

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)	24
10.1	Example Scenario: Research Systems	24
10.2	Scenarios of Use Today, Tomorrow and the Day After	27
11	Future Research (Brian)	29
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 [13].

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, 24, 38]; 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 [23].

77 The process of analyzing the patterns within these digital breadcrumbs is called reality
 78 mining [12, 28], 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 [29], or whether we are the sort of person who will pay back loans [30]. 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 [26]. 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 [15].

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” [20]. 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 [28]. 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 [22, 29]. 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 [27]. 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

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

The ultimate goal is to provide the society tools to analyze and understand what needs to be done, and to reach the consensus on how to do it. This goes beyond just creating more communication platforms. The assumption that more interactions between users will result in better decisions being made, may be very misleading. Although in the recent years we have seen some great examples of using social networks for better organization in society, for example during political protests [5,16], we are not even close to the point where we can start reaching consensus about the big problems: epidemics, climate change, pollution. The discussions must be data driven, involving both experts and wisdom of the crowds. The problems we are dealing with as a now global society are not easy. We are responsible for many of them, and being able to tackle them on a global scale is necessary for our, mankind, survival.

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

It has long been recognized that the first step to promoting liquidity in land and commodity markets is to guarantee ownership rights so that people can safely buy and sell. Similarly, the first step toward creating greater idea and idea flow ('idea liquidity') is to define ownership rights. The only politically viable course is to give individual citizens rights over data that are about them and in fact, in the European Union these rights flow directly from the constitution **AS: Citation? There is no 'EU constitution' per se.** . We need to recognize personal data 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 [39]. 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 [39]. 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 [39], 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 [39] 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 [34] – 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
 248 the trust network **AS: This sentence is hard to understand. Missing verb?** . When
 249 a trust network involves use of personal data, then the user permissions and corresponding
 250 limits on use are fundamental to the trust model. In this context, the permissions, including
 251 the provenance of the data, should require appropriate levels of audit. A well designed trust
 252 network, elegantly integrating computer and legal rules, allows automatic auditing of data use
 253 and allows individuals to change their permissions and withdraw data.

254 Having system rules applicable to the networks, applications, and data as well as all the
 255 services providers other intermediaries, and the users themselves is the mechanism for estab-

lishing 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 multiparty shared architectural and contractual rules that share the generic characteristic of creating binding obligations and enforceable expectations on all participants in scalable networks. Another common characteristic of the system rules design pattern is that the participants in the network can be widely distributed across very heterogeneous business ownership boundaries, legal governance structures, and technical security domains. Yet, the parties need not agree to conform all or most aspects of their basic roles, relationships, and activities in order to connect to to systems of a trust network. Cross-domain trusted systems must, by their nature, focus mandatory and enforceable rules narrowly upon the critical items that must be commonly agreed in order for that network to achieve it's purpose.

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

A system like this has made the interbank money transfer system among the safest systems in the world and the daily backbone for exchanges of trillions of dollars, but until recently such systems were only for the 'big guys'. To give individuals a similarly safe method of managing personal data, the Human Dynamics research group at MIT, in partnership with the Institute for Data Driven Design, co-founded by John Clippinger and one author (Pentland), have helped build open Personal Data Store (openPDS) [10]. See <http://openPDS.media.mit.edu>

for project information and <https://github.com/HumanDynamics/openPDS> for the open source code.

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

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

The capacity to apply the appropriate methods of enforcement for a trust network depend upon a clear understanding and agreement among parties about the purpose of the trusted system and the respective roles or expectations of those connecting as participants. Therefore, 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)

The New Deal on Data restates the controls and expectations people have with respect to their private property and personal assets. Institutional controls must align with the New Deal on Data by providing 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 to governments and institutions providing customer services. Phone providers own mobility traces for their users,

309 while music services store and use data on musical preferences.

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

319 Recently, we have shown how challenging, but also feasible, it is to open such institutional
320 Big Data. In the Data For Development (D4D) Challenge <http://www.d4d.orange.com/home>,
321 the telecom operator Orange opened access to a large dataset of call detail records (CDRs) from
322 the Ivory Coast. Working with the data as part of a challenge, teams of researchers came up
323 with life-changing insights for the country. For example, one team developed a model for how
324 disease spread in the country and demonstrated that information campaigns based on one-to-one
325 phone conversations among members of social groups can be an effective countermeasure [21]. In
326 releasing and analysing this data, the privacy of the people who generated the data was protected
327 not only by the technical means, such as removal of the Personally Identifiable Information
328 (PIIs), but also by legal means, with the researchers signing an agreement they will not use the
329 data for re-identification or other nefarious purposes. As we have seen in several cases, such as
330 the Netflix Prize privacy disaster [25] and other similar privacy breaches [33], true anonymization
331 is extremely hard. Some of the weight of privacy protection must rest on the legal framework.

