

Operational Framework: Institutional Controls

Daniel "Dazza" Greenwood^{1,*}, Arek 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)	1
2	The New Realities of Living in a Big Data Society (Arek)	1
3	The New Deal on Data (Arek)	3
4	Personal Data: Emergence of a New Asset Class (Thomas)	5
5	Enforcing the New Deal on Data (Dazza)	8
6	Essential Elements of the New Deal of Data (Brian)	11
7	Transitioning End-User Assent Practices (Arek)	15
8	Business, Legal and Technical Dimensions of Big Data Systems (Dazza)	17
9	Big Data and Personal Data Institutional Controls (Thomas)	18
10	Scenarios of Use in Context (Dazza)	22
10.1	Example Scenario: Research Systems	23
10.2	Scenarios of Use Today, Tomorrow and the Day After	26
11	Future Research (Brian)	27
11.1	Research on Design and Deployment of Big Data Systems	28

22	11.2 Research on Big Data for Design of Institutions	30
----	--	----

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. The
 27 control points traditionally relied upon as part of corporate governance, management oversight,
 28 legal compliance, and enterprise architecture must evolve and expand to match operational
 29 frameworks for Big Data. An operational framework used for a Big Data-driven organization
 30 requires a balanced set of institutional controls. These institutional controls must support and
 31 reflect greater user control over personal data and large scale interoperability for data sharing
 32 between and among institutions. Core capabilities of these controls include responsive rule-based
 33 systems governance and fine-grained authorizations for distributed rights management. In the
 34 following sections we explore the emergence of the Big Data Society, outline the ways to support
 35 it in the institutional context, and draft the future directions of research and development.

36 **2 The New Realities of Living in a Big Data Society (Arek)**

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, creating huge
 39 social and environmental problems. The remedy then was to build centralized networks that
 40 delivered clean water and safe food, enabled commerce, removed waste, provided energy, fa-
 41 cilitated transportation, and offered access to centralized healthcare, police, and educational
 42 services. Those networks formed a backbone of the society as we know it today.

43 These century-old solutions are however becoming increasingly obsolete and inefficient. We
 44 have cities jammed with traffic, world-wide outbreaks of disease that are seemingly unstoppable,
 45 and political institutions that are deadlocked and unable to act. We face the challenges of

46 global warming, uncertain energy, water, and food supplies, and a rising population, driving
47 urbanization that will require paving 5 billion square meters of road by 2025 in China alone [1].

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

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

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

71 The process of analyzing the patterns within these digital breadcrumbs is called reality
72 mining [2,3], and through it we can learn an enormous amount about who we are. The Human

73 Dynamics research group at MIT have found that we can use them to tell if we are likely to
 74 get diabetes [4], or whether we are the sort of person who will pay back loans [5]. By analyzing
 75 these patterns across many people, we are discovering that we can begin to explain many things
 76 — crashes, revolutions, bubbles — that previously appeared to be random acts of God [6]. For
 77 this reason the magazine Technology Review named our development of reality mining as one
 78 of the ten technologies that will change the world [7].

79 **3 The New Deal on Data (Arek)**

80 The digital breadcrumbs we leave behind provide clues about who we are and what we want. This
 81 makes these personal data immensely valuable, both for public good and for private companies.
 82 As European Consumer Commissioner, Meglena Kuneva said recently, “Personal data is the
 83 new oil of the Internet and the new currency of the digital world” [8]. This new ability to see
 84 the details of every interaction can be however used for good or for ill. Therefore, maintaining
 85 protection of personal privacy and freedom is critical to our future success as a society. On one
 86 hand, we need to enable even more data sharing for the public good; on the other, we need to
 87 do a much better job in protecting the privacy of the individuals.

