

Operational Framework: Institutional Controls

Daniel "Dazza" Greenwood^{1,*}, Arkadiusz Stopczynski^{1,2}, Brian Sweatt¹, Thomas Hardjono¹,

Alex Sandy Pentland¹

1 MIT

2 DTU

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

Contents

1	Introduction and Overview (Arek)	2
2	The New Realities of Living in a Big Data Society (Arek)	2
3	The New Deal on Data (Arek)	4
4	Personal Data: Emergence of a New Asset Class (Thomas)	6
5	Enforcing the New Deal on Data (Dazza)	9
6	Essential Elements of the New Deal of Data (Brian)	12
7	Transitioning End-User Assent Practices (Arek)	16
8	Business, Legal and Technical Dimensions of Big Data Systems (Dazza)	18
9	Big Data and Personal Data Institutional Controls (Thomas)	19
10	Scenarios of Use in Context (Dazza)	23
10.1	Example Scenario: Research Systems	24
10.2	Scenarios of Use Today, Tomorrow and the Day After	27
11	Future Research (Brian)	28
11.1	Research on Design and Deployment of Big Data Systems	29

22	11.2 Research on Big Data for Design of Institutions	31
----	--	----

23 **1 Introduction and Overview (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. This
 27 chapter describes how the common good can be served by framing these types of institutional
 28 rules and processes to ensure a greater user control over personal data, as well as large scale risk
 29 management and interoperability for data sharing between and among institutions.

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 op-
 32 erational frameworks for Big Data. An operational framework used for a Big Data-driven or-
 33 ganization requires a balanced set of institutional controls. These institutional controls must
 34 support and reflect greater user control over personal data and large scale interoperability for
 35 data sharing between and among institutions. Core capabilities of these controls include re-
 36 sponsive rule-based systems governance and fine-grained authorizations for distributed rights
 37 management. In the following sections we explore the emergence of the Big Data Society, out-
 38 line the ways to support it in the institutional context, and draft the future directions of research
 39 and development.

40 **2 The New Realities of Living in a Big Data Society (Arek)**

41 Sustaining a healthy, safe, and efficient society is a scientific and engineering challenge going
 42 back to the 1800s, when the Industrial Revolution spurred rapid urban growth, creating huge
 43 social and environmental problems. The remedy then was to build centralized networks that
 44 delivered clean water and safe food, enabled commerce, removed waste, provided energy, fa-
 45 cilitated transportation, and offered access to centralized healthcare, police, and educational

46 services. Those networks formed the backbone of the society as we know it today.

47 These century-old solutions are however becoming increasingly obsolete and inefficient. We
48 have cities jammed with traffic, world-wide outbreaks of disease that are seemingly unstoppable,
49 and political institutions that are deadlocked and unable to act. We face the challenges of global
50 warming, uncertain energy, water, and food supplies, and a rising population and urbanization,
51 that will add 350 million people to the urban population by 2025 in China alone [12].

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

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

67 The engine driving this new nervous system is Big Data: the newly ubiquitous digital data,
68 now available about all aspects of human life. We can analyze patterns of human experience and
69 ideas exchange within the *digital breadcrumbs* that we all leave behind as we move through the
70 world: call records, credit card transactions, GPS location fixes, among others. By recording
71 our choices, these data tell the story of our lives. And this may be very different from what
72 we decide to put on Facebook or Twitter; our postings there are what we choose to tell people,

73 edited according to the standards of the day and filtered to match the persona we are building.
 74 Mining social networks can give some great insights about human nature [3, 23, 37]; who we
 75 really are is however even more accurately determined by where we spend our time and which
 76 things we buy, rather than just what we say we do [22].

77 The process of analyzing the patterns within these digital breadcrumbs is called reality
 78 mining [11, 27], and through it we can learn an enormous amount about who we are. The
 79 Human Dynamics research group at MIT have found that we can use them to tell if we are
 80 likely to get diabetes [28], or whether we are the sort of person who will pay back loans [29]. By
 81 analyzing these patterns across many people, we are discovering that we can begin to explain
 82 many things — crashes, revolutions, bubbles — that previously appeared to be random acts of
 83 God [25]. For this reason the magazine Technology Review named our development of reality
 84 mining as one of the ten technologies that will change the world [14].

85 **3 The New Deal on Data (Arek)**

86 The digital breadcrumbs we leave behind provide clues about who we are and what we want. This
 87 makes these personal data immensely valuable, both for public good and for private companies.
 88 As European Consumer Commissioner, Meglena Kuneva said recently, “Personal data is the
 89 new oil of the Internet and the new currency of the digital world” [19]. This new ability to see
 90 the details of every interaction can be however used for good or for ill. Therefore, maintaining
 91 protection of personal privacy and freedom is critical to our future success as a society. On one
 92 hand, we need to enable even more data sharing for the public good; on the other, we need to
 93 do a much better job in protecting the privacy of the individuals.

