accenture technology

Accenture Labs

The Ethics of Data Sharing: A guide to best practices and governance

High performance. Delivered.

Strategy | Consulting | Digital | Technology | Operations

With the growth of the digital economy, data sharing has become an essential business practice—whether between different groups within the same organization, between partners in larger platform endeavors, or even, as in growing open data movements, with the public. Sharing enables new insights from existing data, and lets organizations make full use of this core resource. But it also introduces new ethical risks. This paper presents a best-practice approach for data sharing, to ensure that ethics are properly considered throughout the process, and that risks are appropriately identified and mitigated.

Data is the lifeblood of organizations and broader business ecosystems. It helps companies thrive when the economy is up, succeed when the economy is stable, and maintain when the economy falters. Every organization should seek data-sharing best practices, and other resources to maximize the value of data while ensuring that sharing pays due consideration to ethical concerns.

Sharing or aggregating data generates myriad security, ethics, and privacy risks. These risks are inextricable from the scientific and economic power of big data and require attention from the leaders of any organization that engages in data sharing, aggregating, or analytics. For example, when Experian added property managers as a data aggregation source, it obtained informed consent from tenants prior to collecting data.[1]

Experian only uses rental payment data in an individual's credit scoring if it will improve credit scores, not lower them, which ensures that the benefits accrued to their scoring service do not pose undue risks to individuals.

Data sharing: at the heart of the digital economy

Technologist Stewart Brand is often misquoted claiming that "all information wants to be free." Although this catchy maxim has provided a useful ethical principle for open internet advocates, the rest of Brand's statement shows that it can more accurately be described as a paradox about the value of information as it moves through the world. One iteration of the longer quote reads: "Information wants to be free. Information also wants to be expensive. Information wants to be free because it has become so cheap to distribute, copy, and recombine—too cheap to meter. It wants to be expensive because it can be immeasurably valuable to the recipient. That tension will not go away."[2]

This fundamental tension between the value of information and the tendency of information to propagate inexpensively is at the heart of an aspect of business that is critically important in the digital economy: data sharing.

The open data movement, advocated by many governments and nonprofits, makes an effort to formalize and standardize methods for placing useful datasets into as many hands as can potentially make use of it. Data sharing is often framed in terms of these norms of open data, the unrestricted sharing of data with anyone. Open data advocates have articulated a strong case for unrestricted sharing in many settings, particularly in connection with publicly-funded scientific and governmental data.[3] The ethos of openness—often presented as a pure good—informs most discussions of the social, business, political and infrastructural dynamics of the internet.

However, these robust efforts at building norms of open data have left an important middle ground largely unaddressed. In a space between closed and open data, sharing occurs between a limited set of parties, such as between industry and selected non-profit members, or between government and vetted partners in private industry. Sharing is also often subject to limited durations, introducing complications that are largely absent in open data contexts. Where there are good reasons to share data in a limited fashion, but where legal, financial, and ethical reasons prevent open sharing, what norms should be used to guide such exchanges? How can the utility of data to do good be maximized when sharing the data with the whole world is undesirable or infeasible?

Why share?

The most distinct characteristic of "big data" is not simply the surging size of aggregate datasets, but the wealth of insights now available from it via advanced analytics. Sharing data helps to create richer multivariate datasets, which in turn allows for more significant insights to be extracted from data. The rise of inexpensive networked computing resources in the public cloud has enabled many organizations to run advanced analytics without investing in expensive infrastructure. Such resources allow them to store data indefinitely, move it

At its crux, big data is about sharing, where "sharing" implies putting data to multiple re-uses across different contexts in the hands of various teams or organizations.

unpredictably and re-analyze it repeatedly. The long-range benefit of these new technologies lies in the ability to share and merge data within or between organizations, generating tremendous economic, strategic, and humanitarian potential.

These opportunities must, however, be considered alongside ethical issues. Given the human, financial, and technical resources devoted to collecting and managing data, there's an ethical obligation to repurpose and share data in a manner that maximizes the good that can be achieved. There are also concerns: for example, focusing solely on the financial benefit of data sharing (to the exclusion of all else) can erode trust from partners or those providing the data. That could negate a tremendous opportunity for good. With proper consideration, however, both corporate and ethical opportunities can be achieved.

Ethical arguments for sharing data

Where the subjects of data have sacrificed resources or borne some risk to make that data available, the obligation to maximize its use is amplified. This argument is particularly strong if data is produced using public or nonprofit resources that are already targeted at furthering the common good (e.g. medical and scientific research). The longstanding debates about data sharing and open data in the sciences have shaped discussions around why and how to share other types of data. The National Institutes of Health (NIH) and the National Science Foundation (NSF) have, for a decade, had data-sharing policies requiring grant recipients to share data in public repositories. The NSF instituted a new requirement in 2011 that researchers must submit a supplementary Data Management Plan (DMP) with every grant, in part due to pressure from policymakers to ensure maximum social value is squeezed from taxpayer-funded datasets.[4,5]

Debates about data-sharing practices have been especially vibrant in genomics, a field that has led the pack in data-intensive biology ever since it arose from the race between public and private ventures to offer the first draft of the human genome.[6] There has long been a push and pull between activist scientists and patient advocates, who view maximally open and participatory genomic data as an ethical obligation to speed research into diseases, and government agencies that lean toward closed or tiered access policies for ethically sensitive datasets in accordance with their readings of human-subjects protections.[7,8] Research has indicated that open data policies do not currently have a significant effect on whether most human subjects decide to participate in specific research projects, but that people do prefer to have a sense of control over how data about them is used.[9]

The long-range benefit of these new technologies lies in the ability to share and merge data within or between organizations, generating tremendous economic, strategic, and humanitarian potential.