88 A successful data-driven society must be able to guarantee that our data will not be abused;
 89 perhaps especially that government will not abuse the power conferred by access to such fine-
 90 grain data. To achieve the positive possibilities of the new society, we require the *New Deal on*
 91 *Data*, workable guarantees that the data needed for public good are readily available while at the
 92 same time protecting the citizenry [3]. We must develop much more powerful and sophisticated
 93 tools to use personal data to both build a better society and to protect the rights of the citizens.

94 The key insight that motivates the creation of the New Deal on Data is that our data are
 95 worth more when shared, because these aggregated data inform improvements in systems such
 96 as public health, transportation, and government. For instance, we have demonstrated that
 97 data about the way we behave and where we go can be used to minimize the spread of infectious
 98 disease [4, 9]. Our research has reported how we were able to use these digital breadcrumbs to

99 track the spread of influenza from person to person on an individual level. And if we can see it,
100 we can stop it. Here the result of sharing our personal data is that we can build a world where
101 the threat of infectious pandemics is greatly diminished.

102 Similarly, if we are worried about global warming, these shared, aggregated data can show
103 us how patterns of mobility relate to productivity [10]. In turn, this provides us with the ability
104 to design cities that are more productive and, at the same time, more energy efficient. But in
105 order to be able to obtain these results and make a greener world, we need to be able to see
106 the people moving around; this depends on many people willing to contribute their data, even
107 if only anonymously and in aggregate.

108 While concrete examples such as better health systems and more energy efficient transporta-
109 tion systems motivate the New Deal on Data, there is an even greater public good that can be
110 achieved by efficient and safe data sharing. To enable sharing of personal data and experiences,
111 we need secure technology and regulation that allow individuals to safely and conveniently share
112 personal information with each other, with corporations, and with government. Consequently,
113 the heart of the New Deal on Data must be to provide both regulatory standards and financial
114 incentives that entice owners to share data, while at the same time serving the interests of both
115 individuals and society at large. We must promote greater idea flow among individuals, not just
116 corporations or government 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
119 the form of mobility patterns, financial transactions, phone and Internet communications, etc.
120 These data must not remain the exclusive domain of private companies, because then they are
121 less likely to contribute to the common good. These private organizations must be thus 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 Ultimately, the entities who should be empowered to share and make decisions about their data,

126 are people themselves: users, participants, citizens.

127 The ultimate goal is to provide the society tools to analyze and understand what needs
 128 to be done, and to reach the consensus how to do it. This goes beyond the creation of more
 129 communication platforms. The assumption that more interactions between users will result in
 130 better decisions being made, may be very misleading. Although in the recent years we have
 131 seen some great examples of using social networks for better organization in society, for example
 132 during political protests [11,12], we are not even close to the point where we can start reaching
 133 consensus about the big problems: epidemics, climate change, pollution. The discussions must
 134 be data driven, involving both experts and wisdom of the crowds. The problems we are dealing
 135 with as a now global society are not easy. We are responsible for many of them, and being able
 136 to tackle them on a global scale is necessary for our, mankind, survival.

137 **4 Personal Data: Emergence of a New Asset Class (Thomas)**

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

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

- 147 • You have the right to possess data about you. Regardless of what entity collects the data,
 148 the data belong to you, and you can access your data at any time. Data collectors thus
 149 play a role akin to a bank, managing the data on behalf of their customers.
- 150 • You have the right to full control over the use of your data. The terms of use must be opt-

151 in and clearly explained in plain language. If you are not happy with the way a company
 152 uses your data, you can remove the data, just as you would close your account with a bank
 153 that is not providing satisfactory service.

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

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

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

164 In 2007, one author (Pentland) first proposed the New Deal on Data to the World Economic
 165 Forum [13]. Since then, this idea has run through various discussions and eventually helped
 166 shape the 2012 Consumer Data Bill of Rights in the United States, along with a matching
 167 declaration on Personal Data Rights in the EU. These new regulations hope to accomplish the
 168 combined trick of breaking data out of the current silos, thus enabling public goods, while at
 169 the same time giving individuals greater control over data about them. But, of course this is
 170 still a work in progress and the battle for individual control of personal data rages onward.