94 A successful data-driven society must be able to guarantee that our data will not be abused;
 95 perhaps especially that government will not abuse the power conferred by access to such fine-
 96 grain data. To achieve the positive possibilities of the new society, we require the *New Deal on*
 97 *Data*, workable guarantees that the data needed for public good are readily available while at
 98 the same time protecting the citizenry [27]. For this, we must develop much more powerful and

99 sophisticated tools to use personal data to both build a better society and to protect the rights
100 of the citizens.

101 The key insight that motivates the idea of the New Deal on Data is that our data are worth
102 more when shared, because these aggregated data inform improvements in systems such as
103 public health, transportation, and government. For instance, we have demonstrated that data
104 about the way we behave and where we go can be used to minimize the spread of infectious
105 disease [21, 28]. Our research has reported how we were able to use these digital breadcrumbs
106 to track the spread of influenza from person to person on an individual level. And if we can see
107 it, we can stop it. The result of sharing our personal data here, is that we can build a world
108 where the threat of infectious pandemics is greatly diminished.

109 Similarly, if we are worried about global warming, these shared, aggregated data can show
110 us how patterns of mobility relate to productivity [26]. In turn, this provides us with the ability
111 to design cities that are more productive and, at the same time, more energy efficient. But in
112 order to be able to obtain these results and make a greener world, we need to be able to see
113 the people moving around; this depends on many people willing to contribute their data, even
114 if only anonymously and in aggregate.

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

124 Unfortunately, today most personal data are siloed off in private companies and therefore
125 largely unavailable. Private organizations collect the vast majority of the personal data in

126 the form of mobility patterns, financial transactions, phone and Internet communications, etc.
 127 These data must not remain the exclusive domain of private companies, because then they are
 128 less likely to contribute to the common good. These private organizations must be thus the key
 129 players in the New Deal on Data framework for privacy and data control. Likewise, these data
 130 should not become the exclusive domain of the government, as this will not serve the public
 131 interest of transparency; we should be suspicious of trusting the government with such power.
 132 Ultimately, the entities who should be empowered to share and make decisions about their data,
 133 are people themselves: users, participants, citizens.

134 The ultimate goal is to provide the society tools to analyze and understand what needs to
 135 be done, and to reach the consensus on how to do it. This goes beyond just creating more
 136 communication platforms. The assumption that more interactions between users will result in
 137 better decisions being made, may be very misleading. Although in the recent years we have
 138 seen some great examples of using social networks for better organization in society, for example
 139 during political protests [5,15], we are not even close to the point where we can start reaching
 140 consensus about the big problems: epidemics, climate change, pollution. The discussions must
 141 be data driven, involving both experts and wisdom of the crowds. The problems we are dealing
 142 with as a now global society are not easy. We are responsible for many of them, and being able
 143 to tackle them on a global scale is necessary for our, mankind, survival.

144 4 Personal Data: Emergence of a New Asset Class (Thomas)

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

152 services.

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

- 155 • You have the right to possess data about you. Regardless of what entity collects the data,
156 the data belong to you, and you can access your data at any time. Data collectors thus
157 play a role akin to a bank, managing the data on behalf of their customers.
- 158 • You have the right to full control over the use of your data. The terms of use must be opt-
159 in and clearly explained in plain language. If you are not happy with the way a company
160 uses your data, you can remove the data, just as you would close your account with a bank
161 that is not providing satisfactory service.
- 162 • You have the right to dispose of or distribute your data. You have the option to have data
163 about you destroyed or redeployed elsewhere.

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

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

172 In 2007, one author (Pentland) first proposed the New Deal on Data to the World Economic
173 Forum [38]. Since then, this idea has run through various discussions and eventually helped
174 shape the 2012 Consumer Data Bill of Rights in the United States, along with a matching
175 declaration on Personal Data Rights in the EU. These new regulations hope to accomplish the
176 combined trick of breaking data out of the current silos, thus enabling public goods, while at

177 the same time giving individuals greater control over data about them. But, of course this is
178 still a work in progress and the battle for individual control of personal data rages onward.

179 The World Economic Forum (WEF) has dubbed personal data as the “New Oil” or resource
180 of the 21st century [38]. The discovery of oil and the subsequent development of the oil industry
181 over the past 100 years has spurred not only the development of the automobile industry but also
182 the creation of the global transportation infrastructure, including the massive freeway networks
183 that we see today in the developed nations. The “personal data sector” of the economy today is
184 still in its infancy, its state akin to the oil industry at the late 1890s prior to the development of
185 the Model-T Ford automobile. The productive collaboration between the Government (building
186 the state owned freeways), the private sector (mining and refining oil, building automobiles) and
187 the citizen (the user-base of these services) allowed the developed nations to expand its economies
188 by creating new markets adjacent to the automobile and oil industries.