Ethical concerns in sharing data

Even in fields that have yet to emphasize data sharing as an obligation, there's a concerted push toward increased sharing and collaboration. Yet, the technical and ethical dynamics driving us toward ever more data sharing also set the stage for big data's ethical risks. If data can be reused indefinitely to discover unpredictable correlations, then controls meant to protect the subjects of datasets should ideally cover a future timespan (and the possible uses of data during that period). It's not correct to assume that "open," "shared" and "public" are synonymous with the "public good"—particularly when datasets are combined with others for purposes that could not have been predicted when the data was first collected. Those who generate and control public datasets should account for whose good is served by specific kinds of openness, and take responsibility for fostering equitable uses of their data. Researchers and practitioners have found any number of surprising correlations that can disclose sensitive information about persons in datasets available to the public.[10]

Sharing poses special structural challenges for ethical data practices because the familiar procedures for protecting people from abuse are physically and temporally removed from the situations in which data is most valuable. Consent, whether it is informed consent in a medical context or end-user license agreements for internet services, typically occurs as an obligatory passage point at the beginning of data collection (for an in-depth discussion of informed consent, see Informed Consent and Data in Motion). So, insofar as data is increasingly used outside the initial context of collection, and the process of consent and other human subjects protections are largely restrained to the time and context of collection, there's a troublesome gap between the tools of data analytics that rely on sharing and the tools used to protect subjects from harms that may be caused by sharing.

For example, a Freedom of Information Act (FOIA) request allowed a programmer to receive and publicize the entire anonymized dataset of the New York City Taxi Commission's trip records.[11] Such a dataset could be applied to any number of highly useful research projects about civic planning and transportation infrastructure. However, from this dataset and in relation to auxiliary datasets, analysts could also determine the likely religion of certain cab drivers, which rides were taken by celebrities and how much they tipped, and the likely identity of individuals frequenting strip clubs. They could also de-anonymize the names of drivers based on medallion numbers, which could then be correlated with other private details, including income.[12,13,14,15] Although some attributed this incident entirely to the dataset being poorly anonymized, it fits a well-established pattern: anonymization is either not meaningfully protective or may not be technically viable in common circumstances.[16]

Anonymization is either not meaningfully protective or may not be technically viable in common circumstances.

Data sharing also amplifies subtle and complex questions of interpretation, transparency, collaboration and trust that are at the heart of data ethics concerns. Contrary to the common discourse of data as a neutral arbiter of truth, data requires significant interpretive work to create useful knowledge.[17] How data is collected and how datasets are architected subtly shapes what the data "means" down the line, often in ways that are non-obvious and prone to bias.

Data mining can replicate systemic inequality by "discovering" patterns of bias that were there all along but not accounted for within the database architecture.[18] What looks like a neutral, data-driven decision free of human bias can actually be a reflection of longstanding inequality, and without deep inquiry it's hard to tell when this is the case. Some scholars have called predictions about the future behavior of a person or a group based on data that reflects an unequal social structure and/or inaccurate data a "predictive harm."[19] The subtle ways in which bias can be built into data are amplified by machine-learning techniques that enable "mutation machine" algorithms to restructure and interpret databases with minimal human input.

As these examples and potential pitfalls demonstrate, it's problematic to treat sharing merely as a matter of lobbing self-contained databases back and forth between partners. Those that are closest to the creation and maintenance of a dataset may have a good sense of its archaeology; those that are further removed are more likely to take a database at face value. Furthermore, data sharing does not merely imply sharing of datasets—rather, it typically also includes the sharing of interpretive resources that make data analytics possible, such as methods and models. In other words, data sharing is not simply the sharing of data, it's also the sharing of interpretation. Sharing therefore expands obligations to account for the possibility of implicit and explicit biases, and the downstream harms that can come from them. Requiring substantial commitments to ongoing collaboration, partners should be mutually accountable for how they handle and interpret databases in order to protect against interpretive harms. As a recent report from the Centre for Information Policy Leadership indicated, new models of accountability and transparency within and between organizations, to protect individuals and communities from such harms, will be a necessary component of a sustainable information economy.[20]

In other words, data sharing is not simply the sharing of data, it's also the sharing of interpretation.

Data sharing taxonomy and practices

Getting a handle on the specific ethical risks faced by an organization requires some attention to where their activities fall on the map of data-sharing practices. This can be challenging. While much attention has been paid to open data as an ethos, with well-defined standards and norms, data sharing encompasses a wide diversity of practices and lacks any equivalent coherence. The following discussion examines a taxonomy for data-sharing practices, with an eye to the ethical nuances involved at different levels of openness.