171 The World Economic Forum (WEF) has dubbed personal data as the “New Oil” or resource
 172 of the 21st century [13]. The discovery of oil and the subsequent development of the oil industry
 173 over the past 100 years has spurred not only the development of the automobile industry but also
 174 the creation of the global transportation infrastructure, including the massive freeway networks
 175 that we see today in the developed nations. The “personal data sector” of the economy today is
 176 still in its infancy, its state akin to the oil industry at the late 1890s prior to the development of
 177 the Model-T Ford automobile. The productive collaboration between the Government (building

178 the state owned freeways), the private sector (mining and refining oil, building automobiles) and
 179 the citizen (the user-base of these services) allowed the develop nations to expand its economies
 180 by creating new markets adjacent to the automobile and oil industries.

181 If personal data as the new oil is to reach its global economic potential, there needs to be
 182 a productive collaboration between all the stakeholders in the establishment of a *personal data*
 183 *ecosystem*. As mentioned in [13] a number of fundamental questions about privacy, property,
 184 global governance, human rights - essentially around who should benefit from the products and
 185 services built upon personal data - are major uncertainties shaping the opportunity. The rapid
 186 rate of technological change and commercialization in using personal data is undermining end
 187 user confidence and trust.

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

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

202 The report of the World Economic Forum [13] also suggest a way forward by recommending
 203 a number of areas where efforts could be directed:

- 204 • Alignment of key stakeholders: Citizens, the private sector and the public sector need to

work in support of one another. Efforts such as NSTIC [?] – 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 [?] 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 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

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

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

246 Having system rules applicable to the networks, applications and data as well as all the ser-
247 vices providers other intermediaries, and the users themselves is the mechanism for establishing
248 and operating a trust network. System rules are sometimes called operating regulations in the
249 credit card context, or known as trust frameworks in the identity federations context, or trading
250 partner agreements in a supply value chain context. There are many general examples of multi-
251 party shared architectural and contractual rules that share the generic characteristic of creating
252 binding obligations and enforceable expectations on all participants in scalable networks. An-
253 other common characteristic of the system rules design pattern is that the participants in the
254 network can be widely distributed across very heterogeneous business ownership boundaries,
255 legal governance structures and technical security domains. Yet, the parties need not agree to
256 conform all or most aspects of their basic roles, relationships and activities in order to connect

257 to to systems of a trust network. Cross-domain trusted systems must, by their nature, focus
258 mandatory and enforceable rules narrowly upon the critical items that must be commonly agreed
259 in order for that network to achieve it's purpose.

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

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

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

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

282 The capacity to apply the appropriate methods of enforcement for a trust network depend
283 upon a clear understanding and agreement among parties about the purpose of the trusted

system and the respective roles or expectations of those connecting is as participants. Therefor, an anchor is needed to a clear context of a big data operational framework and institutional controls appropriate for access and confidentiality or privacy. The following section posits the trust model and signature traits of such a context, through the lens of the New Deal on Data.of those connecting is as participants. Therefor, an anchor is needed to a clear context of a big data operational framework and institutional controls appropriate for access and confidentiality or privacy. The following section posits the trust model and signature traits of such a context, through the lens of the New Deal on Data.

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

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

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

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

Today, all of the data people generate are stored in closed silos belonging insitutions providing customer services. Phone providers own mobility traces for their users, while music services store and use data on musical preferences.

For these data to be useful to society, the silos must be opened, and the data must be integrated across institutions far more often than they are today. If access to data for the purpose of creating value—either for the user or the society—is very limited, it does not matter how big the data is. The value of the data lies not just in the fact that they exist. Rather, it is

the knowledge, understanding, and wisdom we gain from them that makes the data valuable. It is an even bigger challenge to open up the data from multiple silos. Accessing the multi-faced data, which exist under multiple jurisdictions, about people may be prohibitively difficult. Silos are hard to crack open. Such data, not just Big but Deep, covering multiple facets of a person's life, may be invaluable for research.

