, managing the data on behalf of their customers.

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

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

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

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

168        In 2007, one author (Pentland) first proposed the New Deal on Data to the World Economic  
 169        Forum [44]. Since then, this idea has run through various discussions and eventually helped  
 170        shape the 2012 Consumer Data Bill of Rights in the United States, along with a matching  
 171        declaration on Personal Data Rights in the EU. These new regulations hope to accomplish the  
 172        combined trick of breaking data out of the current silos, thus enabling the public good, while  
 173        at the same time giving individuals greater control over data about them. But, of course this is  
 174        still a work in progress and the battle for individual control of personal data rages onward.

175        The World Economic Forum (WEF) has dubbed personal data as the “New Oil” or resource  
 176        of the 21st century [44]. The discovery of oil and the subsequent development of the oil industry  
 177        over the past 100 years has spurred not only the development of the automobile industry but also  
 178        the creation of the global transportation infrastructure, including the massive freeway networks

179 that we see today in the developed nations. The “personal data sector” of the economy today is  
180 still in its infancy, its state akin to the oil industry at the late 1890s prior to the development of  
181 the Model-T Ford automobile. The productive collaboration between the Government (building  
182 the state owned freeways), the private sector (mining and refining oil, building automobiles)  
183 and the citizen (the user-base of these services) allowed the developed nations to expand their  
184 economies by creating new markets adjacent to the automobile and oil industries.

185 If personal data, as the new oil, is to reach its global economic potential, there needs to be  
186 a productive collaboration between all the stakeholders in the establishment of a *personal data*  
187 *ecosystem*. As mentioned in [44], a number of fundamental questions about privacy, property,  
188 global governance, human rights — essentially around who should benefit from the products  
189 and services built upon personal data — are major uncertainties shaping the opportunity. The  
190 rapid rate of technological change and commercialization in using personal data is undermining  
191 end user confidence and trust.

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

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



206 Recently, we have shown how challenging, but also feasible, it is to open such institutional  
 207 Big Data. In the Data For Development (D4D) Challenge <http://www.d4d.orange.com/home>,  
 208 the telecom operator Orange opened access to a large dataset of call detail records (CDRs) from  
 209 the Ivory Coast. Working with the data as part of a challenge, teams of researchers came up  
 210 with life-changing insights for the country. For example, one team developed a model for how  
 211 disease spread in the country and demonstrated that information campaigns based on one-to-one  
 212 phone conversations among members of social groups can be an effective countermeasure [26].  
 213 In releasing and analyzing this data, the privacy of the people who generated the data was  
 214 protected not only by technical means, such as removal of Personally Identifiable Information  
 215 (PIIs), but also by legal means, with the researchers signing an agreement they will not use the  
 216 data for re-identification or other nefarious purposes. As we have seen in several cases, such as  
 217 the Netflix Prize privacy disaster [30] and other similar privacy breaches [38], true anonymization  
 218 is extremely hard. In the Unique in the Crowd [10], de Montjoye et al. showed that even though  
 219 human beings are highly predictable [36], we are also very unique. Having access to one dataset  
 220 may be enough to uniquely fingerprint someone based on just a few datapoints, and use this  
 221 fingerprint to discover their true identity. The higher the resolution of the data, the easier it  
 222 gets to identify a person from this type of data.

223 The report of the World Economic Forum [44] also suggest a way forward by recommending  
 224 a number of areas where efforts could be directed:

- 225 • Alignment of key stakeholders: Citizens, the private sector and the public sector need to  
 226 work in support of one another. Efforts such as NSTIC [39] — albeit still in its infancy —  
 227 represent a promising direction for a global collaboration.
- 228 • Viewing “data as money”: There needs to be a new change in mindset where an individual’s  
 229 personal data items are viewed and treated in the same way as their money. These personal  
 230 data items would reside in an “account” (like a bank account) where it would be controlled,  
 231 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 experiences. Efforts such as the *User managed Access* (UMA) initiative [2] point in the right direction by designing systems that are user-centric and managed by the user.

Opening data from the silos by publishing static datasets — collected at some point and unchanging — is important, but it is only the first step. We can do even more substantial things when the data is available in real time and can become part of a society’s nervous system. Epidemics can be monitored and prevented in real time [34], underperforming students can be helped, and people with health risks can be treated before they get sick [9]. The same data can potentially be used for stalking, burglarizing one’s home, and as justification to charge people more for an insurance policy.