189 If personal data as the new oil is to reach its global economic potential, there needs to be
190 a productive collaboration between all the stakeholders in the establishment of a *personal data*
191 *ecosystem*. As mentioned in [38] a number of fundamental questions about privacy, property,
192 global governance, human rights - essentially around who should benefit from the products and
193 services built upon personal data - are major uncertainties shaping the opportunity. The rapid
194 rate of technological change and commercialization in using personal data is undermining end
195 user confidence and trust.

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

202 For many participants, the risks and liabilities exceed the economic returns. Besides not
203 having the infrastructure and tools to manage personal data, many end-users simply do not see

the benefit of fully participating in the ecosystem. The current focus of many Internet-based service providers is to capture as much personal data from the end-user and to sell this data into the advertising industry. Personal privacy concerns are thus inadequately addressed at best, or simply overlooked in the majority of the cases. The current technologies and laws fall short of providing the legal and technical infrastructure needed to support a well-functioning digital economy.

The report of the World Economic Forum [38] 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 [33] – 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 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.

5 Enforcing the New Deal on Data (Dazza)

How can we enforce this New Deal? The threat of legal action alone is important, but insufficient, because if you cannot see abuses then you cannot prosecute them. Moreover, who wants more lawsuits anyway? Enforcement can be addressed in significant ways without prosecution of public statute or regulation at all. In many fields, companies and governments rely upon multi-party frameworks of agreed rules governing common business, legal and technical practices to

229 create effective self-organization and enforcement. These approaches hold promise as a method
230 for using institutional controls to form a reliable operational framework balancing the needs for
231 big data, privacy and access.

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

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

253 Having system rules applicable to the networks, applications and data as well as all the ser-
254 vices providers other intermediaries, and the users themselves is the mechanism for establishing
255 and operating a trust network. System rules are sometimes called operating regulations in the

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

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

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

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

286 The Human Dynamics Lab has applied the system rules approach to development of inte-
 287 grated business, technical architecture and rules large scale institutional use of personal data
 288 stores, available as an example under MIT's creative commons license by MIT, at: github.com/HumanDynamics,

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

299 **6 Essential Elements of the New Deal of Data (Brian)**

300 The New Deal on Data restates the controls and expectations people have with respect to their
 301 private property and personal affects and applies it to their personal data and online affects.
 302 Institutional controls must align with the New Deal on Data by providing responsive, rule-based
 303 systems governance and fine grained authorizations for distributed rights management.

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

308 Today, all of the data people generate are stored in closed silos belonging to institutions

309 providing customer services. Phone providers own mobility traces for their users, while music
 310 services store and use data on musical preferences.

311 For these data to be useful to society, the silos must be opened, and the data must be
 312 integrated across institutions far more than they are today. If access to data for the purpose
 313 of creating value—either for the user or the society—is very limited, it does not matter how big
 314 the data is. The value of the data lies not just in the fact that they exist, but rather the
 315 knowledge, understanding, and wisdom we gain from them. It is an even bigger challenge to
 316 open up the data from disparate silos. Accessing multi-faceted data, which exist under multiple
 317 jurisdictions, about people may be prohibitively difficult. Silos are hard to crack open. Despite
 318 these difficulties, such data, not just big, but deep, covering multiple facets of a person’s life,
 319 may be invaluable for research.

320 Recently, we have shown how challenging, but also feasible, it is to open such institutional
 321 Big Data. In the Data For Development (D4D) Challenge ¹, the telecom operator Orange
 322 opened access to a large dataset of call detail records (CDRs) from the Ivory Coast. Working
 323 with the data as part of a challenge, teams of researchers came up with life-changing insights for
 324 the country. For example, one team developed a model for how disease spread in the country
 325 and demonstrated that information campaigns based on one-to-one phone conversations among
 326 members of social groups can be an effective countermeasure [20]. In releasing and analysing this
 327 data, the privacy of the people who generated the data was protected not only by the technical
 328 means, such as removal of the Personally Identifiable Information (PIIs), but also by legal means,
 329 with the researchers signing an agreement they will not use the data for re-identification or other
 330 nefarious means. As we have seen in several cases, such as the Netflix Prize privacy disaster [24]
 331 and other similar privacy breaches [32], true anonymization is extremely hard. Some of the
 332 weight of privacy protection must rest on the legal framework.

333 Opening data from the silos by publishing static datasets is important, but it is only the first
 334 step. We can do even more substantial things when the data is available in real time and can

¹<http://www.d4d.orange.com/home>

335 become part of a society's nervous system. Epidemics can be monitored and prevented in real
336 time [28], underperforming students can be helped, and people with health risks can be treated
337 before they get sick [8]. The same data can potentially be used for stalking, burglarizing one's
338 home, and as justification to charge people more for an insurance policy.

339 In the Unique in the Crowd project [9], de Montjoye et al. showed that even though human
340 beings are highly predictable [30], we are also very unique. Having access to one dataset, it is
341 easy to uniquely fingerprint someone based on just few datapoints, and use this fingerprint to
342 discover their true identity. The higher the resolution of the data, the easier it gets to identify
343 a person from this type of data.