Recently, we have shown how challenging, but also possible, it is to open such institutional Big Data. In the Data For Development (D4D) Challenge ¹, the telecom operator Orange opened access to a large dataset of call detail records (CDRs) from the Ivory Coast. Working with the data as part of a challenge, teams of researchers came up with life-changing insights for the country. The privacy of the people was protected not only by the technical means, such as removal of the Personally Identifiable Information (PIIs), but also by legal means, with the researchers signing an agreement they will not use the data for reidentification or other nefarious means. As we have seen in several cases, such as the Netflix Prize privacy disaster [14] and other similar privacy breaches [15], true anonymization is extremely hard. Some of the weight of privacy protection must rest on the legal framework.

Opening data from the silos by publishing static datasets is important, but it is only the first step. We can do even more important things when the data is available in real time and can become part of a nervous system of a society. Epidemics can be monitored and prevented in real time [4], underperforming students can be helped, and people with health risks can be treated before they get sick [16]. The same data can potentially be used for stalking, burglarizing one's home, and as justification to charge people more for an insurance policy.

In the Unique in the Crowd project [17], we have shown that even though human beings are highly predictable [18], we are also very unique. Having access to one dataset, it is easy to uniquely fingerprint someone based on just few datapoints, and use this fingerprint to discover their true identity. The higher the resolution of the data, the better the data, the easier it gets.

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

¹<http://www.d4d.orange.com/home>

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

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

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

357 Under such a system, users can accumulate data about themselves from multiple sources.
 358 Information on healthcare records, mobility patterns, favorite movies, etc., all belong to the user
 359 and can be accessed based on their authorization. This changes how and what data that can be
 360 obtained for the purpose of research and providing services. Rather than gaining access to the
 361 movements of millions of people from a telcom operator, one can potentially gain access to a

²<http://www.healthit.gov/bluebutton>

362 smaller number but of much richer datasets describing the users from the mobility, health, and
363 shopping perspectives. New startups do not have to build the user profile from scratch, but can
364 jump in offering competitive services based on the user's previously-collected data. Users can
365 immediately get better services, using their data in new places.

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

377 **7 Transitioning End-User Assent Practices (Arek)**

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

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

400 This gap between the interface, the single click, and the effect, can render the data owner-
401 ship meaningless; the click may wrench people and their data into systems and rules that are
402 antithetical to fair information practices, such as is prevalent with today's end-user licenses in
403 cloud services or applications. Managing the potentially long term and opposite dynamics fueled
404 by old deal systems operating simultaneously with the new deal systems is an important design
405 and migration challenge during the transition to a Big Data economy. During this transition
406 and after the New Deal on Data is no longer new, personal data must continue to flow in order
407 to be useful. Protecting the data of people outside of the user-controlled domain is very hard
408 without a combination of cost effective and useful business practices, legal rules, and technical
409 solutions. For these reasons, the Human Dynamics group has focused upon and collaborated
410 with partners to support the clarification of business, legal, and technical short- and longer-term
411 viable solutions.

412 We envision Living Informed Consent, where the user is entitled to know what data is being
413 collected about her by which entities, empowered to understand the implications of data sharing,
414 and finally put in charge of the sharing authorizations. We suggest the readers ask themselves a

415 question: *Which services know which city I am in today?* Google? Apple? Twitter? Facebook?
 416 Flickr? This small application we have authorized a few years ago to access our Facebook
 417 check-ins and forgot since then? This is an example of a fundamental question related to user
 418 privacy and assent, and yet finding the answer to it may be surprisingly difficult in today's
 419 ecosystem. We can hope that most of the services treat the data responsibly and according to
 420 user authorizations. In the complex network of data flows however, it is relatively easy for the
 421 data to leak to services careless with it or simply malicious [19].