Although the central motivations are similar, the practices and standards of data sharing and open data can differ significantly. A recent exchange in *The Guardian* newspaper highlighted this. In an editorial, *The Guardian* called attention to the privacy consequences of the UK National Health Service's (NHS) plan to centralize and sell large amounts of citizens' personal medical data to health management and pharmaceutical companies as private partners.[21] The editorial encouraged the UK government to attend more rigorously to the ways in which open data policies can cause harm. Other critics noted that the initiative was deeply problematic because the NHS provided citizens with a very short window to opt out of this plan for sharing their own medical data, and to disallow future removal of either their own data or their children's data if they changed their minds.[22] *The Guardian's* editorial called this a "smash and grab scheme" that proved the necessity of moving slowly on open data initiatives.

Jeni Tenninson of the Open Data Institute responded in an open letter, arguing that, far from an example of open data policy, the NHS initiative was rather data sharing gone awry and misleadingly categorized as open data by the NHS.[23] Were it an example of open data policies, Tenninson argued, these medical datasets would have been shared as a public good and thoroughly anonymized. This case illustrates some of the critical ethical dynamics that must be addressed in data-sharing contracts between a limited set of actors. Should access be restricted for those not paying for the data? Will data be used to further the general public good, and is the public good better served by a different arrangement? Are ethical obligations, such as gaining genuine consent and offering opt-outs, upheld in both spirit and practice?

This NHS case also illustrates the ways in which open data has a clear ethos that defines the goals and prescribes ethical actions, often linked to notions that we consider intrinsically good, such as open and democratic governance.[24] Sharing data in a narrower context does not have a well-established ethos that can justify and order data-sharing efforts. However, it's important to note that data sharing between a limited set of organizations, even if it includes monetized datasets or is not fully open, does not *preclude* having rigorous ethical conditions. The fact that such arrangements currently exist without those conditions is merely an indication that such conditions have not yet been defined and promulgated.

Conversations about ethical data sharing in enterprises and private-public partnerships have begun to surface. The Governance Lab (GovLab) at New York University has offered the term "data collaborative" to define the middle ground between open government datasets and closed corporate data.[25] "The term data collaborative refers to a new *form* of collaboration, beyond the public-private partnership model, in which participants from different sectors (including private companies, research institutions, and government agencies) can exchange data to help solve public problems."[26,27] Framing data sharing as a "collaborative" turns attention toward the platform dynamics of sharing. *Open* or *closed* or *shared*

ultimately refers to the status of a dataset, not what is done with it. GovLab's terminology is particularly useful for focusing on *how* data gets shared to solve problems, rather than beginning and ending the conversation at the dataset's status as open or closed. They identify the strengths of private-public collaborations at a platform level as:

- **Data-driven decision-making** that invites all actors with useful data to participate
- **Information exchange and coordination** between trusted parties and
- **Shared standards and frameworks to enable multi-actor, multi-sector participation** that enables efficient sharing across a variety of sources and users.

Noting that there's scant public discussion of how to effectively facilitate data sharing, GovLab has published requests for private partners to share cases of data sharing done well and done poorly. Stefan Verhulst of GovLab also sketches out a taxonomy of data collaboratives, identifying six types:

1. **Research partnerships,** in which corporations share data with universities and other research organizations.
2. **Prizes and challenges,** in which companies make data available to qualified applicants who compete to develop new apps or discover innovative uses for the data.
3. **Trusted intermediaries,** where companies share data with a limited number of known partners.
4. **Application programming interfaces (APIs),** which allow developers and others to access data for testing, product development, and data analytics.
5. **Intelligence products,** where companies share (often aggregated) data that provides general insight into market conditions, customer demographic information, or other broad trends.
6. **Corporate data cooperatives or pooling,** in which corporations—and other important dataholders such as government agencies—group together to create "collaborative databases" with shared data resources.[28]

Verhulst notes the need for substantive research on the value proposition for data sharing and the methods that could be used to track and mitigate harm.

In July 2014, the Responsible Data Forum co-hosted with UN Global Pulse and the Data & Society Research Institute a workshop on private data sharing for the public good. Participants discussed experiences negotiating and facilitating the release of corporate data for use in public and nonprofit contexts. Among their outputs was a Data Risk Checker wiki as a general tool for assessing risks posed by sharing. They suggest that efforts to check and mitigate risk proceed through a series of steps that consider economic, physical, and psycho-social/emotional harms:

1. Identify the persons at risk in the event of exposure.
2. Identify knowledge assets that can be extracted from the data collected.
3. Evaluate the importance of each knowledge asset to the campaign.
4. For each type of harm, evaluate probability and severity of harm.

