

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

3 Daniel "Dazza" Greenwood^{1,*}, Arkadiusz Stopczynski^{1,2}, Brian Sweatt¹, Thomas Hardjono¹,
4 Alex Sandy Pentland¹

5 **1 MIT**

6 **2 DTU**

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

8 **Contents**

9	1 Introduction (Arek)	2
10	2 The New Realities of Living in a Big Data Society (Arek)	2
11	3 The New Deal on Data (Arek)	4
12	4 Personal Data: Emergence of a New Asset Class (Thomas)	6
13	5 Enforcing the New Deal on Data (Dazza)	9
14	6 Essential Elements of the New Deal of Data (Brian)	12
15	7 Transitioning End-User Assent Practices (Arek)	16
16	8 Business, Legal, and Technical Dimensions of Big Data Systems (Dazza)	17
17	9 Big Data and Personal Data Institutional Controls (Thomas)	19
18	10 Scenarios of Use in Context (Dazza)	23
19	10.1 Example Scenario: Research Systems	24
20	10.2 Scenarios of Use Today, Tomorrow and the Day After	26

21	11 Future Research (Brian)	28
22	11.1 Research on Design and Deployment of Big Data Systems	29
23	11.2 Research on Big Data for Design of Institutions	31

24 **1 Introduction (Arek)**

25 To realize the promise and prospects of a Big Data society and avoid its security and confiden-
 26 tiality perils, institutions are updating operational frameworks governing business, legal, and
 27 technical dimensions of their internal organization and interactions with the outside world. This
 28 chapter describes how the common good can be served by framing these types of institutional
 29 rules and processes to ensure a greater user control over personal data, as well as large scale risk
 30 management and interoperability for data sharing between and among institutions.

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

41 **2 The New Realities of Living in a Big Data Society (Arek)**

42 Sustaining a healthy, safe, and efficient society is a scientific and engineering challenge going
 43 back to the 1800s, when the Industrial Revolution spurred rapid urban growth, creating huge
 44 social and environmental problems. The remedy then was to build centralized networks that

45 delivered clean water and safe food, enabled commerce, removed waste, provided energy, fa-
46 cilitated transportation, and offered access to centralized healthcare, police, and educational
47 services. Those networks formed the backbone of the society as we know it today.

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

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

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

68 The engine driving this new nervous system is Big Data: the newly ubiquitous digital data,
69 now available about all aspects of human life. We can analyze patterns of human experience and
70 ideas exchange within the *digital breadcrumbs* that we all leave behind as we move through the
71 world: call records, credit card transactions, GPS location fixes, among others. By recording

our choices, these data tell the story of our lives. And this may be very different from what we decide to put on Facebook or Twitter; our postings there are what we choose to tell people, edited according to the standards of the day and filtered to match the persona we are building. Mining social networks can give some great insights about human nature [4, 26, 40]; who we really are is however even more accurately determined by where we spend our time and which things we buy, rather than just what we say we do [25].

The process of analyzing the patterns within these digital breadcrumbs is called reality mining [13, 30], and through it we can learn an enormous amount about who we are. The Human Dynamics research group at MIT have found that we can use them to tell if we are likely to get diabetes [31], or whether we are the sort of person who will pay back loans [32]. By analyzing these patterns across many people, we are discovering that we can begin to explain many things — crashes, revolutions, bubbles — that previously appeared to be random acts of God [28]. For this reason the magazine *Technology Review* named our development of reality mining as one of the ten technologies that will change the world [16].

3 The New Deal on Data (Arek)

The digital breadcrumbs we leave behind provide clues about who we are, what we do and want. This makes these personal data immensely valuable, both for public good and for private companies. As European Consumer Commissioner, Meglena Kuneva said recently, “Personal data is the new oil of the Internet and the new currency of the digital world” [21]. This new ability to see the details of every interaction can be however used for good or for ill. Therefore, maintaining protection of personal privacy and freedom is critical to our future success as a society. We need to enable even more data sharing for the public good; at the same time, we need to do a much better job in protecting the privacy of the individuals.