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

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

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

358 Within the existing legal frameworks, it is possible to change the vantage point of the data
359 ownership and put the user, the entity about whom the data are, in control. This may be
360 achieved by providing a copy of the data to a personal silo, which is provided by or on behalf of
361 the user. The user would become the owner of their copy of the data, or whenever possible the

original, in the old Common Law sense with the right to use, transfer, and delete the data. An example of such a mechanism in an institutional context is the Blue Button initiative ², where the patients can get a copy of their health records. Once the copy is with the user, they can do with it as they wish: give it to someone, make it public, do research on it, destroy it.

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

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

²<http://www.healthit.gov/bluebutton>

388 7 Transitioning End-User Assent Practices (Arek)

389 The way the user grants authorizations to the data she owns is not a trivial matter. The flow of
 390 personal information, such as location data, purchases, health records, etc. can be very complex.
 391 Every tweet, every geo-tagged picture, every phone call, and every purchase with credit card,
 392 provide the user's location not only to the primary service, but also to all the applications and
 393 services that have been authorized to access and re-use these data. The authorizations may
 394 come from the end-user or, often, be granted by the collecting service, based on an umbrella
 395 terms of service, allowing the re-use of the data. Implementation of such flows was a crucial
 396 part of the Web 2.0 revolution, realized with RESTful APIs, mashups, and authorization-based
 397 access. The way the data travel between the services has however become arguably too complex
 398 for a user to handle and manage.

399 Increasing the amount of data the user controls and granularity of this control is meaningless
 400 if it cannot be exercised in an informed way. For many years, the End User License Agreements
 401 (EULAs), long incomprehensible texts have been accepted blindly by the end-user, trusting they
 402 have not agreed to anything that could harm them. The process of granting the authorizations
 403 cannot be too complex, as it would prevent the user from understanding her decisions. At
 404 the same time, it cannot be too simplistic, as it may not sufficiently convey the weight of the
 405 privacy-related decisions. It is a challenge in itself, to build the end-user assent systems that
 406 allow the user to understand and adjust their privacy settings. Complex EULAs do not promote
 407 the privacy of the users, effectively pushing them to press *I Agree* in every presented window.
 408 The consequences of those assent actions are not emphasized; as the data being collected is
 409 becoming increasingly complex and our computations more sophisticated, every act of sharing
 410 can lead to great benefits to the society, but also make the users vulnerable.

411 This gap between the interface, the single click, and the effect, can render the data owner-
 412 ship meaningless; the click may wrench people and their data into systems and rules that are
 413 antithetical to fair information practices, such as is prevalent with today's end-user licenses in
 414 cloud services or applications. Managing the potentially long term and opposite dynamics fueled

415 by old deal systems operating simultaneously with the new deal systems is an important design
416 and migration challenge during the transition to a Big Data economy. During this transition
417 and after the New Deal on Data is no longer new, personal data must continue to flow in order
418 to be useful. Protecting the data of people outside of the user-controlled domain is very hard
419 without a combination of cost effective and useful business practices, legal rules, and technical
420 solutions. For these reasons, the Human Dynamics group has focused upon and collaborated
421 with partners to support the clarification of business, legal, and technical short- and longer-term
422 viable solutions.

423 We envision Living Informed Consent, where the user is entitled to know what data is being
424 collected about her by which entities, empowered to understand the implications of data sharing,
425 and finally put in charge of the sharing authorizations. We suggest the readers ask themselves a
426 question: *Which services know which city I am in today?*. Google? Apple? Twitter? Facebook?
427 Flickr? This small application we have authorized a few years ago to access our Facebook
428 check-ins and forgot since then? This is an example of a fundamental question related to user
429 privacy and assent, and yet finding the answer to it may be surprisingly difficult in today's
430 ecosystem. We can hope that most of the services treat the data responsibly and according to
431 user authorizations. In the complex network of data flows however, it is relatively easy for the
432 data to leak to services careless with it or simply malicious [6].

433 It is clear that the promise of the Big Data can only be realized when the data is shared,
434 available even more than it is today. For this, the user herself should be put in the driver's
435 seat and made decisions about who is authorized to see what and for what purpose. To realize
436 this, the solutions for making the user decisions well thought-through must be designed and
437 implemented.

438 8 Business, Legal and Technical Dimensions of Big Data Sys- 439 tems (Dazza)

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

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

460 Once a systems view is adopted, there is a tractable starting point to narrow or broaden
461 the scope of view to see the smaller and larger systems and to make better and more effective
462 use and control of big data. Within a given institution, there may in fact be many different
463 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.

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.

9 Big Data and Personal Data Institutional Controls (Thomas)

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, per se, has been in the field of environmental regulatory frameworks.