## 4 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 upon 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

257 violation of the permissions. For example, in such a system all personal data can have attached  
258 labels specifying what the data can and cannot be used for. These labels are exactly matched  
259 by the network's system rules and terms in legal contracts between all the participants, stating  
260 penalties for not obeying the permission labels. These rules can, and often do, reference or  
261 require audits of relevant systems and data use, demonstrating how traditional internal controls  
262 can be leveraged as part of the transition to more novel trust models.

263 Complete tracking and regulation of every aspect of a trust network is not the goal or  
264 even desirable in order to achieve effective enforcement. Rather, the rules for a trust network  
265 align enforcement with the highest priority issues and those upon which trust of participants is  
266 premised. The relevant issues for a given trust network arise from that systems underlying trust  
267 models and the contextual scenarios within which the networked data and the relationships of  
268 parties occur.

269 When a trust network involves use of personal data, then the user permissions and corre-  
270 sponding limits on use are fundamental to the trust model. In this context, the permissions,  
271 including the provenance of the data, should require appropriate levels of audit. A well designed  
272 trust network, elegantly integrating computer and legal rules, allows automatic auditing of data  
273 use and allows individuals to change their permissions and withdraw data.

274 Having system rules applicable to the networks, applications, and data as well as all the ser-  
275 vices providers other intermediaries, and the users themselves is the mechanism for establishing  
276 and operating a trust network. System rules are sometimes called operating regulations in the  
277 credit card context, or known as trust frameworks in the identity federations context, or trading  
278 partner agreements in a supply value chain context. There are many general examples of multi-  
279 party shared architectural and contractual rules that share the generic characteristic of creating  
280 binding obligations and enforceable expectations on all participants in scalable networks. An-  
281 other common characteristic of the system rules design pattern is that the participants in the  
282 network can be widely distributed across very heterogeneous business ownership boundaries,  
283 legal governance structures, and technical security domains. Yet, the parties need not agree

284 to conform to all or most aspects of their basic roles, relationships, and activities in order to  
285 connect to systems of a trust network. Cross-domain trusted systems must, by their nature,  
286 focus mandatory and enforceable rules narrowly upon the critical items that must be commonly  
287 agreed in order for that network to achieve its purpose.

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

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

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

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

311 `//github.com/HumanDynamics/SystemRules.`

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

## 318 **5 Transitioning End-User Assent Practices**

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

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

337 promote the privacy of the users, effectively pushing them to press *I Agree* in every presented  
338 window.

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

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

## 360 6 Business, Legal, and Technical Dimensions of Big Data Sys- 361 tems

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

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

382 Defining as a ‘system’ the thing to which institutional controls apply provides an achievable  
383 and measurable basis for balancing privacy, access and other interests in Big Data. Within a  
384 given institution, there may in fact be many different discernable organizations and correspond-  
385 ing systems. Meanwhile the system of one institution frequently exists across many different

external institutions. The application of Big Data institutional controls can be applied across the board to a unit of a given institution or targeted by agreement to certain types of data or particular transactions spanning many institutions. Once a systems view is adopted, there is a tractable starting point to narrow or broaden the scope of view, to focus on material dimensions of a system and therefore enable more effective use and control of Big Data.

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

## 7 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 understood by examining how the concept has been applied to other domains. The most prevalent use of institutional controls 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 Environmental Protection Agency (EPA). This following definition is instructive, and is part of the Institutional Control Glossary of Terms [41]:

“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



411 controls: governmental controls; proprietary controls; enforcement tools; and infor-  
412 mational devices.”

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

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

419 In domains of information technology, this approach is most commonly reflected as “enter-  
420 prise controls” related to security. See, for example, the Juniper Networks enterprise security  
421 report [23] stating: “Enterprise mobility technologies, especially those designed to retrofit en-  
422 terprise controls on top of consumer mobile devices, are rapidly evolving. This was a message  
423 we heard loud and clear in the study.” This study and analysis also reveals much about the  
424 internal controls needed to accommodate mobile device use by employees. In both capacities as  
425 employee, consumer, and other roles, the use of mobile devices triggers myriad legal, policy, and  
426 other implications for institutional controls.

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

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

450 See, for example, Florida Department of Environmental Protection (FDEP), Division of  
451 Waste Management [16] which states that “...RMO III does contemplate contamination beyond  
452 the Property boundaries, which would require agreement by the adjacent owners to put an RC  
453 on their properties as well.”

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

463 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* [41]. Also see: *Institutional Controls Bibliography: Institutional Control, Remedy Selection, and Post-Construction Completion Guidance and Policy, December 2005* [40].