A successful data-driven society must be able to guarantee that our data will not be abused; perhaps especially that government will not abuse the power conferred by access to such fine-grain data. To achieve the positive possibilities of the new society, we require the *New Deal on*

98 *Data*, workable guarantees that the data needed for public good are readily available while at
99 the same time protecting the citizenry [30]. For this, we must develop much more powerful and
100 sophisticated tools to use personal data to both build a better society and to protect the rights
101 of the citizens.

102 The key insight that motivates the idea of the New Deal on Data is that our data are worth
103 more when shared, because these aggregated data inform improvements in systems such as
104 public health, transportation, and government. For instance, we have demonstrated that data
105 about the way we behave and where we go can be used to minimize the spread of infectious
106 disease [24, 31]. Our research has reported how we were able to use these digital breadcrumbs
107 to track the spread of influenza from person to person on an individual level. And if we can see
108 it, we can stop it.

109 Similarly, if we are worried about global warming, these shared, aggregated data can show
110 us how patterns of mobility relate to productivity [29]. 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
 126 form of mobility patterns, financial transactions, phone and Internet communications. These
 127 data must not remain the exclusive domain of private companies, because then they are less
 128 likely to contribute to the common good. These private organizations must be thus the key
 129 players in the New Deal on Data framework for privacy and data control. Likewise, these data
 130 should not become the exclusive domain of the government, as this will not serve the public
 131 interest of transparency; we should be suspicious of trusting the government with such power.
 132 Ultimately, the entities who should be empowered to share and make decisions about their data,
 133 are people themselves: users, participants, citizens.

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

145 4 Personal Data: Emergence of a New Asset Class (Thomas)

146 It has long been recognized that the first step to promoting liquidity in land and commodity
 147 markets is to guarantee ownership rights so that people can safely buy and sell. Similarly, the
 148 first step toward creating greater idea and idea flow (“idea liquidity”) is to define ownership rights.
 149 The only politically viable course is to give individual citizens rights over data that are about
 150 them and in fact, in the European Union these rights flow directly from the constitution **AS:**

151 **Citation? There is no 'EU constitution' per se.** . We need to recognize personal data
 152 as a valuable asset of the individual that is given to companies and government in return for
 153 services.

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

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

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

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

173 In 2007, one author (Pentland) first proposed the New Deal on Data to the World Economic
 174 Forum [41]. Since then, this idea has run through various discussions and eventually helped
 175 shape the 2012 Consumer Data Bill of Rights in the United States, along with a matching
 176 declaration on Personal Data Rights in the EU. These new regulations hope to accomplish the

combined trick of breaking data out of the current silos, thus enabling public goods, while at the same time giving individuals greater control over data about them. But, of course this is still a work in progress and the battle for individual control of personal data rages onward.

The World Economic Forum (WEF) has dubbed personal data as the “New Oil” or resource of the 21st century [41]. The discovery of oil and the subsequent development of the oil industry over the past 100 years has spurred not only the development of the automobile industry but also the creation of the global transportation infrastructure, including the massive freeway networks that we see today in the developed nations. The “personal data sector” of the economy today is still in its infancy, its state akin to the oil industry at the late 1890s prior to the development of the Model-T Ford automobile. The productive collaboration between the Government (building the state owned freeways), the private sector (mining and refining oil, building automobiles) and the citizen (the user-base of these services) allowed the developed nations to expand its economies by creating new markets adjacent to the automobile and oil industries.

If personal data, as the new oil, is to reach its global economic potential, there needs to be a productive collaboration between all the stakeholders in the establishment of a *personal data ecosystem*. As mentioned in [41], a number of fundamental questions about privacy, property, global governance, human rights – essentially around who should benefit from the products and services built upon personal data – are major uncertainties shaping the opportunity. The rapid rate of technological change and commercialization in using personal data is undermining end user confidence and trust.

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

For many participants, the risks and liabilities exceed the economic returns. Besides not

204 having the infrastructure and tools to manage personal data, many end-users simply do not see
 205 the benefit of fully participating in the ecosystem. The current focus of many Internet-based
 206 service providers is to capture as much personal data from the end-user and to sell this data into
 207 the advertising industry. Personal privacy concerns are thus inadequately addressed at best,
 208 or simply overlook in the majority of the cases. The current technologies and laws fall short
 209 of providing the legal and technical infrastructure needed to support a well-functioning digital
 210 economy.