A good example of how this concept supports and reflects the goals and objectives of environmental regulation can be found in the policy documents of the EPA. This following definition is instructive, and is part of the Institutional Control Glossary of Terms [35]:

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

Going deeper, the article by DeMeo and Doar [10] defines institutional controls thusly:

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

494 In domains of information technology, this approach is most commonly reflected as “enter-
495 prise controls” related to security. See, for example, the report [18] stating: “Enterprise mobility
496 technologies, especially those designed to retrofit enterprise controls on top of consumer mobile
497 devices, are rapidly evolving. This was a message we heard loud and clear in the study.” This
498 study and analysis also reveals much about the internal controls needed to accommodate mobile
499 device use by employees. In both capacities as employee, consumer and other roles, the use of
500 mobile devices triggers myriad legal, policy and other implications for institutional controls.

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

510 Consider again the analogy of institutional controls in the context of environmental law, and
511 how these types of measures can be applied in the big data, privacy and access context to digital
512 environments. Given the relatively mature and stable state of environmental regulation, there is
513 much to be learned by examining this context of institutional controls. Environmental regulatory
514 compliance with waste management cleanup requirements could include institutional controls
515 restricting land use on adjacent property. In these situations, it is possible that the remediation

516 strategy requires significant use of land outside the property boundaries of the cleanup site.
517 In these cases, the regulators and the land owner responsible for the regulated property must
518 find ways to ensure a common approach among multiple owners and across multiple property
519 environments. Use of measures such as a clauses on the relevant deeds, an enforceable consent
520 order or regulations and zoning rules are examples of more severe institutional controls that
521 can be employed to ensure consistent and effective actions are taken across ownership and real
522 property boundaries.

523 See, for example, FDEP, Division of Waste Management [13] which states that “...RMO III
524 does contemplate contamination beyond the Property boundaries, which would require agree-
525 ment by the adjacent owners to put an RC on their properties as well.”

526 The concept of an “institutional control boundary” is especially clarifying and powerful when
527 applied to the networked and digital boundaries of an institution. In the context of Florida’s
528 environmental regulation frameworks, the phrase is applied to describe the various types of
529 combinations risk management levels related to target cleanup standards and extend beyond
530 the area of a physical property boundary. Also see a recent University of Florida report on
531 Development of Cleanup Target Levels (CTLs) [7] stating “Risk Management Options Level
532 III, like Level II, allows concentrations above the default groundwater CTLs to remain on site.
533 However, in some rare situations, the institutional control boundary at which default CTLs must
534 be met can extend beyond the site property boundary.”

535 The EPA provides considerable information on the nature and use of institutional controls,
536 including situations when the situational scope extends to adjacent properties owned by third
537 parties. See, generally, *EPA Hazardous Waste Corrective Action Guidance on Institutional Con-*
538 *trols* [35]. Also see: *Institutional Controls Bibliography: Institutional Control, Remedy Selection,*
539 *and Post-Construction Completion Guidance and Policy, December 2005* [34].

540 When institutional controls would apply to “separately owned neighboring properties” a
541 number of issues arise. Engagement with affected third parties, requiring the party responsible
542 for site cleanup to use “best efforts” to attain agreement by third parties to institute the relevant

543 institutional controls, use of third party neutrals to resolve disagreements regarding the applica-
544 tion with institutional controls or forcing an acquisition of the neighboring land by forcing the
545 party responsible to purchase the property of by purchase of the property directly by the EPA.
546 See [36].

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

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

566 Following the recommendation of the World Economic Forum recommendations of treating
567 personal data stores in the manner of bank accounts [38], there are a number of infrastructure
568 improvements that need to be realized if the personal data ecosystem is to flourish and deliver
569 new economic opportunities. We believe the following infrastructure improvements are necessary

570 for the coming personal data ecosystem:

- 571 • *New global data provenance network*: In order for personal data to be treated like bank
 572 accounts, the origin information regarding data items coming into the data store must be
 573 maintained [17]. In other words, the provenance of all data items must be accounted for
 574 by the IT infrastructure upon which the personal data store operates. The heterogeneous
 575 provenance databases must then be interconnected in order to provide a resilient and
 576 scalable platform for audit and accounting systems to track and reconcile the movement
 577 of personal data from the respective data stores.

- 578 • *Trust network for computational law*: In order for trust to be established between parties
 579 who wish to exchange personal data, we foresee that some degree of “computational law”
 580 technologies may have to be integrated into the design of personal data systems. Such
 581 technologies should not only verify terms of contracts (e.g. terms of data use) against user-
 582 defined policies but also have mechanisms built-in to ensure non-repudiation of entities
 583 who have accepted these digital contracts. Efforts such as [1, 2] are beginning to bring
 584 non-repudiation and enforceability of contracts into the technical protocol flows.

- 585 • *Development of Institutional Controls for Digital Institutions*: Currently there are a number
 586 of proposal for the creation of virtual currencies (e.g. BitCoin [4], Ven [31]) in which the
 587 systems have the potential to evolve into self-governing “digital institutions” [16]. Such
 588 systems and insitutions that operate on them will necessitate the development of a new
 589 paradigm to understand the aspects of institutional control within their context.