332 Opening data from the silos by publishing static datasets is important, but it is only the first
333 step. We can do even more substantial things when the data is available in real time and can
334 become part of a society’s nervous system. Epidemics can be monitored and prevented in real
335 time [29], underperforming students can be helped, and people with health risks can be treated

336 before they get sick [8]. The same data can potentially be used for stalking, burglarizing one's
 337 home, and as justification to charge people more for an insurance policy.

338 In the Unique in the Crowd [9], de Montjoye et al. showed that even though human beings
 339 are highly predictable [31], we are also very unique. Having access to one dataset, it may be
 340 easy to uniquely fingerprint someone based on just few datapoints, and use this fingerprint to
 341 discover their true identity. The higher the resolution of the data, the easier it gets to identify
 342 a person from this type of data.

343 The question of privacy in this context effectively becomes a question of control: Who can
 344 release the data of one's movements? To whom? How much and how often?

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

355 **AS: What about this one?** [Footnote: HEW Report, OECD rendition, EU Directive,
 356 DHS/NSTIC version, MGL FIPA and culminating in New Deal on Data adaptation].

357 Within the existing legal frameworks, it is possible to change the vantage point of the data
 358 ownership and put the user, the entity about whom the data are, in control. This may be
 359 achieved by providing a copy of the data to a personal store, which is provided by or on behalf
 360 of the user. The user would become the owner of their copy of the data, or whenever possible,
 361 the original, in the old Common Law sense with the right to use, transfer, and delete the data.
 362 An example of such a mechanism in an institutional context is the Blue Button initiative [http:](http://)

363 `//www.healthit.gov/bluebutton`, where the patients can get a copy of their health records.
364 Once the copy is with the user, they can do with it as they wish: give it to someone, make it
365 public, do research on it, destroy it.

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

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

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 very vulnerable.

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

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

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

It is clear that the promise of the Big Data can only be realized when the data is shared, available even more than it is today. For this, the user herself should be put in the driver's seat and made decisions about who is authorized to see what and for what purpose. To realize this, the solutions for making the user decisions well thought-through must be designed and 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 other-
441 wise – the traditional container of an institution makes less and less sense. Institutional controls
442 apply, by definition by or to some type of institutional entity such as a business, governmental,
443 or religious organization. A combined view of the business, legal, and technical facts and cir-
444 cumstances surrounding big data is necessary to know what access, confidentiality, and other
445 expectations exist. The relevant contextual aspects of Big Data of one institutional is often
446 profoundly different from that of another. As more and more organizations use and rely upon
447 big data, a single formula for institutional controls will not work for increasingly heterogeneous
448 business, 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 of
451 actors and interactions. The business models, interface-point operational practices and relevant
452 assumptions must be consistent and frequently carefully agreed upon 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
457 goals and purposes for the system and data. When technical standards are selected, configured
458 and applied to systems they too must support and reflect the business and legal dimensions and
459 be 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. **AS: The paragraph above is hard to understand I think.**

Many organizations are structured with clear leadership on business, legal, and technical issues functionally assigned to top level executive roles. Business issues are typically allocated 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 has been in the field of environmental regulatory frameworks.

A good example of how this concept supports and reflects the goals and objectives of environmental regulation can be found in the policy documents of the Environmental Protection Agency (EPA). This following definition is instructive, and is part of the Institutional Control Glossary of Terms [36]:

“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 infor-

489 mational devices.”

490 Going deeper, the article by DeMeo and Doar [11] defines institutional controls thusly:

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

496 In domains of information technology, this approach is most commonly reflected as “enter-
497 prise controls” related to security. See, for example, the report [19] stating: “Enterprise mobility
498 technologies, especially those designed to retrofit enterprise controls on top of consumer mobile
499 devices, are rapidly evolving. This was a message we heard loud and clear in the study.” This
500 study and analysis also reveals much about the internal controls needed to accommodate mobile
501 device use by employees. In both capacities as employee, consumer, and other roles, the use of
502 mobile devices triggers myriad legal, policy, and other implications for institutional controls.