Matt Stempeck, Microsoft's Director of Civic Technology, frames corporate data sharing as a matter of philanthropy that ultimately expands business opportunities.[29] He writes, "Data philanthropy achieves many of the goals sought by traditional corporate social philanthropy: it allows companies to give back in a way that produces meaningful impact, and reflects the business's core competencies while preserving or expanding value for shareholders." He also sketches a handful of core ethical questions, such as accounting carefully for who could benefit most from the information and what formats they can make use of, stripping data of anything identifiable and considering whether data is best shared via real-time API or in occasional bulk releases.

The World Health Organization (WHO) held a meeting in September 2015 to discuss developing norms for data sharing during public health emergencies, which was prompted by the previous summer's Ebola outbreaks. Their report details consensus solutions to the concerns that led to slow or reduced sharing of data during the outbreaks.[30] Common among these concerns were fears that sharing pre-publication data would make journals less likely to publish research and that intellectual property owned by pharmaceutical companies could be compromised. By getting commitments from all stakeholders to accommodate the needs of other ecosystem members and focus on a central, urgent public health task, the WHO was able to develop early stages of what is essentially a coordinated data-sharing platform. The WHO's effort is echoed by collaborative data sharing between public and private entities to develop global climate resilience and further pharmaceutical research for drugs important to public health.[31,32]

Open data scholar, Tim Davies, has identified the need for standardized contracting processes—not standardized contracts—as a major impediment to data-sharing agreements.[33] The specifics of any contract will vary significantly according to the parameters of the datasets, the ethical risks posed by the data, and the needs of the partners. It's therefore fairly fruitless to seek standardized contracts. Instead, the most promising route to facilitate data sharing is to develop processes that can be replicated and common standards that can be propagated, such as ensuring that data is machine-readable and datasets utilize common architectures. He also notes that in the history of open data, the development of standards and practices is often linked to a significant rethinking of information infrastructures—nations that adopt open data standards often end up significantly improving their backend infrastructures as a result of the deliberate changes required to meet the new standards.

Risk management approach to data sharing and privacy considerations

Executives have long used risk management to make strategic and operational decisions that positively impact the long- and short-term success of their organization. Risk management has become part of compliance law and is recognized as a standard practice; it is also increasingly applied to questions of data privacy, protection, and sharing. Organizations are thus well-advised to incorporate data sharing into their existing risk management frameworks.

The Centre for Information Policy Leadership has advocated the application to data privacy of risk management practices accepted in other areas of decision-making; it has also provided suggestions for a more standardized approach to risk management in this area. In its paper, *The Role of Risk Management in Data Protection*, the Centre concludes that risk management is "a valuable tool for calibrating the implementation of and compliance with privacy requirements, prioritizing action, raising and informing awareness about risks (and) identifying appropriate mitigation measures."[34]

To achieve successful risk management practices in the context of data-sharing contracts, it's necessary to operate with a shared classification or taxonomy of risks. One example of such an approach is the matrix provided by the Centre in an earlier paper, *A Risk-based Approach to Privacy: Improving Effectiveness in Practice*. This delineates a list of risk categories so that those engaging in risk assessment can ensure they're considering generally accepted relevant concerns.[35] While the matrix provided by the Centre is in draft form and does not purport to be a complete classification, it offers a helpful tool for organizations to customize to their own specific situation. (continued on page 13)

Figure 1: Sample risk assessment framework (credit: Centre for Innovation Policy Leadership)

DRAFT: Risk Matrix

	Unjustifiable Collection			Inappropriate Use			Security Breach			Aggregate
				Inaccuracies Not expected by individual Viewed as unreasonable Viewed as unjustifiable			Lost data Stolen data Access Violation			
Risks	**Likely**	**Serious**	**Score**	**Likely**	**Serious**	**Score**	**Likely**	**Serious**	**Score**	**Risk Rank**
Tangible Harm										
Bodily harm	0	0	0	0	0	0	0	0	0	0
Loss of liberty or freedom	0	0	0	0	0	0	0	0	0	0
Financial loss	0	0	0	0	0	0	0	0	0	0
Other tangible loss	0	0	0	0	0	0	0	0	0	0
Intangible Distress										
Excessive surveillance	0	0	0	0	0	0	0	0	0	0
Suppress free speech	0	0	0	0	0	0	0	0	0	0
Suppress associations	0	0	0	0	0	0	0	0	0	0
Embarrassment/ anxiety	0	0	0	0	0	0	0	0	0	0
Discrimination	0	0	0	0	0	0	0	0	0	0
Excessive state power	0	0	0	0	0	0	0	0	0	0
Loss of social trust	0	0	0	0	0	0	0	0	0	0

LEGEND:

Rank "Likely" from 10 (high) to 1 (low) based on the highest score for any component

Rank "Serious" from 10 (high) to 1 (low) based on the highest score for any component

AGGREGATE RISK RATE:

Highest score is 300

Lowest score is 0

Levels of sharing and ethical obligations

To ask the right questions about data sharing, it's imperative to know the extent of the data supply chain and how far shared data could reach. Discussed in the main body of this paper, the middle tier—data sharing amongst partners (see table below)—is both the most complicated and the least discussed. The "Entity" tier can largely be handled by internal procedures that may already exist in most organizations; if procedures do not exist, there's already significant literature available on the subject to aid with their development. The "Open" tier has well-articulated standards and norms. Therefore, our recommendations here focus on best practices to improve governance within the "Shared" tier.