211 The report of the World Economic Forum [41] also suggest a way forward by recommending
 212 a number of areas where efforts could be directed:

- 213 • Alignment of key stakeholders: Citizens, the private sector and the public sector need to
 214 work in support of one another. Efforts such as NSTIC [36] – albeit still in its infancy –
 215 represents a promising direction for a global collaboration.
- 216 • Viewing “data as money”: There needs to be a new change in mindset where an individual’s
 217 personal data items are viewed and treated in the same way as their money. These personal
 218 data items would reside in an “account” (like a bank account) where it would be controlled,
 219 managed, exchanged and accounted for just like personal banking services operate today.
- 220 • End-user centricity: All entities in the ecosystem need to recognize that end-users are
 221 vital and independent stakeholders in the co-creation and value exchange of services and
 222 experiences. Efforts such as the *User managed Access* (UMA) initiative [2] point in the
 223 right direction by designing systems that are user-centric and managed by the user.

224 5 Enforcing the New Deal on Data (Dazza)

225 How can we enforce this New Deal? The threat of legal action alone is important, but insufficient,
 226 because if you cannot see abuses then you cannot prosecute them. Moreover, who wants more
 227 lawsuits anyway? Enforcement can be addressed in significant ways without prosecution of

228 public statute or regulation at all. In many fields, companies and governments rely upon multi-
229 party frameworks of agreed rules governing common business, legal, and technical practices to
230 create effective self-organization and enforcement. These approaches hold promise as a method
231 for using institutional controls to form a reliable operational framework balancing the needs for
232 big data, privacy, and access.

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

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

255 Having system rules applicable to the networks, applications, and data as well as all the
256 services providers other intermediaries, and the users themselves is the mechanism for estab-
257 lishing and operating a trust network. System rules are sometimes called operating regulations
258 in the credit card context, or known as trust frameworks in the identity federations context, or
259 trading partner agreements in a supply value chain context. There are many general examples of
260 multiparty shared architectural and contractual rules that share the generic characteristic of cre-
261 ating binding obligations and enforceable expectations on all participants in scalable networks.
262 Another common characteristic of the system rules design pattern is that the participants in
263 the network can be widely distributed across very heterogeneous business ownership boundaries,
264 legal governance structures, and technical security domains. Yet, the parties need not agree to
265 conform all or most aspects of their basic roles, relationships, and activities in order to connect
266 to to systems of a trust network. Cross-domain trusted systems must, by their nature, focus
267 mandatory and enforceable rules narrowly upon the critical items that must be commonly agreed
268 in order for that network to achieve it's purpose.

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

278 A system like this has made the interbank money transfer system among the safest systems
279 in the world and the daily backbone for exchanges of trillions of dollars, but until recently such
280 systems were only for the 'big guys'. To give individuals a similarly safe method of managing
281 personal data, the Human Dynamics research group at MIT, in partnership with the Insti-

282 tute for Data Driven Design, co-founded by John Clippinger and one author (Pentland), have
 283 helped build open Personal Data Store (openPDS) [11]. See <http://openPDS.media.mit.edu>
 284 for project information and <https://github.com/HumanDynamics/openPDS> for the open source
 285 code.

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

289 The Human Dynamics Lab has applied the system rules approach to development of in-
 290 tegrated business, technical architecture, and rules large scale institutional use of personal
 291 data stores, available as an example under MIT’s creative commons license by MIT, at <https://github.com/HumanDynamics/SystemRules>.
 292

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

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

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

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

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

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

320 Recently, we have shown how challenging, but also feasible, it is to open such institutional
321 Big Data. In the Data For Development (D4D) Challenge <http://www.d4d.orange.com/home>,
322 the telecom operator Orange opened access to a large dataset of call detail records (CDRs) from
323 the Ivory Coast. Working with the data as part of a challenge, teams of researchers came up
324 with life-changing insights for the country. For example, one team developed a model for how
325 disease spread in the country and demonstrated that information campaigns based on one-to-one
326 phone conversations among members of social groups can be an effective countermeasure [23]. In
327 releasing and analysing this data, the privacy of the people who generated the data was protected
328 not only by the technical means, such as removal of the Personally Identifiable Information
329 (PIIs), but also by legal means, with the researchers signing an agreement they will not use the
330 data for re-identification or other nefarious purposes. As we have seen in several cases, such as
331 the Netflix Prize privacy disaster [27] and other similar privacy breaches [35], true anonymization
332 is extremely hard. Some of the weight of privacy protection must rest on the legal framework.