When institutional controls would apply to “separately owned neighboring properties” a number of issues arise that are very relevant to the problems associated with managing personal and big data across legal, business and other systemic boundaries. Requiring the party responsible for site cleanup to use “best efforts” to attain agreement by third parties to institute the relevant institutional controls is perhaps the most direct and least prescriptive approach. When direct negotiated agreement is not successful, then use of third party neutrals to resolve disagreements regarding institutional controls can be required. If necessary, environmental regulation can force an acquisition of neighboring land by compelling the party responsible to purchase the other property or by purchase of the property directly by the EPA [42].

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 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 owners 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-

491 users is a way to ensure intended operational frameworks conform to applicable institutional  
 492 controls. The example of Living Informed Consent described in this chapter, demonstrates how  
 493 institutional controls comprised of legal and definite workflow measures, in concert with technical  
 494 methods, can result in a higher level of performance, while appropriately balancing legitimate  
 495 interests of various parties regarding use and access to personal data.

496 Following the World Economic Forum recommendations of treating personal data stores in  
 497 the manner of bank accounts [44], there are a number of infrastructure improvements that need to  
 498 be realized, if the personal data ecosystem is to flourish and deliver new economic opportunities.  
 499 We believe the following infrastructure improvements are necessary for the coming personal data  
 500 ecosystem:

- 501 • *New global data provenance network*: In order for personal data to be treated like bank  
 502 accounts, the origin information regarding data items coming into the data store must be  
 503 maintained [22]. In other words, the provenance of all data items must be accounted for  
 504 by the IT infrastructure upon which the personal data store operates. The heterogeneous  
 505 provenance databases must then be interconnected in order to provide a resilient and  
 506 scalable platform for audit and accounting systems to track and reconcile the movement  
 507 of personal data from the respective data stores.
- 508 • *Trust network for computational law*: In order for trust to be established between parties  
 509 who wish to exchange personal data, we foresee that some degree of “computational law”  
 510 technologies may have to be integrated into the design of personal data systems. Such  
 511 technologies should not only verify terms of contracts (e.g. terms of data use) against user-  
 512 defined policies but also have mechanisms built-in to ensure non-repudiation of entities who  
 513 have accepted these digital contracts. Efforts such as [1,2] are beginning to bring better  
 514 evidentiary proof and enforceability of contracts into the technical protocol flows.
- 515 • *Development of institutional controls for digital institutions*: Currently there are a number  
 516 of proposals for the creation of virtual currencies (e.g. BitCoin [5], Ven [37]) in which the

517 systems have the potential to evolve into self-governing “digital institutions” [21]. Such  
 518 systems and institutions that operate on them will necessitate the development of a new  
 519 paradigm to understand the aspects of institutional control within their context.

## 520 8 Scenarios of Use in Context

521 Development of frameworks for Big Data that effectively balance economic, legal, security and  
 522 other interests requires an understanding of the relevant context and applicable scenarios within  
 523 which the Big Data exists. Although Big Data straddles multiple business, legal and technical  
 524 boundaries it will nonetheless have one or more institutions that are capable of, or in some  
 525 situations required to, manage and control it. The public good referred to in the title of this  
 526 book can be articulated through the use of system, service and software modeling, requirements  
 527 setting, development, testing and certification processes. Discrete use cases of actors and actions  
 528 is one approach to model business, legal and technical requirements in a way that can objectively  
 529 be agreed in advance and traceably be tested against implemented systems and components.  
 530 However, user cases are typically atomic or very low level of granularity and operate deep within  
 531 layers of assumed context. Higher level contexts and corresponding scenarios of multiple use  
 532 cases can describe fundamental expectations about matters like interests in property, rights to  
 533 liberty and honoring the social compact.

534 Institutional controls and other system requirements or safeguards are important methods  
 535 to ensure context-appropriate outcomes consistent with clearly applicable system scenarios that  
 536 set the contours and underpinnings for a greater public good. The New Deal on Data can  
 537 be achieved in part by sets of institutional controls involving governance, business, legal, and  
 538 technical aspects of Big Data and interoperating systems. The following scenarios demonstrate  
 539 signature features of the New Deal on Data in various contexts and serve as an anchor to evaluate  
 540 what institutional controls are well aligned.