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

513 Consider again the analogy of institutional controls in the context of environmental law, and
514 how these types of measures can be applied in the Big Data, privacy, and access context to digital

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

See, for example, Florida Department of Environmental Protection (FDEP), Division of Waste Management [14] which states that “...RMO III does contemplate contamination beyond the Property boundaries, which would require agreement by the adjacent owners to put an RC on their properties as well.”

The concept of an “institutional control boundary” is especially clarifying and powerful when applied to the networked and digital boundaries of an institution. In the context of Florida’s environmental regulation frameworks, the phrase is applied to describe the various types of combinations risk management levels related to target cleanup standards and extend beyond the area of a physical property boundary. Also see a recent University of Florida report on Development of Cleanup Target Levels (CTLs) [7] 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 Con-*

542 trols [36]. Also see: *Institutional Controls Bibliography: Institutional Control, Remedy Selection,*
 543 *and Post-Construction Completion Guidance and Policy, December 2005* [35].

544 When institutional controls would apply to “separately owned neighboring properties” a
 545 number of issues arise. Engagement with affected third parties, requiring the party responsible
 546 for site cleanup to use “best efforts” to attain agreement by third parties to institute the relevant
 547 institutional controls, use of third party neutrals to resolve disagreements regarding the appli-
 548 cation with institutional control,s or forcing an acquisition of the neighboring land by forcing
 549 the party responsible to purchase the property of by purchase of the property directly by the
 550 EPA [37].

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

564 The use of system rules and integrated participation agreements by developers and end-users
 565 is a way to ensure intended operational frameworks conform to applicable institutional controls.
 566 The example of “living consent” described in this chapter, demonstrates how institutional con-
 567 trols comprised of legal and definite workflow measures, in concert with technical methods, can
 568 result in a higher level of performance, while appropriately balancing legitimate interests of

569 various parties regarding use and access to personal data.

570 Following the World Economic Forum recommendations of treating personal data stores in
 571 the manner of bank accounts [39], there are a number of infrastructure improvements that need to
 572 be realized, if the personal data ecosystem is to flourish and deliver new economic opportunities.
 573 We believe the following infrastructure improvements are necessary for the coming personal data
 574 ecosystem: **AS: We should remove the bullets, turn them into continuous text.**

- 575 • *New global data provenance network*: In order for personal data to be treated like bank
 576 accounts, the origin information regarding data items coming into the data store must be
 577 maintained [18]. In other words, the provenance of all data items must be accounted for
 578 by the IT infrastructure upon which the personal data store operates. The heterogeneous
 579 provenance databases must then be interconnected in order to provide a resilient and
 580 scalable platform for audit and accounting systems to track and reconcile the movement
 581 of personal data from the respective data stores.
- 582 • *Trust network for computational law*: In order for trust to be established between parties
 583 who wish to exchange personal data, we foresee that some degree of “computational law”
 584 technologies may have to be integrated into the design of personal data systems. Such
 585 technologies should not only verify terms of contracts (e.g. terms of data use) against user-
 586 defined policies but also have mechanisms built-in to ensure non-repudiation of entities
 587 who have accepted these digital contracts. Efforts such as [1, 2] are beginning to bring
 588 non-repudiation and enforceability of contracts into the technical protocol flows.
- 589 • *Development of institutional controls for digital institutions*: Currently there are a number
 590 of proposal for the creation of virtual currencies (e.g. BitCoin [4], Ven [32]) in which the
 591 systems have the potential to evolve into self-governing “digital institutions” [17]. Such
 592 systems and institutions that operate on them will necessitate the development of a new
 593 paradigm to understand the aspects of institutional control within their context.

594 10 Scenarios of Use in Context (Dazza)

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

602 10.1 Example Scenario: Research Systems

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

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

617 This year we have deployed a 1,000 phones study at Technical University of Denmark, where
 618 we handed out mobile phones to freshmen students in order to study their networks and so-
 619 cial behavior in the important change moment of their lives, when they join the university.

620 The study, called SensibleDTU, uses not only data collected from the mobile phones (location,
621 Bluetooth-based proximity, call and sms logs etc.) but also data collected from social networks,
622 questionnaires filled out by participants, behavior in economic games and so on. As the data
623 is collected in the context of the university, there is potentially a big issues of students feeling
624 obliged to participate in the study, feeling that their grades may depend on it, or that the data
625 may influence their grades. In this context, we see the implementation of Living Informed Con-
626 sent not only as a technical mean to put participants in control of the data we collect, but also
627 to convey the message about the opt-in nature of the study, the boundaries of the data usage,
628 and parties accessing the data.