Entity-level Data

- level 1: insular within team
- level 2: kept within the division
- level 3: kept within the organization

Shared Data

- level 4: shared with alliance partners
- level 5: shared with partners
- level 6: shared under contract with an exchange of value

Open Data

- level 7: semi-open data: tiered levels of access for vetted users
- level 8: open data, available to all but requires transformation
- level 9: open data

Guidelines for sharing data among external partners:

1. Ongoing collaboration and mutual accountability are necessary between data-sharing partners. Datasets are not static or neutral—they can reflect bias and require ongoing interpretive work. Partners should be accountable to each other for sensitive, high-quality interpretive work that seeks to address possible bias and potential harms. Recording decisions that affect analysis can help with transparency among partners.

2. Build common contracting procedures, but treat every contract and dataset as unique. The wide variety of possible uses and potential harms for each dataset suggests that there is no "one size fits all" universal criteria for ethical data sharing and governance. Rather, standardized contracting procedures can build community norms about what review processes are necessary complements to data sharing.

3. Develop ethical review procedures between partners. Partners should determine in advance how ethics concerns can be escalated and resolved both within their own organizations and between their organizations. Everyone handling and interpreting the data should be aware of these procedures.

4. Be mutually accountable for interpretive resources. When methods and models are developed for machine-learning systems and other modes of data analytics, assumptions should be enumerated. This includes hypothesis generation, initial experiment results, and training datasets. Each step of the data transformations should be enumerated.

5. Minimalist approaches to data sharing are preferable. Although there is a general obligation to make the best use of data through sharing, it's also the case that increased levels of openness generally increase risk to data originators and each partner. This is particularly true where the subjects of the data have a reasonable expectation of privacy. Data holders and recipients should carefully audit the datasets for such risks before sharing all or some of the data under consideration. The terms of the relevant contract should be explicit about the value of data to each participant organization.

6. Identify potential risks of sharing data within sharing agreements. The goal of this is to learn how to more accurately identify risks. The assumption is that identified risks at the outset of a sharing relationship will not sufficiently describe the universe of risks, and practitioners will learn to be more systemic in risk management through experience.

7. Repurposed data requires special attention. The hardest type of harm to predict and mitigate is that which can result from future repurposing of data. Data that appears innocuous in one context may potentially be very damaging when combined with other datasets. Data-sharing partners should have explicit agreements on the parameters of repurposing.

8. When ethical principles or regulations are unclear, emphasize process and transparency. Particularly as the industry matures, data analytics will often operate in gray areas without much—if any—precedent. Where there is no existing industry standard, transparent and consistent decision-making processes are the best buffer against harms and a way to reinforce public trust.

9. Published research requires additional attention. As tech and social networking firms build ever-bigger datasets, data scientists inside industry are increasingly seeking routes for publishing generalized scientific knowledge derived from their datasets. If research is to be published, all involved parties should agree to the publication in advance, and ensure they've undertaken reasonable attempts to protect the data-subjects from harm and obtained their informed consent.

10. Treat trust as a networked phenomenon. When drafting documents such as terms of service agreements, privacy policies, and end-user license agreements, be sensitive to the tensions between legal compliance and trust with your users, other partners and the public. The document that puts your organization in the safest legal position may not build the strongest trust with other parties. Users deserve a document that is clearly understandable and helps them to protect themselves and their own interests.

(continued from page 9)

Governance models

The ethical dynamics of data sharing involve both open-ended problems discussed above—what are the risks of repurposing sensitive data—and organizational challenges—how should an organization account for and mitigate those risks? The organizational challenges have sparked new conversations about ethics governance in both industry and nonprofits. Many organizations are discovering that the peculiar challenges posed by the acquisition, repurposing, and sharing of data stretch the capacity of existing compliance processes: data ethics requires organizations to adopt responsive, ongoing, and collaborative governance. To maximize utility while ensuring ethical practices, organizations must build on new guidelines that consider the specific challenges of data sharing.

As shared data norms begin to reach more sectors, organizations of all kinds should be aware that ensuring shared data does good in the world requires careful attention to governance. For example, social media companies including Twitter and Facebook have begun to offer outside computer science and social science academics limited access to their data; enabling that access has required significant efforts in the development of contracts and governance. Conducting rigorous data-driven research on social media has always posed dilemmas for academics because social media companies typically have rules governing the use of API's that run contrary to norms of peer-review, such as Twitter limiting data-scraping to the previous seven days and not allowing sharing of entire research datasets. Twitter offered a single round of "data grants" in 2014, which were functionally subsidized access to their historical data troves for academics, available through a data reseller program. Yahoo recently established a method for safely sharing large amounts of anonymized login data with security researchers.[36] Meanwhile Facebook has focused on developing partnerships with academics through in-house scholarships for students and faculty, enabling researchers to have secured access to their datasets.