541 The basic common law inspired ownership tenants of the New Deal on Data are general prin-  
 542 ciples that guide and inform basic relationships and expectations. However, the dynamic bundle

of recombinant rights and responsibilities constituting "ownership" interests in personal data and expectations pertaining to Big Data vary significantly from context to context and even from one scenario to another within a given general context. The applicable scenario within which the data exists can provide a method and mechanisms of sorts to establish the basic ownership, control and other expectations of the key parties. For example, it may not be sufficient to describe the exchange of money and financial information because the nature of the transaction and their respective data and systems are not identified enough to predict the rights and obligations or other outcomes reasonably expected by individuals and organizations that engage in the activity of a financial exchange. The sale of used cars via an app, the conduct of a counseling session via Google Hangout and the earning of a masters degree via an online university all represent scenarios wherein the use case of a financial exchange takes place. However, each of these scenarios occurs in contexts that are easily identifiable, involving the sale of goods and deeper access to financial information if the car is financed, or involving the practice of therapy by a licensed professional involving confidential mental health data or involving elearning services and protected educational records and possibly deeper financial information if the program is funded by scholarship or loans. Identifying the people (a consumer and a used car dealer) the transaction (purchase of a used car) the data (sales and title data, finance information, etc) and the systems (the third party app and its relevant services or functions, state DMV services, credit card and bank services, etc) provide enough context to establish generally what existing consumer rights under the relevant state lemon laws, the Uniform Commercial Code and other applicable rules will govern when duties arise or are terminated, what must be promised, what can be repudiated, by whom data must be kept secure and other requirements or constraints on the use of personal data and Big Data. These and other factors vary when a transaction that is otherwise identical seeming operates within different scenarios, and even scenarios will differ depending upon which contexts apply.

Which scenarios are relevant and what lower level use cases apply are knowable in detail only with reference to the relevant context of a factually based situation. Relevant scenario of

570 use are comprised of people conducting transactions through systems in which personal data  
571 and Big Data exists or flows. It is possible to test whether frameworks for engagement success-  
572 fully address Big Data, privacy and the public good by testing outcomes of relevant scenarios.  
573 Scenarios are capable of adequately defining these high level goals and objectives when they  
574 identify each of the following four elements:

- 575 1. Who are the people in the scenario (eg who are the parties involved and what are their  
576 respective roles and relationships)?
- 577 2. What are the relevant interactions (eg: what transactions or other actions are conducted  
578 by or with the people involved)?
- 579 3. What are the relevant data and data sets (eg: what types of data are created, stored,  
580 computed, transmitted, modified or deleted)?
- 581 4. What are the relevant systems (eg: what services or other software is used by the people,  
582 for the transactions or with the data)?

583 Retail marketing is a common context within which personal data is important. Personal  
584 data is critical to many different scenarios in the context of retail marketing. Consider the  
585 scenario whereby a merchant conducts an online promotion for an app or service by using a  
586 purchased direct marketing database of consumers who have expressed interest in similar prod-  
587 ucts. Data such as the names, email addresses, phone numbers and other personal information  
588 can be used to lower costs and increase revenue by better targeting promotional messages and  
589 increasing sales. However, there are risks to the merchant and consumer alike, including the  
590 potential of a data breach and resulting identity theft and fraud. There is also risk that some  
591 consumers will feel annoyed or violated when their personal information is used in this man-  
592 ner without their prior knowledge or consent. The information available from such third party  
593 marketing lists and databases may be out of date and lead to the waste of marketing dollars and  
594 the failure to inform potentially interested consumers of a product they might have purchased if  
595 the solicitation had gone to their current email or appropriate network. Imagine that the same  
596 consumers had individual personal data stores and were able to "intent-cast" their interest in

597 the product. This can be done without revealing all the other personal data of that person. The  
598 The openPDS system could be configured to provide permission based answers to questions such  
599 as whether the consumer is over the age of 18 or lives in a city, suburb or rural area. Sectors  
600 such as real estate could be transformed by such intent-casting by qualified buyers.

601 Another common context involving personal data is governmental transactions with the  
602 public. Government filings, registrations, permits and other such public sector transactions with  
603 the individuals or organizations create a large volume and variety of personal data flow. Consider  
604 the scenario whereby a person runs a small business and must comply with tax, employee  
605 related, licensing and other rules by filing forms with multiple government agencies at the federal,  
606 state and local levels. Individuals names, addresses, occupations, dates of birth, social security  
607 numbers and many other types of personal information are common elements of such filings.  
608 Similarly to the retail marketing scenario above, the parties to government filing transactions  
609 also risk unauthorized access to the personal data by interception during transmission or by  
610 breach of data storage systems. In addition, the costs associated with requiring the same data  
611 by many different agencies and updating or correcting data are born by both the filer and the  
612 regulator. What if the people who own or operate such businesses had access to the services  
613 and functions of a personal data store for themselves individually and also for the corporate  
614 entity they operated? Routine changes in status, such as a change of address or name, could  
615 be accomplished in a secure manner once via their own data service and leveraged again and  
616 again by the many faces of government requiring that data. When the authoritative source  
617 of such information can be deemed to be housed within or logically connected to a person's  
618 data store, then the laborious task of address verification and tedious forms and other processes  
619 required by each government entity could be avoided. The saving of direct and indirect costs,  
620 the regaining of time spent by each agency and business and avoidance of delays and uncertainty  
621 are of significant value to all parties (See: <http://kansasbusinesscenter.com> and see the data  
622 files at <https://github.com/kansasbusinesscenter>)

623 The scenario below describes deeper fact-based situations and circumstances in the context



of social science research and studies involving personal data and Big Data. Note how the roles of people, their interactions, the use of data and the design of the corresponding systems reflect and support the New Deal on Data in ways that deliberately provide immediate and increasing value to the stakeholders than is typical or expected typically.