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

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

647 what kind of value they can get out of it.

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

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

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

674 music recommendation may get better, because your music service can make a use of this extra
675 data. Your data.

676 10.2 Scenarios of Use Today, Tomorrow and the Day After

677 By inquiring into and noting the four facets of relevant context described above, it is possible
678 to describe the basic material contours of any scenario within which big data exists such that
679 the operational framework and adequate approaches to access, use, confidentiality and other key
680 interests can be sustainably balanced. In a commercial scenario the relevant people might be a
681 consumer, merchants, banks, products manufacturers, third party app developers and individual
682 members of that consumers bowling team. The relevant transactions might be a purchase of
683 goods by the consumer from the merchant and the corresponding app that was embedded in
684 the goods and the downstream transaction of involving the consumer now transacting with the
685 merchant bowling alley and interacting with a bowling team, with whom activity and sports
686 performance data are shared and aggregated and further mashed up. The rest of the con-
687 text can be described for any given scenario and this all could be expressed specifically rather
688 than by role simply by running a report from the system to indicate it was in fact John Doe,
689 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
691 same techniques, used for scenarios in other economic sectors and social endeavors shed light
692 on the fundamental nature and implications of big data and options for the use of operational
693 frameworks acting across domains to balance privacy and access, among other intersts.

694 This book represents a high value opportunity to take stock of the current state and domi-
695 nant trends related to big data and help to illuminate important choices at a moment of early
696 adoption, dynamic innovation and wide open possibilities. By contemplating the relevant con-
697 texts of todays scenarios of use in, say, the fields of education, entertainment, government,
698 manufacturing, transportation and many other core anchors of human activity, we have traction
699 to postulate how todays prevailing trends are likely to result and what changes perhaps quite

700 small but of profound long term impact could lead to materially different better outcomes.
701 Consider that if the essence of the New Deal on Data were accepted today, or soon, the na-
702 ture, tenor, capabilities and experience of living by future generations could be unrecognizably
703 better. Simply extrapolate from the current anomalous practices regarding personal data and
704 individual identity and push forward the timeline by 5, 10, 20 years and beyond. The current
705 trajectory ends up with dystopian scenarios that effectively reverse hard fought but easily lost
706 constitutional deal of the United States and social compact of common law societies.

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

726 11 Future Research (Brian)

727 Our traditional methods of testing and improving government, organizations, and so on are of
728 limited use in building a data driven society. Even the scientific method as we normally apply it
729 doesn't work as well as we might expect, because there are so many potential connections that
730 our standard statistical tools generate less than useful results.

731 The reason is that with such rich data, you can easily uncover misleading or unactionable
732 correlations. For instance, lets imagine we discover that people who are unusually active are
733 more likely to get the flu. This is a real example: when we examined the minute-by-minute
734 behavior of a small university community a real-time flow of gigabytes per day for an entire
735 year we noticed that an unusual level of running around often predicted onset of the flu [22].
736 But if we can only analyze the data using traditional statistical methods, we have the problem
737 of discerning why this is true. Is it because the flu virus makes us more active in order to spread
738 itself more quickly? While it is more likely that interacting with many more people than usual
739 makes you more likely to catch the flu, you can't be sure that this is the true cause based on
740 the real-time stream of data alone.

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

746 11.1 Research on Design and Deployment of Big Data Systems

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

752 People and/or systems can determine the right rules to apply to data when the right infor-
753 mation is reliably attached to or logically associated with that data in a standard manner. Any
754 system that can make, use, receive or share big data must be capable of associating provenance
755 and purpose for all data in a common and actionable manner. Requiring a lot of narrative
756 documentation and background about the nuances and circumstances surrounding every data
757 set is both impractical and counterproductive. By contrast, a small amount of metadata listing
758 or reliably linking the parties, transactions, systems and provenance of the data would suffice.
759 This relevant context together with the data forms the basis for accountable analysis on big
760 personal data.

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

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

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

779 all the harder to make when the systems are intended for big data.

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

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

798 **11.2 Research on Big Data for Design of Institutions**

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

804 But with the coming of big data, we are going to be operating very much out of our old,

805 familiar ballpark. These data are often indirect and noisy, and so interpretation of the data
 806 requires greater care than is usual. Even more importantly, a great deal of the data is about
 807 human behavior, and the questions are ones that seek to connect physical conditions to social
 808 outcomes. Until we have a solid, well-proven and quantitative theory of social physics, we wont
 809 be able to formulate and test hypotheses in the way we can when we design bridges or develop
 810 new drugs.