590 10 Scenarios of Use in Context (Dazza)

591 Supporting the effective development of institutional controls for big data requires an under-
 592 standing of how to define and work with the applicable context surrounding the scenarios within
 593 which the big data exists. In particular, the New Deal on Data will require a set of Institutional
 594 Controls involving governance, business, legal and technical aspects that are knowable only with

reference to the relevant context of a factually based scenario of use. The following scenarios demonstrate signature features of the New Deal on Data in various contexts and serve as an anchor to evaluate what Institutional Controls are well aligned.

10.1 Example Scenario: Research Systems

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

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

This year we have deployed a 1,000 phones study at Technical University of Denmark, where we handed out mobile phones to freshmen students in order to study their networks and social behavior in the important change moment of their lives, when they join the university. The study, called SensibleDTU, uses not only data collected from the mobile phones (location, Bluetooth-based proximity, call and sms logs etc.) but also data collected from social networks, questionnaires filled out by participants, behavior in economic games and so on. As the data is collected in the context of the university, there is potentially a big issues of students feeling obliged to participate in the study, feeling that their grades may depend on it, or that the data

621 may influence their grades. In this context, we see the implementation of Living Informed Con-
622 sent not only as a technical mean to put participants in control of the data we collect, but also
623 to convey the message about the opt-in nature of the study, the boundaries of the data usage,
624 and parties accessing the data.

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

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

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

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

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

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

672 10.2 Scenarios of Use Today, Tomorrow and the Day After

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

690 This book represents a high value opportunity to take stock of the current state and domi-
 691 nant trends related to big data and help to illuminate important choices at a moment of early
 692 adoption, dynamic innovation and wide open possibilities. By contemplating the relevant con-
 693 texts of todays scenarios of use in, say, the fields of education, entertainment, government,
 694 manufacturing, transportation and many other core anchors of human activity, we have traction
 695 to postulate how todays prevailing trends are likely to result and what changes perhaps quite
 696 small but of profound long term impact could lead to materially different better outcomes.
 697 Consider that if the essence of the New Deal on Data were accepted today, or soon, the na-
 698 ture, tenor, capabilities and experience of living by future generations could be unrecognizably

699 better. Simply extrapolate from the current anomalous practices regarding personal data and
 700 individual identity and push forward the timeline by 5, 10, 20 years and beyond. The current
 701 trajectory ends up with dystopian scenarios that effectively reverse hard fought but easily lost
 702 constitutional deal of the United States and social compact of common law societies.

703 By contrast, by adopting the New Deal on Data now it is possible to set conditions that
 704 promote prosperity and invention even before the New Deal on Data frameworks are formally
 705 launched. This is because the uncertainty and confusion about the basic premises and expecta-
 706 tions around personal data and identity will be resolved and so investment and risk taking on
 707 a firm foundation can be unleashed. The value of big data can be accessed at less direct cost
 708 and lower risk when uncertainties about privacy liability are addressed and significant the new
 709 value is created by enabling wide scale permission based access to personal data and compu-
 710 tations about such data. Adopting use of personal data services in phases, such one economic
 711 sector, transaction type or data type at a time enables access to the lower costs and new value
 712 in a reasonable manner that allows for time to prepare for and stage each phase of adoption.
 713 By staging and phasing the New Deal on Data typical objections to change based on grounds
 714 of cost, disruption or over regulation can be addressed. Policy incentives can further address
 715 these objections, such as allowing safe harbor protections for conduct of organizations operating
 716 under the rules of a trust network. Policy makers can resolve other difficulties by combina-
 717 tions of strategic transition management methods like allowing safe harbor compliance delays,
 718 or approving alternative adoption paths and granting other non-substantive waivers to ease any
 719 burdens of migrating to new business methods. The key point is change management can be
 720 designed to achieve enough value at every phase for every key stakeholder group such that self
 721 interests and the broader interests are all aligned with the public good.

722 11 Future Research (Brian)

723 Our traditional methods of testing and improving government, organizations, and so on are of
 724 limited use in building a data driven society. Even the scientific method as we normally apply it

725 doesn't work as well as we might expect, because there are so many potential connections that
726 our standard statistical tools generate less than useful results.

727 The reason is that with such rich data, you can easily uncover misleading or unactionable
728 correlations. For instance, lets imagine we discover that people who are unusually active are
729 more likely to get the flu. This is a real example: when we examined the minute-by-minute
730 behavior of a small university community a real-time flow of gigabytes per day for an entire
731 year we noticed that an unusual level of running around often predicted onset of the flu [21].
732 But if we can only analyze the data using traditional statistical methods, we have the problem
733 of discerning why this is true. Is it because the flu virus makes us more active in order to spread
734 itself more quickly? While it is more likely that interacting with many more people than usual
735 makes you more likely to catch the flu, you can't be sure that this is the true cause based on
736 the real-time stream of data alone.

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

742 **11.1 Research on Design and Deployment of Big Data Systems**

743 In order to acheive low risk, high value outcomes efficiently, design and deployment of the coming
744 global wave of big data systems should apply top current research. To understand and address
745 the unique problems and prospects associated with big personal data, the relevant context must
746 be identified and corresponding rules-driven capabilities must be designed into the underlying
747 systems.