Recent scholarship on this matter has illustrated some of the basic elements of data ethics governance. These are instructive as a starting point. The core themes that emerge include the need for independent review, some degree of transparency or public-facing elements, an emphasis on collaboration, reciprocal trust between organizations, and chains of accountability. We recommend that anyone seeking to build data ethics processes consider these suggestions.

Legal theorist Ryan Calo has proposed the notion of "Consumer Subjects Review Boards" (CSRB) to guide ethical decision-making inside big data enterprises. Calo argues that the fundamental

As shared data norms begin to reach more sectors, organizations of all kinds should be aware that ensuring shared data does good in the world requires careful attention to governance.

ethical challenge for big data is that corporate bodies have vastly superior capabilities than consumers. Noting that the power dynamics of big data have driven conversation away from privacy alone and toward the broad, basic principles of data and information ethics, he suggests that review boards modeled on university-based Institutional Review Boards (IRB) but scaled to the business cycle have the potential to redress some of this disparate power balance. Both the White House and the Federal Trade Commission have flagged CSRBs as a potentially important idea for future regulation.[38,39]

Data ethicists Jules Polonetsky, Omar Tene and Joseph Jerome have considered some of the legal and structural dynamics of CSRBs.[40] They emphasize the importance of establishing founding principles for CSRBs, just as the Belmont Report[41] historically established principles for IRBs that have served as the core principles of research ethics, particularly in biomedicine. Establishing broad principles for engagement with ethical issues focuses discussions and provides at least a degree of common ground for the parties involved. To the Belmont Report's set of Respect for Persons, Beneficence and Justice, they add Respect for Law. Additionally, they emphasize that despite the many complaints that academics level against IRBs, having a reliable and largely transparent system of ethics review has provided the highly valuable intangible benefit of fostering community trust in the system as a whole. Without this, human-subjects medical research could be largely stymied.

Polonetsky *et al* advocate a two-track model for review boards that has one internal board (with at least one accountable executive) and one external board of experts and community representatives. Internal and external ethics advice can address different sorts of problems, have different levels of access to sensitive business information, and operate on different time-scales. They also define best practices for maintaining independence and credibility, including adequate funding and resources that are not dependent on particular outcomes.

Dove *et al* recently identified the core characteristics of successful ethics review processes in international data-intensive research.[42] While the details might differ significantly in enterprise data-sharing, these principles are still a useful guide when developing a governance plan. They identify reciprocity (collaborating entities accept each other's judgements), delegation (collaborating entities delegate responsibilities in advance), and federation (ethics review is done centrally when possible) as organizational priorities when establishing ethics review processes in contexts where entities are

operating across diverse regulatory ethical regimes.

Facebook is publicly discussing its internal ethics review practices, marking them as the first major tech company to build a review process analogous to IRB's.[43] Although the particularities of Facebook's process would not be universal, there are some general governance lessons that organizations will likely find useful. Most importantly, their review process is baked into the engineering workflow through their internal task management system. They have identified common flags for projects that require more review due to potential risks to users or the general public, and empower project managers to determine whether projects need to be flagged. Perhaps their most important decision was to ignore the traditional distinction between research for publication and product development. Historically, the former has been far more likely to receive ethics review than the latter, and Facebook's decision to subject them to the same scrutiny is an important departure from the norm.

If a company depends on the long-term trust of its users and the public at large, some variety of ethics review process will likely be a condition of doing business in the data/tech sector, particularly if the company engages in the sort of machine learning and data analytics research that provokes public concern.[44] The sphere of research ethics regulation has largely been oriented toward biomedicine and there's little reason to expect that data or tech companies will fall under its remit. However, that does not mean the purpose of ethical review is not shared across sectors.[45] Indeed, a number of tech leaders have begun to adopt review boards as core business strategy[46], such as Google's "right to be forgotten board"[47] and AI ethics board[48], Palantir's Council on Privacy and Civil Liberties[49], and Facebook's efforts to develop internal controls on its research agenda[50].

To end where we began, "big data" analytics undeniably expands the power of data in science and business. And data sharing undeniably expands the power of data analytics by providing a broader set of datapoints for mining insights. But data analytics in general, and data sharing in particular, pose unique and largely unfamiliar ethical and governance challenges. While there are good ethical reasons to share in the long-run, the consequences of sharing in any given case are unpredictable and possibly harmful.

Navigating such challenges requires the development of flexible, collaborative governance structures that enable organizations to make consistent, actionable decisions about uncertain outcomes. Those that succeed in this objective stand to gain significant advantages – both in business benefits and in securing and sustaining building public trust.

Data analytics in general, and data sharing in particular, pose unique and largely unfamiliar ethical and governance challenges.

100/365-day plans

To enhance best practices and governance for your organization, over the next three months, a cross-functional team should:

- Learn about the governance structures your organization already has in place. Search for governance boards involving ethics and/or data.
- Take inventory of which datasets you have implicated in data-sharing agreements.
- Catalog any existing governance procedures.
- Audit your data-sharing contracts for compliance with the self-imposed guidelines identified above (top-down audit).