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

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

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

830 The specific research questions we are exploring depend upon a set of personal data services
 831 designed to enable users to collect, store, manage, disclose, share and use data about themselves.

832 These data can be used for the personal self-empowerment of each member, or (when aggre-
 833 gated) for the improvement of the community through data commons that enable social network
 834 incentives. The ability to share data safely should enable better idea flow among individuals,
 835 companies, and government, and we want to see if these tools can in fact increase productivity
 836 and creative output at the scale of an entire city.

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

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

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

855 References

- 856 1. Binding obligations on User-Managed Access (UMA) participants. Technical Specifica-
 857 tions draft-maler-oauth-umatrust-01, Kantara Initiative, July 2013.

- 858 2. User-Managed Access (UMA) profile of OAuth2.0. Technical Specifications draft-
859 hardjono-oauth-umacore-08, Kantara Initiative, December 2013.
- 860 3. Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social
861 networks. *Science*, 337(6092):337–341, 2012.
- 862 4. Simon Barber, Xavier Boyen, Elaine Shi, and Ersin Uzun. Bitter to Better – how to
863 make Bitcoin a better currency. In *Proceedings Financial Cryptography and Data Security
864 Conference (Lecture Notes in Computer Science Volume 7397)*, pages 399–414, April 2012.
- 865 5. Ellen Barry. Protests in moldova explode, with help of twitter. *New York Times*, 8, 2009.
- 866 6. Nick Bilton. Girls around me: An app takes creepy to a new level. *The New York Times*.
- 867 7. Center for Environmental & Human Toxicology University of Florida. Development of
868 Cleanup Target Levels (CTLs) For Chapter 62-777, F.A.C. Technical report, Division of
869 Waste Management Florida Department of Environmental Protection, February 2005.
- 870 8. Paul Lukowicz Bert Arnrich Cornelia Setz Gerhard Troster David Tacconi, Oscar Mayora
871 and Christian Haring. Activity and emotion recognition to support early diagnosis of
872 psychiatric diseases. pages 100–102. IEEE, 2008.
- 873 9. Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel.
874 Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.
- 875 10. Yves-Alexandre de Montjoye, Samuel S Wang, Alex Pentland, Dinh Tien Tuan Anh, An-
876 witaman Datta, Kevin W Hamlen, Lalana Kagal, Murat Kantarcioglu, Vaibhav Khadilkar,
877 Kerim Yasin Oktay, et al. On the trusted use of large-scale personal data. *IEEE Data
878 Eng. Bull.*, 35(4):5–8, 2012.
- 879 11. Ralph A. DeMeo and Sarah Meyer Doar. Restrictive covenants as institutional controls
880 for remediated sites: Worth the effort? *The Florida Bar Journal*, 85(2), 2011.

- 881 12. Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Per-*
882 *sonal and ubiquitous computing*, 10(4):255–268, 2006.
- 883 13. Jonathan Woetzel et al. Preparing for china’s urban billion. 2009.
- 884 14. Florida Department of Environmental Protection - Division of Waste Management. Insti-
885 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)
886 [_topics/publications/wc/csf/icpg.pdf](http://www.dep.state.fl.us/waste/quick_topics/publications/wc/csf/icpg.pdf), June 2012.
- 887 15. Kate Greene. Reality mining. *Technology Review*, 2008.
- 888 16. Lev Grossman. Iran protests: Twitter, the medium of the movement. *Time Magazine*,
889 17, 2009.
- 890 17. Thomas Hardjono, Patrick Deegan, and John Clippinger. On the Design of Trustworthy
891 Compute Frameworks for Self-Organizing Digital Institutions. In *Proceedings of the 16th*
892 *International Conference on Human-Computer Interaction*, 2014.
- 893 18. Thomas Hardjono, Daniel Greenwood, and Alex Pentland. Towards a trustworthy digital
894 infrastructure for core identities and personal data stores. In *Proceedings of the ID360*
895 *Conference on Identity*. University of Texas, April 2013.
- 896 19. Juniper Networks. Secure Data Access Anywhere and Anytime: Current Landscape and
897 Future Outlook of Enterprise Mobile Security. A forrester consulting thought leadership
898 paper commissioned by att and juniper networks, Forrester Research, October 2012.
- 899 20. Meglena Kuneva. Roundtable on Online Data Collection, Targeting and Profiling . [http:](http://europa.eu/rapid/press-release_SPEECH-09-156_en.htm)
900 [//europa.eu/rapid/press-release_SPEECH-09-156_en.htm](http://europa.eu/rapid/press-release_SPEECH-09-156_en.htm), 2009.
- 901 21. Antonio Lima, Manlio De Domenico, Veljko Pejovic, and Mirco Musolesi. Exploiting
902 cellular data for disease containment and information campaigns strategies in country-
903 wide epidemics. School of computer science university of birmingham technical report
904 csr-13-01, University of Birmingham, May 2013.