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

427 8 Business, Legal and Technical Dimensions of Big Data Sys- 428 tems (Dazza)

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

438 Looking at an institution as a business, legal and technical system is one effective approach
 439 for dealing with the inherent complexity of managing heterogeneous and distributed networks

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

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

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 [20]:

”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 [21] defines institutional controls thusly:

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

In domains of information technology, this approach is most commonly reflected as “enterprise controls” related to security. See, for example, the report [22] stating: “Enterprise mobility technologies, especially those designed to retrofit enterprise controls on top of consumer mobile devices, are rapidly evolving. This was a message we heard loud and clear in the study.” This study and analysis also reveals much about the internal controls needed to accommodate mobile device use by employees. In both capacities as employee, consumer and other roles, the use of mobile devices triggers myriad legal, policy and other implications for institutional controls.

In the legal domain, this concept frequently emerges under the moniker “regulatory compliance” or “legal compliance” anchored in legal and regulatory frameworks such as HIPAA and

492 Sarbanes-Oxley (SOX). These statutory legal frameworks require covered organizations to es-
493 tablished integrated sets of governance, legal, transactional, security and other internal controls
494 to avoid violating the rules. The institutional controls are accomplished in tight integration with
495 engineering and other measures in order to ensure compliance and to control legal and security
496 risk. The use of institutional controls of this type are fundamental methods for achieving and
497 maintaining the transition to a digital, networked and big data footing for any private company,
498 government agency or other organization.

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

512 See, for example, FDEP, Division of Waste Management [23] which states that “...RMO III
513 does contemplate contamination beyond the Property boundaries, which would require agree-
514 ment by the adjacent owners to put an RC on their properties as well.”

515 The concept of an “institutional control boundary” is especially clarifying and powerful when
516 applied to the networked and digital boundaries of an institution. In the context of Florida’s
517 environmental regulation frameworks, the phrase is applied to describe the various types of
518 combinations risk management levels related to target cleanup standards and extend beyond

the area of a physical property boundary. See the Final Technical Report: Development of Cleanup Target Levels (CTLs) for Ch. 62-777, F.A.C. [24] stating “Risk Management Options Level III, like Level II, allows concentrations above the default groundwater CTLs to remain on site. However, in some rare situations, the institutional control boundary at which default CTLs must be met can extend beyond the site property boundary.”

The EPA provides considerable information on the nature and use of institutional controls, including situations when the situational scope extends to adjacent properties owned by third parties. See, generally, *EPA Hazardous Waste Corrective Action Guidance on Institutional Controls* citeEPA2007. Also see: *Institutional Controls Bibliography: Institutional Control, Remedy Selection, and Post-Construction Completion Guidance and Policy, December 2005* [25].

When institutional controls would apply to “separately owned neighboring properties” a number of issues arise. Engagement with affected third parties, requiring the party responsible for site cleanup to use “best efforts” to attain agreement by third parties to institute the relevant institutional controls, use of third party neutrals to resolve disagreements regarding the application with institutional controls or forcing an acquisition of the neighboring land by forcing the party responsible to purchase the property or by purchase of the property directly by the EPA. See [26].

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

546 a general model are one widely understood, accepted and efficient method for defining, agreeing
 547 and enforcing institutional and other controls across business, legal and technical domains of
 548 ownership, governance and operation.