## 8.1 Example Scenario: Research System for Computational Social Science

Computational Social Science (CSS) studies are based on data collected often with an extremely high resolution and scale [25]. 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 in these contexts exist (See: [http://www2.imm.dtu.dk/pubdb/views/publication\\_details.php?id=6632](http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6632)), 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

650 in economic games and so on. As the data is collected in the context of the university, there is  
651 potentially a big issue of students feeling obliged to participate in the study, feeling that their  
652 grades may depend on it, or that the data may influence their grades. In this context, we see the  
653 implementation of Living Informed Consent not only as a technical mean to put participants in  
654 control of the data we collect, but also to clearly and comprehensibly convey broader New Deal  
655 on Data principles such as the opt-in nature of the study, the boundaries of the data usage, and  
656 parties accessing the data.

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

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

676 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 reuse participants data from a previous study based on their authorization. When the data are owned by 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.

## 704 8.2 Scenarios of Use Today, Tomorrow and the Day After

705 The New Deal on Data is designed to provide good value to all stakeholders creating, using or  
706 benefiting from personal data, but the entire vision need not be adopted before value starts to  
707 flow. The social science research study scenario (below) demonstrates how researchers and study  
708 participants alike derive value from New Deal on Data principles today. As more researchers  
709 and students use the types of systems described above, the value is predicted to increase based  
710 upon a network effect. The same dynamic is expected in other contexts as well.

711 Adopting New Deal on Data principles on a large scale can be accomplished iteratively, such  
712 one economic sector, transaction type or data type at a time. A reasonable success metric for  
713 adoption of large scale visions such as the New Deal on Data is whether change management  
714 has been designed to achieve enough value at every phase for every key stakeholder group to  
715 make the change worth the effort. Value to all parties participating in the New Deal on Data  
716 increases as direct or indirect use and re-use of personal data is available in greater volumes and  
717 varieties. Such volume and variety of personal data increases as more parties and transaction  
718 types and data sets and systems adopt and interoperate within the New Deal on Data.

719 By staging and phasing adoption of the New Deal on Data typical objections to change based  
720 on grounds of cost, disruption or over regulation can be addressed. Policy incentives can further  
721 address these objections, such as allowing safe harbor protections for conduct of organizations  
722 operating under the rules of a trust network. Policy makers can resolve other difficulties by  
723 combinations of strategic transition management methods like allowing safe harbor compliance  
724 delays, or approving alternative adoption paths and granting other non-substantive waivers to  
725 ease any burdens of migrating to new business methods.

726 Developing relevant context and scenarios defines a clear anchor for measuring whether a  
727 given use of Big Data and personal data is consistent with measurable criteria. Such criteria  
728 can be used to establish compliance with the rules of a Trust Network and for certification by  
729 government for the right to safe harbor or other protections. Criteria applicable to business,  
730 legal and technical aspects of a system or set of systems can be assessed, evaluated and trace-

ably proven. Such criteria can provide a basic lowest common denominator requirements and constraints for work flow, transaction flow, data flow and service flow within the relevant contexts and scenarios of use. The New Deal on Data provides a clear basis routed in common law and broad understandings of the social compact. Therefore, with the New Deal on Data the appropriate bundle of rights and expectations intended to cover privacy and other personal data interests in Big Data can be explicitly enumerated, debated and eventually agreed in ways that fit relevant contexts.

## 9 Future Research

Our traditional methods of testing and improving government, organizations, and so on are of limited use in building a data-driven society. With Big Data, there are so many potential connections that our standard statistical tools generate less than useful results.

The reason is that with such rich data, you can easily uncover misleading or unactionable correlations. For instance, let us 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 [27]. But if we can only analyze the data using traditional statistical methods, we have the problem of discerning why this is true. Is it because the flu virus makes us more active in order to spread itself more quickly? While it is more likely that interacting with many more people than usual makes you more likely to catch the flu, you can't be sure that this is the true cause based on the real-time stream of data alone.