- 905 22. Anmol Madan, Manuel Cebrian, David Lazer, and Alex Pentland. Social sensing for
 906 epidemiological behavior change. In *Proceedings of the 12th ACM international conference*
 907 *on Ubiquitous computing*, pages 291–300. ACM, 2010.
- 908 23. AC Madrigal. Dark social: We have the whole history of the web wrong. *The Atlantic*,
 909 2013.
- 910 24. Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosen-
 911 quist. Pulse of the nation: Us mood throughout the day inferred from twitter. *Accessed*
 912 *November, 22(2011):2011*, 2010.
- 913 25. Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse
 914 datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125.
 915 IEEE, 2008.
- 916 26. Wei Pan, Yaniv Altshuler, and Alex Sandy Pentland. Decoding social influence and
 917 the wisdom of the crowd in financial trading network. In *Privacy, Security, Risk and*
 918 *Trust (PASSAT), 2012 International Conference on and 2012 International Confernece*
 919 *on Social Computing (SocialCom)*, pages 203–209. IEEE, 2012.
- 920 27. Wei Pan, Gourab Ghoshal, Coco Krumme, Manuel Cebrian, and Alex Pentland. Urban
 921 characteristics attributable to density-driven tie formation. *Nature communications*, 4,
 922 2013.
- 923 28. ALEX PENTLAND. Reality mining of mobile communications: Toward a new deal on
 924 data. *The Global Information Technology Report 2008–2009*, page 1981, 2009.
- 925 29. Alex Pentland, David Lazer, Devon Brewer, and Tracy Heibeck. Using reality mining to
 926 improve public health and medicine. *Stud Health Technol Inform*, 149:93–102, 2009.
- 927 30. Vivek K Singh, Laura Freeman, Bruno Lepri, and Alex Sandy Pentland. Classifying
 928 spending behavior using socio-mobile data. *HUMAN*, 2(2):pp–99, 2013.

- 929 31. Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of
930 predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
- 931 32. Stan Stalnaker. The Ven currency, 2013. <http://www.ven.vc>.
- 932 33. Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Fran-*
933 *cisco)*, pages 1–34, 2000.
- 934 34. The White House. National Strategy for Trusted Identities in Cyberspace: Enhancing On-
935 line Choice, Efficiency, Security, and Privacy. The White House, April 2011. Available on
936 http://www.whitehouse.gov/sites/default/files/rss_viewer/NSTICstrategy_041511.pdf.
- 937 35. United States Environmental Protection Agency. Institutional Controls Bibliography.
938 <http://www.epa.gov/superfund/policy/ic/guide/biblio.pdf>, December 2005.
- 939 36. United States Environmental Protection Agency. RCRA Corrective Action Institu-
940 tional Controls - glossary. [http://www.epa.gov/epawaste/hazard/correctiveaction/](http://www.epa.gov/epawaste/hazard/correctiveaction/resources/guidance/ics/glossary1.pdf)
941 [resources/guidance/ics/glossary1.pdf](http://www.epa.gov/epawaste/hazard/correctiveaction/resources/guidance/ics/glossary1.pdf), 2007.
- 942 37. United States Environmental Protection Agency. Institutional Controls: A Guide to Plan-
943 ning, Implementing, Maintaining, and Enforcing Institutional Controls at Contaminated
944 Sites. Technical Report OSWER 9355.0-89 EPA-540-R-09-001, EPA, December 2012.
- 945 38. Jessica Vitak, Paul Zube, Andrew Smock, Caleb T Carr, Nicole Ellison, and Cliff Lampe.
946 It’s complicated: Facebook users’ political participation in the 2008 election. *CyberPsy-*
947 *chology, behavior, and social networking*, 14(3):107–114, 2011.
- 948 39. World Economic Forum. Personal Data: The Emergence of a New
949 Asset Class, 2011. Available on [http://www.weforum.org/reports/](http://www.weforum.org/reports/personal-data-emergence-new-asset-class)
950 [personal-data-emergence-new-asset-class](http://www.weforum.org/reports/personal-data-emergence-new-asset-class).