748 People and/or systems can determine the right rules to apply to data when the right infor-
749 mation is reliably attached to or logically associated with that data in a standard manner. Any
750 system that can make, use, receive or share big data must be capable of associating provenance

751 and purpose for all data in a common and actionable manner. Requiring a lot of narrative
752 documentation and background about the nuances and circumstances surrounding every data
753 set is both impractical and counterproductive. By contrast, a small amount of metadata listing
754 or reliably linking the parties, transactions, systems and provenance of the data would suffice.
755 This relevant context together with the data forms the basis for accountable analysis on big
756 personal data.

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

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

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

776 The four contextual facets of people, interactions, technology and data provide a sound
777 underpinning for the design of new big data and web 2.0 systems. The existing systems design

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

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

11.2 Research on Big Data for Design of Institutions

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

But with the coming of big data, we are going to be operating very much out of our old, familiar ballpark. These data are often indirect and noisy, and so interpretation of the data requires greater care than is usual. Even more importantly, a great deal of the data is about human behavior, and the questions are ones that seek to connect physical conditions to social

804 outcomes. Until we have a solid, well-proven and quantitative theory of social physics, we wont
 805 be able to formulate and test hypotheses in the way we can when we design bridges or develop
 806 new drugs.

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

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

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

826 The specific research questions we are exploring depend upon a set of personal data services
 827 designed to enable users to collect, store, manage, disclose, share and use data about themselves.
 828 These data can be used for the personal self-empowerment of each member, or (when aggre-
 829 gated) for the improvement of the community through data commons that enable social network
 830 incentives. The ability to share data safely should enable better idea flow among individuals,

831 companies, and government, and we want to see if these tools can in fact increase productivity
 832 and creative output at the scale of an entire city.

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

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

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

851 References

- 852 1. Binding obligations on User-Managed Access (UMA) participants. Technical Specifica-
 853 tions draft-maler-oauth-umatrust-01, Kantara Initiative, July 2013.
- 854 2. User-Managed Access (UMA) profile of OAuth2.0. Technical Specifications draft-
 855 hardjono-oauth-umacore-08, Kantara Initiative, December 2013.

- 856 3. Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social
857 networks. *Science*, 337(6092):337–341, 2012.
- 858 4. Simon Barber, Xavier Boyen, Elaine Shi, and Ersin Uzun. Bitter to Better – how to
859 make Bitcoin a better currency. In *Proceedings Financial Cryptography and Data Security
860 Conference (Lecture Notes in Computer Science Volume 7397)*, pages 399–414, April 2012.
- 861 5. Ellen Barry. Protests in moldova explode, with help of twitter. *New York Times*, 8, 2009.
- 862 6. Nick Bilton. Girls around me: An app takes creepy to a new level. *The New York Times*.
- 863 7. Center for Environmental & Human Toxicology University of Florida. Development of
864 Cleanup Target Levels (CTLs) For Chapter 62-777, F.A.C. Technical report, Division of
865 Waste Management Florida Department of Environmental Protection, February 2005.
- 866 8. Paul Lukowicz Bert Arnrich Cornelia Setz Gerhard Troster David Tacconi, Oscar Mayora
867 and Christian Haring. Activity and emotion recognition to support early diagnosis of
868 psychiatric diseases. pages 100–102. IEEE, 2008.
- 869 9. Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel.
870 Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.
- 871 10. Ralph A. DeMeo and Sarah Meyer Doar. Restrictive covenants as institutional controls
872 for remediated sites: Worth the effort? *The Florida Bar Journal*, 85(2), 2011.
- 873 11. Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Per-
874 sonal and ubiquitous computing*, 10(4):255–268, 2006.
- 875 12. Jonathan Woetzel et al. Preparing for china’s urban billion. 2009.
- 876 13. Florida Department of Environmental Protection - Division of Waste Management. Insti-
877 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)
878 [_topics/publications/wc/csf/icpg.pdf](http://www.dep.state.fl.us/waste/quick_topics/publications/wc/csf/icpg.pdf), June 2012.