Normal analysis methods do not suffice to answer these types questions, because we do not know all the possible alternatives, and so we cannot 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 do the experiments in the real world, typically on massive, real-time streams of data.

## 757 9.1 Research on Design and Deployment of Big Data Systems

758 In order to achieve low risk, high value outcomes efficiently, design and deployment of the coming  
759 global wave of Big Data systems should apply relevant research, such as that identified in this  
760 chapter and the book generally. To understand and address the unique problems and prospects  
761 associated with big personal data, the relevant context must be identified and corresponding  
762 rules-driven capabilities must be designed into the underlying systems.

763 Any system that can make, use, receive, or share Big Data must be capable of associat-  
764 ing provenance and purpose for all data in a common and actionable manner. Requiring a  
765 unstructured volumes of narrative documentation and background about the nuances and cir-  
766 cumstances surrounding every data set is both impractical and counterproductive. By contrast,  
767 a small amount of metadata listing or reliably linking the parties, transactions, systems and  
768 provenance of the data would suffice. This relevant context together with the data forms the  
769 basis for accountable analysis on big personal data. People or systems can determine the appro-  
770 priate rules to apply to data when the relevant information is reliably attached to or logically  
771 associated with that data in a standard manner

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

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

784 clinical medicine, education and lifelong learning, social entertainment).

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

791 The four contextual facets of people, interactions, data and systems provide a sound under-  
792 pinning for the design of new Big Data and Web 2.0 systems. The existing systems design and  
793 development processes of establishing business cases, use cases, agile stories, functional require-  
794 ments, etc. do not reliably identify the factors most relevant to use of Big Data, especially in a  
795 Web 2.0 massively distributed environment. The four facets can also be used to analyze appro-  
796 priate, required or prohibited uses for existing Big Data systems. However, it can be difficult  
797 to extract the relevant information from or apply any effective control on systems used for Big  
798 Data but designed to achieve limited purposes in hierarchical closed environments.

799 Big Data, by its nature, represents a new set of business, legal, and technical capabilities and  
800 requirements. Most of the world's systems today are not capable of ingesting, storing, using, or  
801 dynamically flowing Big Data with other systems. Considering that a) Big Data is of high value  
802 immediately and higher value in the short and long terms, and b) the young but competitive  
803 marketplace of Big Data system components, platforms, applications, and other solutions is a  
804 hotbed of innovation it can be predicted that a transition to Big Data systems will continue.  
805 The key observation is that virtually all Big Data systems have yet to be designed, implemented,  
806 customized, or deployed. Institutions that are the current early adopters of today's Big Data  
807 system will soon replace those systems and the rest of the world will adopt Big Data systems in  
808 phases over time. Based upon this observation, it follows that design improvements made now  
809 or soon will have much greater impact than can be had after mass-scale adoption has occurred.

## 810 9.2 Research on Big Data for Design of Institutions

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

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

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

832 An example of such a Living Lab is the ‘open data city’ just launched by one author (Pent-  
 833 land) with the city of Trento in Italy, along with Telecom Italia, Telefonica, the research uni-  
 834 versity Fondazione Bruno Kessler, the Institute for Data Driven Design, and local companies.  
 835 Importantly, this Living Lab has the approval and informed consent of all its participants. Not  
 836 only do these participants consent to sharing of their data, they know that they are part of a



837 gigantic experiment whose goal is to invent a better way of living. This can be a model followed  
838 by many types of systems within and beyond the social science research contexts. More detail  
839 on this Living Lab can be found at <http://www.mobileterritoriallab.eu/>.

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

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

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

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

862 The Trento Living Lab will let us investigate how to deal with the sensitivities of collecting  
863 and using deeply personal data in real-world situations. In particular, the Lab will be used as a

864 pilot for the New Deal on Data and for new ways to give users control of the use of their personal  
 865 data. For example, we will explore different techniques and methodologies to protect the users  
 866 privacy while at the same time being able to use these personal data to generate a useful data  
 867 commons. We will also explore different user interfaces for privacy settings, for configuring the  
 868 data collected, for the data disclosed to applications and for those shared with other users, all  
 869 in the context of a trust framework.

## 870 10 Conclusions

871 Our societies today face unprecedented challenges. Solving these problems will require access  
 872 to personal data, so we can understand how the society works, how we move around, what  
 873 makes us productive, and how everything from ideas to diseases spread. The insights must be  
 874 actionable, available in real-time, and engaging the population, creating the nervous system of  
 875 the society. In this chapter we have reviewed how Big Data collected in institutional context  
 876 can be used for the public good. In many cases, the data needed for creating better society is  
 877 already collected and exists closed in silos of companies and governments. Using well designed  
 878 and implemented sets of institutional controls, covering business, legal, and technical dimensions,  
 879 we described how the silos can be opened. The framework for doing this — the New Deal on  
 880 Data — postulates that the primary driver of the change must be by recognizing ownership of  
 881 personal data rests with the people about whom that data is about. This ownership, the right  
 882 to use, transfer, and remove the data ensures that the data is available for public good, while  
 883 at the same time protecting the privacy of the citizens.