- Audit instances where data has been shared and highlight instances where data was shared without data-specific contracts between in place (bottom-up audit).
- To work with "dirty data," create policies requiring data originators to opt-in after being notified of the purpose of the research. These agreements should be shared with a third-party governance body.
- Create a "dirty data" list describing the types of data that should always be removed from data sets (e.g., National Identification Number).

Within a year, the following actions could be taken:

- Ensure that all executives, data practitioners, and project, program, and product managers are aware of the procedures for escalating ethics concerns about data-sharing practices.
- Identify relevant industry or non-profit bodies which are guiding data-sharing practices and determine how your organization can contribute and benefit.
- After learning about governance structures already in place, find an opportunity to improve and propose new governance initiatives that would be appropriate for your organization.
- Construct training courses for data and legal professionals to ensure a common set of best practices.
- Establish a data ethics committee that will periodically review sharing practices and policies, and to be available for consultation when dilemmas arise.
- Implement a "Risk Management" plan. Utilize an accepted taxonomy for risk assessment such as the one provided by *The Centre for Information Policy Leadership* (see above).

References

1. Haller, E. (2016, September 22). Executive Vice President & Global Head, Experian DataLabs. (S. Tiell, Interviewer)

2. The Media Lab: Inventing the Future at M. I. T. (1988) Reprint edition. New York, N.Y., U.S.A: Penguin Books, pg 202.

3. For example, see GovLab's recent Open Data Impact report: http://odimpact.org/.

4. Metcalf J (2016) Data Management Plan: A Background Report. Council for Big Data, Ethics, and Society. Available from: http://bdes.datasociety.net/council-output/data-management-plan-a-background-report/ (accessed 25 January 2016).

5. Compliance rates with DMPs among NSF grant recipients has questioned widely, leading the NSF to consider revising and strengthening the requirement.

6. See: Wade N (2002) Life Script: How the Human Genome Discoveries Will Transform Medicine and Enhance Your Health. Simon and Schuster; Reardon J (2004) Race to the Finish: Identity and Governance in an Age of Genomics. Princeton: Princeton University Press.

7. Ball MP, Bobe JR, Chou MF, et al. (2014) Harvard Personal Genome Project: lessons from participatory public research. Genome Medicine 6(2): 10.

8. Paltoo DN, Rodriguez LL, Feolo M, et al. (2014) Data use under the NIH GWAS Data Sharing Policy and future directions. Nature Genetics 46(9): 934–938.

9. Cummings JA, Zagrodney JM and Day TE (2015) Impact of Open Data Policies on Consent to Participate in Human Subjects Research: Discrepancies between Participant Action and Reported Concerns. PLoS ONE 10(5): e0125208; McGuire AL, Basford M, Dressler LG, et al. (2011) Ethical and practical challenges of sharing data from genome-wide association studies: The eMERGE Consortium experience. Genome Research 21(7): 1001–1007.

10. Such concerns have been discussed in Nissenbaum, H. Privacy in Context: Technology, Policy, and the Integrity of Social Life, (Cambridge: MIT Press), and even prior to that in Helen Nissenbaum, "Protecting Privacy in an Information Age: The Problem of Privacy in Public," Law and Philosophy, 17: 559-596, 1998.

11. http://www.andresmh.com/nyctaxitrips/

12. Franceschi-Bicchierai L (2015) Finding Muslim NYC Cabbies in Trip Data. Mashable. Available from: http://mashable.com/2015/01/28/redditor-muslim-cab-drivers/#0_uMsT8dnPqP (accessed 6 November 2015).

13. Tockar A. Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset. Neustar Research. Available from: http://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/ (accessed 6 November 2015).

14. ibid

15. Pandurangan V (2014) On Taxis and Rainbows : Lessons from NYC's improperly anonymized taxi logs. Medium. Available from: https://medium.com/@vijayp/of-taxis-and-rainbows-f6bc289679a1 (accessed 10 November 2015).

16. For some scholarly work on the limits of anonymization, see: Ohm P (2009) Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. Available from: http://papers.ssrn.com/abstract=1450006 (accessed 13 November 2015); and de Montjoye Y-A, Hidalgo CA, Verleysen M, et al. (2013) Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports 3. Available from: http://www.nature.com/articles/srep01376 (accessed 13 November 2015); and Montjoye Y-A de, Radaelli L, Singh VK, et al. (2015) Unique in the shopping mall: On the re-identifiability of credit card metadata. Science 347(6221): 536–539.

17. Bowker GC (2005) Memory practices in the sciences. Mit Press Cambridge, MA. Available from: http://www.peterasaro.com/writing/bowker1.pdf (accessed 6 November 2015); Gitelman L (ed.) (2013) Raw Data Is an Oxymoron. Cambridge, MA: MIT Press.

18. Barocas S and Selbst AD (2015) Big Data's Disparate Impact. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. Available from: http://papers.ssrn.com/abstract=2477899 (accessed 9 February 2016).

19. Crawford K and Schultz J (2014) Big data and due process: Toward a framework to redress12 ¬¬ predictive privacy harms. BCL Rev. 55: 93.