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

555 Following the recommendation of the World Economic Forum recommendations of treating
 556 personal data stores in the manner of bank accounts [13], there are a number of infrastructure
 557 improvements that need to be realized if the personal data ecosystem is to flourish and deliver
 558 new economic opportunities. We believe the following infrastructure improvements are necessary
 559 for the coming personal data ecosystem:

- 560 • *New global data provenance network*: In order for personal data to be treated like bank
 561 accounts, the origin information regarding data items coming into the data store must be
 562 maintained. In other words, the provenance of all data items must be accounted for by
 563 the IT infrastructure upon which the personal data store operates. The heterogeneous
 564 provenance databases must then be interconnected in order to provide a resilient and
 565 scalable platform for audit and accounting systems to track and reconcile the movement
 566 of personal data from the respective data stores.
- 567 • *Trust network for computational law*: In order for trust to be established between parties
 568 who wish to exchange personal data, we foresee that some degree of “computational law”
 569 technologies may have to be integrated into the design of personal data systems. Such
 570 technologies should not only verify terms of contracts (e.g. terms of data use) against
 571 user-defined policies but also have mechanisms built-in to ensure non-repudiation of entities
 572 who have accepted these digital contracts. Efforts such as [27, 28] are beginning to bring

non-repudiation and enforceability of contracts into the technical protocol flows.

- *Development of Institutional Controls for Digital Institutions:* Currently there are a number of proposals for the creation of virtual currencies (e.g. BitCoin [29], Ven [30]) in which the systems have the potential to evolve into self-governing “digital institutions” [?]. Such systems and institutions that operate on them will necessitate the development of a new paradigm to understand the aspects of institutional control within their context.

10 Scenarios of Use in Context (Dazza)

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

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

597 consent exist[TODO: reference], they are often not very well suited for the large-scale studies,
598 where the amount and sensitivity of the data calls for sophisticated privacy controls. As the
599 scale of the studies grows, in terms of the number of participants, collected bits per user, and
600 duration, the EULA-style informed consent is no longer sufficient and makes it hard to claim
601 that participants in fact expressed informed consent.

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

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

621 As the study will last for several years, hopefully allowing us to see the life of a student from
622 the very first friendships made until the graduation party, the consent must remain alive. It is
623 again a matter of balance: we do not want the participants to feel under constant surveillance

624 (as they are not, the data is used mostly in aggregated form), at the same time to remember that
625 in fact, the data is being collected and used. We are still trying to understand how to achieve
626 this equilibrium: how often should we remind the users about the collection effort? should they
627 re-authorize applications from time to time? We see a great hope in the applications we create
628 for the users to provide certain services, simple such as life-logging where they can see how
629 active they are, what are their top places etc. and more advanced, such as artistic visualizations
630 of their social networks. Making the user aware of the data by transforming them into value,
631 can greatly benefit the privacy, making users constantly aware what is being collected, but also
632 what kind of value they can get out of it.

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

643 This touches upon the re-usability of data, as the new experiments may require additional
644 data to be collected, but also have access to all the existing data, based on user authorization.
645 We can imagine going even further, where entirely different studies can re-use participants data
646 from a previous study based on their authorization. When the data are owned by the users,
647 they are free to authorize access to them to any party that requests it. We can see a New Deal on
648 Data pattern here: rather than services (studies) talking to each other about the user data, they
649 talk directly to the users, seeking their authorization. This can address a very important problem
650 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.

10.2 Scenarios of Use Today, Tomorrow and the Day After

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

677 on the fundamental nature and implications of big data and options for the use of operational
678 frameworks acting across domains to balance privacy and access, among other interests.

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

692 By contrast, by adopting the New Deal on Data now it is possible to set conditions that
693 promote prosperity and invention even before the New Deal on Data frameworks are formally
694 launched. This is because the uncertainly and confusion about the basic premises and expecta-
695 tions around personal data and identity will be resolved and so investment and risk taking on
696 a firm foundation can be unleashed. The value of big data can be accessed at less direct cost
697 and lower risk when uncertainties about privacy liability are addressed and significant the new
698 value is created by enabling wide scale permission based access to personal data and compu-
699 tations about such data. Adopting use of personal data services in phases, such one economic
700 sector, transaction type or data type at a time enables access to the lower costs and new value
701 in a reasonable manner that allows for time to prepare for and stage each phase of adoption.
702 By staging and phasing the New Deal on Data typical objections to change based on grounds
703 of cost, disruption or over regulation can be addressed. Policy incentives can further address

these objections, such as allowing safe harbor protections for conduct of organizations operating under the rules of a trust network. Policy makers can resolve other difficulties by combinations of strategic transition management methods like allowing safe harbor compliance delays, or approving alternative adoption paths and granting other non-substantive waivers to ease any burdens of migrating to new business methods. The key point is change management can be designed to achieve enough value at every phase for every key stakeholder group such that self interests and the broader interests are all aligned with the public good.