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

335 become part of a society's nervous system. Epidemics can be monitored and prevented in real
 336 time [31], underperforming students can be helped, and people with health risks can be treated
 337 before they get sick [9]. The same data can potentially be used for stalking, burglarizing one's
 338 home, and as justification to charge people more for an insurance policy.

339 In the Unique in the Crowd [10], de Montjoye et al. showed that even though human beings
 340 are highly predictable [33], we are also very unique. Having access to one dataset, it may be
 341 easy to uniquely fingerprint someone based on just few datapoints, and use this fingerprint to
 342 discover their true identity. The higher the resolution of the data, the easier it gets to identify
 343 a person from this type of data.

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

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

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

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

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

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

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

389 7 Transitioning End-User Assent Practices (Arek)

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

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

412 This gap between the interface – single click – and the effect, can render the data ownership
 413 meaningless; the click may wrench people and their data into systems and rules that are anti-
 414 thetical to fair information practices, such as is prevalent with today's end-user licenses in cloud
 415 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.

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 [7]. We need to build the solutions to help the user to make well thought-through decisions about data sharing.

8 Business, Legal, and Technical Dimensions of Big Data Systems (Dazza)

When it comes to data intended to be accessible over networks – whether big, personal, or otherwise – the traditional container of an institution makes less and less sense. Institutional controls apply, by definition by or to some type of institutional entity such as a business, governmental, or religious organization. A combined view of the business, legal, and technical facts and circumstances surrounding big data is necessary to know what access, confidentiality, and other expectations exist. The relevant contextual aspects of Big Data of one institutional is often

441 profoundly different from that of another. As more and more organizations use and rely upon
 442 big data, a single formula for institutional controls will not work for increasingly heterogeneous
 443 business, legal and technical environments in play.

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

455 Once a systems view is adopted, there is a tractable starting point to narrow or broaden
 456 the scope of view to see the smaller and larger systems and to make better and more effective
 457 use and control of big data. Within a given institution, there may in fact be many different
 458 discernable institutions and corresponding systems and any given system of one institution will
 459 frequently in fact exist across many different discernable institutions. However, defining as a
 460 ‘system’ the thing to which institutional controls apply provides an achievable and measurable
 461 basis for balancing privacy, access and other interests in big data. **AS: The paragraph above**
 462 **is hard to understand I think.**

463 Many organizations are structured with clear leadership on business, legal, and technical
 464 issues functionally assigned to top level executive roles. Business issues are typically allocated
 465 to roles such as CEO, COO or CFO, while leadership on legal issues is commonly assigned to
 466 roles like general counsel and regulatory compliance and technical leads are often the roles of
 467 CIO, CTO or CSO. Having top level leadership for each of the business, legal, and technical

468 aspects of a trust network is a critical success factor.

469 9 Big Data and Personal Data Institutional Controls (Thomas)

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

475 A good example of how this concept supports and reflects the goals and objectives of en-
 476 vironmental regulation can be found in the policy documents of the Environmental Protection
 477 Agency (EPA). This following definition is instructive, and is part of the Institutional Control
 478 Glossary of Terms [38]:

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

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

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

491 In domains of information technology, this approach is most commonly reflected as “enter-
 492 prise controls” related to security. See, for example, the report [20] stating: “Enterprise mobility

493 technologies, especially those designed to retrofit enterprise controls on top of consumer mobile
494 devices, are rapidly evolving. This was a message we heard loud and clear in the study.” This
495 study and analysis also reveals much about the internal controls needed to accommodate mobile
496 device use by employees. In both capacities as employee, consumer, and other roles, the use of
497 mobile devices triggers myriad legal, policy, and other implications for institutional controls.

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

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

520 property boundaries.