884 The New Deal on Data is still new. Here we described our efforts in understanding the  
 885 technical means of how it can be implemented, the legal framework around it, business rami-  
 886 fications, and the direct value that can be derived from researchers, companies, governments,  
 887 and users having more access to the data. It is clear that companies must play the major role  
 888 in the implementation of the New Deal, incentivized by business opportunities and pressured  
 889 by the legislation and demand of the users. Only with such orchestration will it be possible to

change the current feudal system of data ownership and finally put the immense quantities and capabilities of 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*, 2012.
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.

- 911 9. Paul Lukowicz Bert Arnrich Cornelia Setz Gerhard Troster David Tacconi, Oscar Mayora  
912 and Christian Haring. Activity and emotion recognition to support early diagnosis of  
913 psychiatric diseases. pages 100–102. IEEE, 2008.
- 914 10. Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel.  
915 Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.
- 916 11. Yves-Alexandre de Montjoye, Samuel S Wang, Alex Pentland, Dinh Tien Tuan Anh, An-  
917 witaman Datta, Kevin W Hamlen, Lalana Kagal, Murat Kantarcioglu, Vaibhav Khadilkar,  
918 Kerim Yasin Oktay, et al. On the trusted use of large-scale personal data. *IEEE Data*  
919 *Eng. Bull.*, 35(4):5–8, 2012.
- 920 12. Ralph A. DeMeo and Sarah Meyer Doar. Restrictive covenants as institutional controls  
921 for remediated sites: Worth the effort? *The Florida Bar Journal*, 85(2), 2011.
- 922 13. EU Directive. 95/46/ec of the european parliament and of the council of 24 october 1995  
923 on the protection of individuals with regard to the processing of personal data and on the  
924 free movement of such data. *Official Journal of the EC*, 23:6, 1995.
- 925 14. Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Per-*  
926 *sonal and ubiquitous computing*, 10(4):255–268, 2006.
- 927 15. Jonathan Woetzel et al. Preparing for china’s urban billion. 2009.
- 928 16. Florida Department of Environmental Protection - Division of Waste Management. Insti-  
929 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)  
930 [\\_topics/publications/wc/csf/icpg.pdf](http://www.dep.state.fl.us/waste/quick\_topics/publications/wc/csf/icpg.pdf), June 2012.
- 931 17. Kim Gittleson. How big data is changing the cost of insurance. *BBC News*, 2013.
- 932 18. Kate Greene. Reality mining. *Technology Review*, 2008.
- 933 19. Lev Grossman. Iran protests: Twitter, the medium of the movement. *Time Magazine*,  
934 17, 2009.

- 935 20. Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy,  
936 David Lazer, Alan Mislove, and Christo Wilson. Measuring personalization of web search.  
937 In *Proceedings of the 22nd international conference on World Wide Web*, pages 527–538.  
938 International World Wide Web Conferences Steering Committee, 2013.
- 939 21. Thomas Hardjono, Patrick Deegan, and John Clippinger. On the Design of Trustworthy  
940 Compute Frameworks for Self-Organizing Digital Institutions. In *Proceedings of the 16th*  
941 *International Conference on Human-Computer Interaction*, 2014.
- 942 22. Thomas Hardjono, Daniel Greenwood, and Alex Pentland. Towards a trustworthy digital  
943 infrastructure for core identities and personal data stores. In *Proceedings of the ID360*  
944 *Conference on Identity*. University of Texas, April 2013.
- 945 23. Juniper Networks. Secure Data Access Anywhere and Anytime: Current Landscape and  
946 Future Outlook of Enterprise Mobile Security. A forrester consulting thought leadership  
947 paper commissioned by att and juniper networks, Forrester Research, October 2012.
- 948 24. Meglena Kuneva. Roundtable on Online Data Collection, Targeting and Profiling . [http:](http://europa.eu/rapid/press-release_SPEECH-09-156_en.htm)  
949 [//europa.eu/rapid/press-release\\_SPEECH-09-156\\_en.htm](http://europa.eu/rapid/press-release_SPEECH-09-156_en.htm), 2009.
- 950 25. David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi,  
951 Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann,  
952 et al. Life in the network: the coming age of computational social science. *Science (New*  
953 *York, NY)*, 323(5915):721, 2009.
- 954 26. Antonio Lima, Manlio De Domenico, Veljko Pejovic, and Mirco Musolesi. Exploiting  
955 cellular data for disease containment and information campaigns strategies in country-  
956 wide epidemics. School of computer science university of birmingham technical report  
957 csr-13-01, University of Birmingham, May 2013.