11 Future Research (Brian)

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

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

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

730 11.1 Research on Design and Deployment of Big Data Systems

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

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

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

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

755 Some capabilities will likely be essential to all big data systems, such as highly scalable
756 active storage, standard methods for integration with other big data systems and a processing

757 architecture enabling high speed statistical analytics. But there are and will continue to emerge
758 multiple types of big data systems. Some functions or controls will likely be important - or
759 even feasible - only for certain types of future systems. For instance, it is reasonable to expect
760 some systems will specialize in enormous volumes of entirely non-personal data from many real-
761 time sources (e.g. for soil science, materials engineering, astronomy, etc) while other big data
762 systems will hinge upon mass quantities of highly sensitive personal information (e.g. for clinical
763 medicine, education and life-long learning, social entertainment, etc).

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

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

779 Big data, by its nature, represents a new set of business, legal and technical capabilities and
780 requirements. Most of the worlds systems today are not capable of ingesting, storing, using or
781 dynamically flowing big data with other systems. Considering that a) big data is of high value
782 immediately and higher value in the short and long terms, and b) the young but competitive
783 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 today's 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 outcomes. Until we have a solid, well-proven and quantitative theory of social physics, we won't be able to formulate and test hypotheses in the way we can when we design bridges or develop new drugs.

Therefore, we must move beyond the closed, laboratory-based question-and-answering process that we currently use and begin to manage our society in a new way. We have to begin to test connections in the real world far earlier and more frequently than we have ever had to do before, using the methods my research group and I have developed for the Friends and Family study or the Social Evolution study. We need to construct Living Laboratories—communities willing to try a new way of doing things or, to put it bluntly, to be guinea pigs—in order to test and prove our ideas. This is new territory and so it is important for us to constantly try out new ideas in the real world in order to see what works and what doesn't.

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

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

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

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

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

837 using the openPDS system.

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

846 References

- 847 1. et al JW (2009) Preparing for china’s urban billion .
- 848 2. Eagle N, Pentland A (2006) Reality mining: sensing complex social systems. *Personal*
 849 and ubiquitous computing 10: 255–268.
- 850 3. PENTLAND A (2009) Reality mining of mobile communications: Toward a new deal on
 851 data. *The Global Information Technology Report 2008–2009* : 1981.
- 852 4. Pentland A, Lazer D, Brewer D, Heibeck T (2009) Using reality mining to improve public
 853 health and medicine. *Stud Health Technol Inform* 149: 93–102.
- 854 5. Singh VK, Freeman L, Lepri B, Pentland AS (2013) Classifying spending behavior using
 855 socio-mobile data. *HUMAN 2*: pp–99.
- 856 6. Pan W, Altshuler Y, Pentland AS (2012) Decoding social influence and the wisdom of
 857 the crowd in financial trading network. In: *Privacy, Security, Risk and Trust (PASSAT)*,
 858 2012 International Conference on and 2012 International Confernece on Social Computing
 859 (SocialCom). IEEE, pp. 203–209.
- 860 7. Greene K (2008) Reality mining. *Technology Review* .