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

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

534 The EPA provides considerable information on the nature and use of institutional controls,
535 including situations when the situational scope extends to adjacent properties owned by third
536 parties. See, generally, *EPA Hazardous Waste Corrective Action Guidance on Institutional Con-*
537 *trols* [38]. Also see: *Institutional Controls Bibliography: Institutional Control, Remedy Selection,*
538 *and Post-Construction Completion Guidance and Policy, December 2005* [37].

539 When institutional controls would apply to “separately owned neighboring properties” a
540 number of issues arise. Engagement with affected third parties, requiring the party responsible
541 for site cleanup to use “best efforts” to attain agreement by third parties to institute the relevant
542 institutional controls, use of third party neutrals to resolve disagreements regarding the appli-
543 cation with institutional control,s or forcing an acquisition of the neighboring land by forcing
544 the party responsible to purchase the property of by purchase of the property directly by the
545 EPA [39].

546 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 owners to enforcement actions corresponding to the applicable controls. Short of government regulation, the use of system rules as a general model are one widely understood, accepted, and efficient method for defining, agreeing, and enforcing institutional and other controls across business, legal, and technical domains of ownership, governance, and operation.

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

Following the World Economic Forum recommendations of treating personal data stores in the manner of bank accounts [41], there are a number of infrastructure improvements that need to be realized, if the personal data ecosystem is to flourish and deliver new economic opportunities. We believe the following infrastructure improvements are necessary for the coming personal data ecosystem: **AS: We should remove the bullets, turn them into continuous text.**

- *New global data provenance network*: In order for personal data to be treated like bank accounts, the origin information regarding data items coming into the data store must be maintained [19]. In other words, the provenance of all data items must be accounted for by the IT infrastructure upon which the personal data store operates. The heterogeneous

provenance databases must then be interconnected in order to provide a resilient and scalable platform for audit and accounting systems to track and reconcile the movement of personal data from the respective data stores.

- *Trust network for computational law*: In order for trust to be established between parties who wish to exchange personal data, we foresee that some degree of “computational law” technologies may have to be integrated into the design of personal data systems. Such technologies should not only verify terms of contracts (e.g. terms of data use) against user-defined policies but also have mechanisms built-in to ensure non-repudiation of entities who have accepted these digital contracts. Efforts such as [1, 2] 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 proposal for the creation of virtual currencies (e.g. BitCoin [5], Ven [34]) in which the systems have the potential to evolve into self-governing “digital institutions” [18]. 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.

597 10.1 Example Scenario: Research Systems

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

599 Computational Social Science (CSS) studies are based on data collected often with an ex-
600 tremely high resolution and scale [22]. Using computational power combined with mathematical
601 models, such data can be used to provide insights into human nature. Much of the data collected,
602 for example mobility traces are sensitive and private; most individuals would feel uncomfortable
603 sharing them publicly. The need for solutions to ensure the privacy of the individuals has grown
604 alongside the data collection efforts.

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

613 One author (Stopczynski) deployed this year a 1,000 phones study at Technical University
614 of Denmark, freshmen students received mobile phones in order to study their networks and
615 social behavior in the important change moment of their lives, when joining the university.
616 The study, called SensibleDTU (<https://www.sensible.dtu.dk/?lang=en>), uses not only data
617 collected from the mobile phones (location, Bluetooth-based proximity, call and sms logs etc.)
618 but also data collected from social networks, questionnaires filled out by participants, behavior
619 in economic games and so on. As the data is collected in the context of the university, there is
620 potentially a big issue of students feeling obliged to participate in the study, feeling that their
621 grades may depend on it, or that the data may influence their grades. In this context, we see
622 the implementation of Living Informed Consent not only as a technical mean to put participants
623 in control of the data we collect, but also to convey the message about the opt-in nature of the

624 study, the boundaries of the data usage, and parties accessing the data.

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

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

644 When a study of such scale is deployed, the particular experiments and sub-studies may
645 not be exactly defined from the very beginning. The initial deployment is a creation of a
646 testbed, where shorter or longer experiments can take place; for example part of the population
647 may participate in the experiment of quantifying the impact of feedback application on their
648 activity levels. Being able to create such experiments in an efficient way is a huge value for the
649 researchers. To do that in the most frictionless way, we give the users the choice to opt-in to
650 those additional experiments, providing some financial or other benefits. This is only possible

651 if there is a notion of identity of the participants, stronger and more useful than a piece of
 652 paper with a signature. This identity allows us to reach out to people, offer them additional
 653 experiments, and let them agree or disagree to them.