- 879 14. Kate Greene. Reality mining. *Technology Review*, 2008.
- 880 15. Lev Grossman. Iran protests: Twitter, the medium of the movement. *Time Magazine*,
881 17, 2009.
- 882 16. Thomas Hardjono, Patrick Deegan, and John Clippinger. On the Design of Trustworthy
883 Compute Frameworks for Self-Organizing Digital Institutions. In *Proceedings of the 16th*
884 *International Conference on Human-Computer Interaction*, 2014.
- 885 17. Thomas Hardjono, Daniel Greenwood, and Alex Pentland. Towards a trustworthy digital
886 infrastructure for core identities and personal data stores. In *Proceedings of the ID360*
887 *Conference on Identity*. University of Texas, April 2013.
- 888 18. Juniper Networks. Secure Data Access Anywhere and Anytime: Current Landscape and
889 Future Outlook of Enterprise Mobile Security. A forrester consulting thought leadership
890 paper commissioned by att and juniper networks, Forrester Research, October 2012.
- 891 19. Meglena Kuneva. Roundtable on Online Data Collection, Targeting and Profiling . [http:](http://europa.eu/rapid/press-release_SPEECH-09-156_en.htm)
892 [//europa.eu/rapid/press-release_SPEECH-09-156_en.htm](http://europa.eu/rapid/press-release_SPEECH-09-156_en.htm), 2009.
- 893 20. Antonio Lima, Manlio De Domenico, Veljko Pejovic, and Mirco Musolesi. Exploiting
894 cellular data for disease containment and information campaigns strategies in country-
895 wide epidemics. School of computer science university of birmingham technical report
896 csr-13-01, University of Birmingham, May 2013.
- 897 21. Anmol Madan, Manuel Cebrian, David Lazer, and Alex Pentland. Social sensing for
898 epidemiological behavior change. In *Proceedings of the 12th ACM international conference*
899 *on Ubiquitous computing*, pages 291–300. ACM, 2010.
- 900 22. AC Madrigal. Dark social: We have the whole history of the web wrong. *The Atlantic*,
901 2013.

- 902 23. Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosen-
 903 quist. Pulse of the nation: Us mood throughout the day inferred from twitter. *Accessed*
 904 *November, 22(2011):2011*, 2010.
- 905 24. Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse
 906 datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125.
 907 IEEE, 2008.
- 908 25. Wei Pan, Yaniv Altshuler, and Alex Sandy Pentland. Decoding social influence and
 909 the wisdom of the crowd in financial trading network. In *Privacy, Security, Risk and*
 910 *Trust (PASSAT), 2012 International Conference on and 2012 International Conferenece*
 911 *on Social Computing (SocialCom)*, pages 203–209. IEEE, 2012.
- 912 26. Wei Pan, Gourab Ghoshal, Coco Krumme, Manuel Cebrian, and Alex Pentland. Urban
 913 characteristics attributable to density-driven tie formation. *Nature communications*, 4,
 914 2013.
- 915 27. ALEX PENTLAND. Reality mining of mobile communications: Toward a new deal on
 916 data. *The Global Information Technology Report 2008–2009*, page 1981, 2009.
- 917 28. Alex Pentland, David Lazer, Devon Brewer, and Tracy Heibeck. Using reality mining to
 918 improve public health and medicine. *Stud Health Technol Inform*, 149:93–102, 2009.
- 919 29. Vivek K Singh, Laura Freeman, Bruno Lepri, and Alex Sandy Pentland. Classifying
 920 spending behavior using socio-mobile data. *HUMAN*, 2(2):pp–99, 2013.
- 921 30. Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of
 922 predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
- 923 31. Stan Stalnaker. The Ven currency, 2013. <http://www.ven.vc>.
- 924 32. Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Fran-*
 925 *cisco)*, pages 1–34, 2000.

- 926 33. The White House. National Strategy for Trusted Identities in Cyberspace: Enhancing On-
927 line Choice, Efficiency, Security, and Privacy. The White House, April 2011. Available on
928 http://www.whitehouse.gov/sites/default/files/rss_viewer/NSTICstrategy_041511.pdf.
- 929 34. United States Environmental Protection Agency. Institutional Controls Bibliography.
930 <http://www.epa.gov/superfund/policy/ic/guide/biblio.pdf>, December 2005.
- 931 35. United States Environmental Protection Agency. RCRA Corrective Action Institu-
932 tional Controls - glossary. [http://www.epa.gov/epawaste/hazard/correctiveaction/](http://www.epa.gov/epawaste/hazard/correctiveaction/resources/guidance/ics/glossary1.pdf)
933 [resources/guidance/ics/glossary1.pdf](http://www.epa.gov/epawaste/hazard/correctiveaction/resources/guidance/ics/glossary1.pdf), 2007.
- 934 36. United States Environmental Protection Agency. Institutional Controls: A Guide to Plan-
935 ning, Implementing, Maintaining, and Enforcing Institutional Controls at Contaminated
936 Sites. Technical Report OSWER 9355.0-89 EPA-540-R-09-001, EPA, December 2012.
- 937 37. Jessica Vitak, Paul Zube, Andrew Smock, Caleb T Carr, Nicole Ellison, and Cliff Lampe.
938 It's complicated: Facebook users' political participation in the 2008 election. *CyberPsy-*
939 *chology, behavior, and social networking*, 14(3):107–114, 2011.
- 940 38. World Economic Forum. Personal Data: The Emergence of a New
941 Asset Class, 2011. Available on [http://www.weforum.org/reports/](http://www.weforum.org/reports/personal-data-emergence-new-asset-class)
942 [personal-data-emergence-new-asset-class](http://www.weforum.org/reports/personal-data-emergence-new-asset-class).