- 861 8. Kuneva M (2009). Roundtable on Online Data Collection, Targeting and Profiling .
862 http://europa.eu/rapid/press-release_SPEECH-09-156_en.htm.
- 863 9. Madan A, Cebrian M, Lazer D, Pentland A (2010) Social sensing for epidemiological
864 behavior change. In: Proceedings of the 12th ACM international conference on Ubiquitous
865 computing. ACM, pp. 291–300.
- 866 10. Pan W, Ghoshal G, Krumme C, Cebrian M, Pentland A (2013) Urban characteristics
867 attributable to density-driven tie formation. Nature communications 4.
- 868 11. Grossman L (2009) Iran protests: Twitter, the medium of the movement. Time Magazine
869 17.
- 870 12. Barry E (2009) Protests in moldova explode, with help of twitter. New York Times 8.
- 871 13. World Economic Forum (2011). Personal Data: The Emergence of
872 a New Asset Class. Available on [http://www.weforum.org/reports/](http://www.weforum.org/reports/personal-data-emergence-new-asset-class)
873 [personal-data-emergence-new-asset-class](http://www.weforum.org/reports/personal-data-emergence-new-asset-class).
- 874 14. Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. In:
875 Security and Privacy, 2008. SP 2008. IEEE Symposium on. IEEE, pp. 111–125.
- 876 15. Sweeney L (2000) Simple demographics often identify people uniquely. Health (San Fran-
877 cisco) : 1–34.
- 878 16. David Tacconi PLBACSGT Oscar Mayora, Haring C (2008) Activity and emotion recog-
879 nition to support early diagnosis of psychiatric diseases. IEEE, pp. 100-102.
- 880 17. de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: The
881 privacy bounds of human mobility. Scientific reports 3.
- 882 18. Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of predictability in human mobility.
883 Science 327: 1018–1021.

- 884 19. Bilton N girls around me: An app takes creepy to a new level. The New York Times .
- 885 20. United States Environmental Protection Agency (2007).
886 RCRA Corrective Action Institutional Controls - glossary.
887 <http://www.epa.gov/epawaste/hazard/correctiveaction/resources/guidance/ics/glossary1.pdf>.
888
- 889 21. DeMeo RA, Doar SM (2011) Restrictive covenants as institutional controls for remediated
890 sites: Worth the effort? The Florida Bar Journal 85.
- 891 22. Juniper Networks (2012) Secure data access anywhere and anytime: Current landscape
892 and future outlook of enterprise mobile security. Forrester report, Forrester.
- 893 23. Florida Department of Environmental Protection - Division of Waste
894 Management (2012). Institutional Controls Procedures Guidance.
895 http://www.dep.state.fl.us/waste/quick_topics/publications/wc/csf/icpg.pdf.
896
- 897 24. Center for Environmental & Human Toxicology University of Florida (2005) Development
898 of Cleanup Target Levels (CTLs) For Chapter 62-777, F.A.C. Technical report, Division
899 of Waste Management Florida Department of Environmental Protection.
- 900 25. United States Environmental Protection Agency (2005). Institutional Controls Bibliog-
901 raphy. <http://www.epa.gov/superfund/policy/ic/guide/biblio.pdf>.
- 902 26. United States Environmental Protection Agency (2012) Institutional Controls: A Guide
903 to Planning, Implementing, Maintaining, and Enforcing Institutional Controls at Con-
904 taminated Sites. Technical Report OSWER 9355.0-89 EPA-540-R-09-001, EPA.
- 905 27. (2013) User-Managed Access (UMA) profile of OAuth2.0. Technical report, Kantara Ini-
906 tiative.

- 907 28. (2013) Binding obligations on User-Managed Access (UMA) participants. Technical re-
908 port, Kantara Initiative.
- 909 29. Barber S, Boyen X, Shi E, Uzun E (2012) Bitter to Better – how to make Bitcoin a better
910 currency. In: Proceedings Financial Cryptography and Data Security Conference (Lecture
911 Notes in Computer Science Volume 7397). pp. 399-414.
- 912 30. Stalnaker S (2013). The Ven currency. [Http://www.ven.vc](http://www.ven.vc).