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

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

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

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

676 By inquiring into and noting the four facets of relevant context described above, it is pos-

sible 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 on the fundamental nature and implications of Big Data and options for the use of operational frameworks acting across domains to balance privacy and access, among other interests.

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

This book represents a high value opportunity to take stock of the current state and dominant trends related to Big Data and help to illuminate important choices at a moment of early adoption, dynamic innovation, and wide open possibilities. By contemplating the relevant contexts of today's scenarios of use in, say, the fields of education, entertainment, government, manufacturing, transportation, and many other core anchors of human activity, we have traction to postulate how today's prevailing trends are likely to result and what changes - perhaps quite small but of profound long term impact - could lead to materially different better outcomes. Consider that if the essence of the New Deal on Data was accepted today, or soon, the nature, tenor, capabilities, and experience of living by future generations could be unrecognizably

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

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

727 **11 Future Research (Brian)**

728 Our traditional methods of testing and improving government, organizations, and so on are of
 729 limited use in building a data-driven society. Even the scientific method that we normally use

do not work as well as we might expect, because there are so many potential connections that our standard statistical tools generate less than useful results.

The reason is that with such rich data, you can easily uncover misleading or unactionable correlations. For instance, let us imagine we discover that people who are unusually active are more likely to get the flu. This is a real example: when we examined the minute-by-minute behavior of a small university community - a real-time flow of gigabytes per day for an entire year - we noticed that an unusual level of running around often predicted onset of the flu [24]. But if we can only analyze the data using traditional statistical methods, we have the problem of discerning why this is true. Is it because the flu virus makes us more active in order to spread itself more quickly? While it is more likely that interacting with many more people than usual makes you more likely to catch the flu, you can't be sure that this is the true cause based on the real-time stream of data alone.

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

11.1 Research on Design and Deployment of Big Data Systems

AS: I do not understand this paragraph? What is top current research? Where is it applied? In order to achieve low risk, high value outcomes efficiently, design and deployment of the coming global wave of Big Data systems should apply top current research. To understand and address the unique problems and prospects associated with big personal data, the relevant context must be identified and corresponding rules-driven capabilities must be designed into the underlying systems.

People or systems can determine the right rules to apply to data when the right information is reliably attached to or logically associated with that data in a standard manner **AS: I think I**

756 **understand this previous sentences but I' m not sure. What is 'a standard manner'**
 757 **here? What is the right information? It seems it is described in the next sentences,**
 758 **maybe remove this one then? .** Any system that can make, use, receive, or share Big Data
 759 must be capable of associating provenance and purpose for all data in a common and actionable
 760 manner. Requiring a lot of narrative documentation and background about the nuances and
 761 circumstances surrounding every data set is both impractical and counterproductive. By con-
 762 trast, a small amount of metadata listing or reliably linking the parties, transactions, systems
 763 and provenance of the data would suffice. This relevant context together with the data forms
 764 the basis for accountable analysis on big personal data.

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

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

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

779 While some capabilities, such as ingesting and processing astronomical data-sets, will be
 780 unique to only a subset of Big Data systems, it is reasonable to anticipate that data will be
 781 increasingly cross-tabulated, merged, and otherwise shared with other systems and data. It can
 782 be nearly impossible to conclusively predict for the entire life of a system what data will be

received by, created in, or transmitted from that system at the design phase. This prediction is all the harder to make when the systems are intended for Big Data.

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

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

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, ‘plausible hypotheses’.