- 958 27. Anmol Madan, Manuel Cebrian, David Lazer, and Alex Pentland. Social sensing for  
959 epidemiological behavior change. In *Proceedings of the 12th ACM international conference*  
960 *on Ubiquitous computing*, pages 291–300. ACM, 2010.
- 961 28. AC Madrigal. Dark social: We have the whole history of the web wrong. *The Atlantic*,  
962 2013.
- 963 29. Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosen-  
964 quist. Pulse of the nation: Us mood throughout the day inferred from twitter. *Accessed*  
965 *November, 22(2011):2011*, 2010.
- 966 30. Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse  
967 datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125.  
968 IEEE, 2008.
- 969 31. Wei Pan, Yaniv Altshuler, and Alex Sandy Pentland. Decoding social influence and  
970 the wisdom of the crowd in financial trading network. In *Privacy, Security, Risk and*  
971 *Trust (PASSAT), 2012 International Conference on and 2012 International Confernece*  
972 *on Social Computing (SocialCom)*, pages 203–209. IEEE, 2012.
- 973 32. Wei Pan, Gourab Ghoshal, Coco Krumme, Manuel Cebrian, and Alex Pentland. Urban  
974 characteristics attributable to density-driven tie formation. *Nature communications*, 4,  
975 2013.
- 976 33. ALEX PENTLAND. Reality mining of mobile communications: Toward a new deal on  
977 data. *The Global Information Technology Report 2008–2009*, page 1981, 2009.
- 978 34. Alex Pentland, David Lazer, Devon Brewer, and Tracy Heibeck. Using reality mining to  
979 improve public health and medicine. *Stud Health Technol Inform*, 149:93–102, 2009.
- 980 35. Vivek K Singh, Laura Freeman, Bruno Lepri, and Alex Sandy Pentland. Classifying  
981 spending behavior using socio-mobile data. *HUMAN*, 2(2):pp–99, 2013.

- 982 36. Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of  
983 predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
- 984 37. Stan Stalnaker. The Ven currency, 2013. <http://www.ven.vc>.
- 985 38. Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Fran-*  
986 *cisco)*, pages 1–34, 2000.
- 987 39. The White House. National Strategy for Trusted Identities in Cyberspace: Enhancing On-  
988 line Choice, Efficiency, Security, and Privacy. The White House, April 2011. Available on  
989 [http://www.whitehouse.gov/sites/default/files/rss\\_viewer/NSTICstrategy\\_041511.pdf](http://www.whitehouse.gov/sites/default/files/rss_viewer/NSTICstrategy_041511.pdf).
- 990 40. United States Environmental Protection Agency. Institutional Controls Bibliography.  
991 <http://www.epa.gov/superfund/policy/ic/guide/biblio.pdf>, December 2005.
- 992 41. United States Environmental Protection Agency. RCRA Corrective Action Institu-  
993 tional Controls - glossary. [http://www.epa.gov/epawaste/hazard/correctiveaction/](http://www.epa.gov/epawaste/hazard/correctiveaction/resources/guidance/ics/glossary1.pdf)  
994 [resources/guidance/ics/glossary1.pdf](http://www.epa.gov/epawaste/hazard/correctiveaction/resources/guidance/ics/glossary1.pdf), 2007.
- 995 42. United States Environmental Protection Agency. Institutional Controls: A Guide to Plan-  
996 ning, Implementing, Maintaining, and Enforcing Institutional Controls at Contaminated  
997 Sites. Technical Report OSWER 9355.0-89 EPA-540-R-09-001, EPA, December 2012.
- 998 43. Jessica Vitak, Paul Zube, Andrew Smock, Caleb T Carr, Nicole Ellison, and Cliff Lampe.  
999 It’s complicated: Facebook users’ political participation in the 2008 election. *CyberPsy-*  
1000 *chology, behavior, and social networking*, 14(3):107–114, 2011.
- 1001 44. World Economic Forum. Personal Data: The Emergence of a New  
1002 Asset Class, 2011. Available on [http://www.weforum.org/reports/](http://www.weforum.org/reports/personal-data-emergence-new-asset-class)  
1003 [personal-data-emergence-new-asset-class](http://www.weforum.org/reports/personal-data-emergence-new-asset-class).