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

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

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

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

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

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

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

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

861 12 Conclusions

862 Some summary of the chapter. Take the introduction and show that we have proved the thesis.

863 References

- 864 1. Binding obligations on User-Managed Access (UMA) participants. Technical Specifica-
865 tions draft-maler-oauth-umatrust-01, Kantara Initiative, July 2013.
- 866 2. User-Managed Access (UMA) profile of OAuth2.0. Technical Specifications draft-
867 hardjono-oauth-umacore-08, Kantara Initiative, December 2013.
- 868 3. Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. Social fmri: Investi-
869 gating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*,
870 7(6):643–659, 2011.
- 871 4. Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social
872 networks. *Science*, 337(6092):337–341, 2012.
- 873 5. Simon Barber, Xavier Boyen, Elaine Shi, and Ersin Uzun. Bitter to Better – how to
874 make Bitcoin a better currency. In *Proceedings Financial Cryptography and Data Security*
875 *Conference (Lecture Notes in Computer Science Volume 7397)*, pages 399–414, April 2012.
- 876 6. Ellen Barry. Protests in moldova explode, with help of twitter. *New York Times*, 8, 2009.
- 877 7. Nick Bilton. Girls around me: An app takes creepy to a new level. *The New York Times*.
- 878 8. Center for Environmental & Human Toxicology University of Florida. Development of
879 Cleanup Target Levels (CTLs) For Chapter 62-777, F.A.C. Technical report, Division of
880 Waste Management Florida Department of Environmental Protection, February 2005.
- 881 9. Paul Lukowicz Bert Arnrich Cornelia Setz Gerhard Troster David Tacconi, Oscar Mayora
882 and Christian Haring. Activity and emotion recognition to support early diagnosis of
883 psychiatric diseases. pages 100–102. IEEE, 2008.

- 884 10. Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel.
885 Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.
- 886 11. Yves-Alexandre de Montjoye, Samuel S Wang, Alex Pentland, Dinh Tien Tuan Anh, An-
887 witaman Datta, Kevin W Hamlen, Lalana Kagal, Murat Kantarcioglu, Vaibhav Khadilkar,
888 Kerim Yasin Oktay, et al. On the trusted use of large-scale personal data. *IEEE Data*
889 *Eng. Bull.*, 35(4):5–8, 2012.
- 890 12. Ralph A. DeMeo and Sarah Meyer Doar. Restrictive covenants as institutional controls
891 for remediated sites: Worth the effort? *The Florida Bar Journal*, 85(2), 2011.
- 892 13. Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Per-*
893 *sonal and ubiquitous computing*, 10(4):255–268, 2006.
- 894 14. Jonathan Woetzel et al. Preparing for china’s urban billion. 2009.
- 895 15. Florida Department of Environmental Protection - Division of Waste Management. Insti-
896 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)
897 [_topics/publications/wc/csf/icpg.pdf](http://www.dep.state.fl.us/waste/quick_topics/publications/wc/csf/icpg.pdf), June 2012.
- 898 16. Kate Greene. Reality mining. *Technology Review*, 2008.
- 899 17. Lev Grossman. Iran protests: Twitter, the medium of the movement. *Time Magazine*,
900 17, 2009.
- 901 18. Thomas Hardjono, Patrick Deegan, and John Clippinger. On the Design of Trustworthy
902 Compute Frameworks for Self-Organizing Digital Institutions. In *Proceedings of the 16th*
903 *International Conference on Human-Computer Interaction*, 2014.
- 904 19. Thomas Hardjono, Daniel Greenwood, and Alex Pentland. Towards a trustworthy digital
905 infrastructure for core identities and personal data stores. In *Proceedings of the ID360*
906 *Conference on Identity*. University of Texas, April 2013.

- 907 20. Juniper Networks. Secure Data Access Anywhere and Anytime: Current Landscape and
 908 Future Outlook of Enterprise Mobile Security. A forrester consulting thought leadership
 909 paper commissioned by att and juniper networks, Forrester Research, October 2012.
- 910 21. Meglena Kuneva. Roundtable on Online Data Collection, Targeting and Profiling . [http:](http://europa.eu/rapid/press-release_SPEECH-09-156_en.htm)
 911 [//europa.eu/rapid/press-release_SPEECH-09-156_en.htm](http://europa.eu/rapid/press-release_SPEECH-09-156_en.htm), 2009.
- 912 22. David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi,
 913 Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann,
 914 et al. Life in the network: the coming age of computational social science. *Science (New*
 915 *York, NY)*, 323(5915):721, 2009.
- 916 23. Antonio Lima, Manlio De Domenico, Veljko Pejovic, and Mirco Musolesi. Exploiting
 917 cellular data for disease containment and information campaigns strategies in country-
 918 wide epidemics. School of computer science university of birmingham technical report
 919 csr-13-01, University of Birmingham, May 2013.
- 920 24. Anmol Madan, Manuel Cebrian, David Lazer, and Alex Pentland. Social sensing for
 921 epidemiological behavior change. In *Proceedings of the 12th ACM international conference*
 922 *on Ubiquitous computing*, pages 291–300. ACM, 2010.
- 923 25. AC Madrigal. Dark social: We have the whole history of the web wrong. *The Atlantic*,
 924 2013.
- 925 26. Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosen-
 926 quist. Pulse of the nation: Us mood throughout the day inferred from twitter. *Accessed*
 927 *November*, 22(2011):2011, 2010.
- 928 27. Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse
 929 datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125.
 930 IEEE, 2008.

- 931 28. Wei Pan, Yaniv Altshuler, and Alex Sandy Pentland. Decoding social influence and
 932 the wisdom of the crowd in financial trading network. In *Privacy, Security, Risk and*
 933 *Trust (PASSAT), 2012 International Conference on and 2012 International Confernece*
 934 *on Social Computing (SocialCom)*, pages 203–209. IEEE, 2012.
- 935 29. Wei Pan, Gourab Ghoshal, Coco Krumme, Manuel Cebrian, and Alex Pentland. Urban
 936 characteristics attributable to density-driven tie formation. *Nature communications*, 4,
 937 2013.
- 938 30. ALEX PENTLAND. Reality mining of mobile communications: Toward a new deal on
 939 data. *The Global Information Technology Report 2008–2009*, page 1981, 2009.
- 940 31. Alex Pentland, David Lazer, Devon Brewer, and Tracy Heibeck. Using reality mining to
 941 improve public health and medicine. *Stud Health Technol Inform*, 149:93–102, 2009.
- 942 32. Vivek K Singh, Laura Freeman, Bruno Lepri, and Alex Sandy Pentland. Classifying
 943 spending behavior using socio-mobile data. *HUMAN*, 2(2):pp–99, 2013.
- 944 33. Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of
 945 predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
- 946 34. Stan Stalnaker. The Ven currency, 2013. <http://www.ven.vc>.
- 947 35. Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Fran-*
 948 *cisco)*, pages 1–34, 2000.
- 949 36. The White House. National Strategy for Trusted Identities in Cyberspace: Enhancing On-
 950 line Choice, Efficiency, Security, and Privacy. The White House, April 2011. Available on
 951 http://www.whitehouse.gov/sites/default/files/rss_viewer/NSTICstrategy_041511.pdf.
- 952 37. United States Environmental Protection Agency. Institutional Controls Bibliography.
 953 <http://www.epa.gov/superfund/policy/ic/guide/biblio.pdf>, December 2005.

- 954 38. United States Environmental Protection Agency. RCRA Corrective Action Institu-
955 tional Controls - glossary. [http://www.epa.gov/epawaste/hazard/correctiveaction/](http://www.epa.gov/epawaste/hazard/correctiveaction/resources/guidance/ics/glossary1.pdf)
956 [resources/guidance/ics/glossary1.pdf](http://www.epa.gov/epawaste/hazard/correctiveaction/resources/guidance/ics/glossary1.pdf), 2007.
- 957 39. United States Environmental Protection Agency. Institutional Controls: A Guide to Plan-
958 ning, Implementing, Maintaining, and Enforcing Institutional Controls at Contaminated
959 Sites. Technical Report OSWER 9355.0-89 EPA-540-R-09-001, EPA, December 2012.
- 960 40. Jessica Vitak, Paul Zube, Andrew Smock, Caleb T Carr, Nicole Ellison, and Cliff Lampe.
961 It's complicated: Facebook users' political participation in the 2008 election. *CyberPsy-*
962 *chology, behavior, and social networking*, 14(3):107–114, 2011.
- 963 41. World Economic Forum. Personal Data: The Emergence of a New
964 Asset Class, 2011. Available on [http://www.weforum.org/reports/](http://www.weforum.org/reports/personal-data-emergence-new-asset-class)
965 [personal-data-emergence-new-asset-class](http://www.weforum.org/reports/personal-data-emergence-new-asset-class).