20. Centre for Information Policy Leadership (2015) The Role of Enhanced Accountability in Creating a Sustainable Data-driven Economy and Information Society. Discussion Draft, Hunton & Williams LLP. Available from: https://www.informationpolicycentre.com/files/Uploads/Documents/Centre/World_of_Big_Data_Accountability_and_Digital_Responsibility_Sustainable_Data-Driven_Economy_and_Information_Society.pdf (accessed 10 March 2016).

21. Guardian Editorial Board (2014) Open data: slow down. The Guardian, 18th April. Available from: http://www.theguardian.com/commentisfree/2014/apr/18/open-data-whitehall-government-editorial (accessed 25 January 2016).

22. Anderson R (n.d.) Opting out of the latest NHS data grab. Light Blue Touchpaper: Security Research, Computer Laboratory, University of Cambridge. Available from: https://www.lightbluetouchpaper.org/2014/01/08/opting-out-of-the-latest-nhs-data-grab/ (accessed 25 January 2016).

23. Tenninson J (2014) Open data is a public good. It should not be confused with data sharing. The Guardian, 12th May. Available from: http://www.theguardian.com/commentisfree/2014/may/12/response-confuse-open-data-sharing-government (accessed 26 January 2016).

24. Davies TG and Bawa ZA (2012) The Promises and Perils of Open Government Data (OGD). The Journal of Community Informatics 8(2). Available from: http://www.ci-journal.net/index.php/ciej/article/view/929 (accessed 26 January 2016).

25. Noveck BS (2015) Data Collaboratives: Sharing Public Data in Private Hands for Social Good. Forbes. Available from: http://www.forbes.com/sites/bethsimonenoveck/2015/09/24/private-data-sharing-for-public-good/ (accessed 26 January 2016).

26. Verhulst SG (2015) Data Collaboratives: Exchanging Data to Improve People's Lives. Medium. Available from: https://medium.com/@sverhulst/data-collaboratives-exchanging-data-to-improve-people-s-lives-d0fcfc1bdd9a (accessed 26 January 2016).

27. For an extended reading list about data collaboratives hosted by GovLab, see http://thegovlab.org/the-govlab-selected-readings-on-data-collaboratives/.

28. Verhulst SG (2015) Mapping the Next Frontier of Open Data: Corporate Data Sharing: Taxonomy of current corporate data sharing practices, Mapping the Next Frontier. Medium. Available from: https://medium.com/internet-monitor-2014-data-and-privacy/mapping-the-next-frontier-of-open-data-corporate-data-sharing-73b2143878d2#.hn4u9id6a (accessed 26 January 2016). Originally published in Gasser U, Zittrain JL, Faris R, et al. (2014) Internet Monitor 2014: Reflections on the Digital World: Platforms, Policy, Privacy, and Public Discourse. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. Available from: http://papers.ssrn.com/abstract=2538813 (accessed 26 January 2016).

29. Stempeck M (2014) Sharing Data Is a Form of Corporate Philanthropy. Harvard Business Review. Available from: https://hbr.org/2014/07/sharing-data-is-a-form-of-corporate-philanthropy (accessed 26 January 2016).

30. World Health Organization (2015) WHO | Developing global norms for sharing data and results during public health emergencies. Geneva. Available from: http://www.who.int/medicines/ebola-treatment/blueprint_phe_data-share-results/en/ (accessed 26 January 2016).

31. The White House (2014) FACT SHEET: President Obama Announces New Actions To Strengthen Global Resilience To Climate Change And Launches Partnerships To Cut Carbon Pollution. whitehouse.gov. Available from: https://www.whitehouse.gov/the-press-office/2014/09/23/fact-sheet-president-obama-announces-new-actions-strengthen-global-resil (accessed 27 January 2016).

32. Reardon S (2014) Pharma firms join NIH on drug development. Nature. Available from: http://www.nature.com/doifinder/10.1038/nature.2014.14672 (accessed 27 January 2016).

33. Davies T (2014) Unpacking Open Data: Power, Politics and the Influence of Infrastructures. Colloquium, Bermkan Center for Internet and Society, Harvard University. Available from: https://cyber.law.harvard.edu/interactive/events/luncheon/2014/11/davies (accessed 20 January 2015).

34. Bellamy, Bojana et al (2014) The Role of Risk Management in Data Protection The Centre for Information Policy Leadership

35. Bellamy, Bojana et al (2014) A Risk-based Approach to Privacy: Improving Effectiveness in Practice, The Centre for Information Policy Leadership

36. Spice B (2016) Carnegie Mellon, Stanford Researchers Devise Method to Share Password Data Safely: Yahoo! Releases Password Statistics of 70 Million Users For Cybersecurity Studies. Carnegie Mellon News, 22nd February. Available from: http://www.cmu.edu/news/stories/archives/2016/february/sharing-password-data.html.

37. Calo R (2013) Consumer Subject Review Boards: A Thought Experiment. Stanford Law Review Online 66: 97.

38. Executive Office of the President (2014) Big Data: Seizing Opportunities, Preserving Values. The White House. Available from: https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf (accessed 10 November 2015).

39. Federal Trade Commission (2015) Consumer Privacy Bill of Rights (Administration Discussion Draft 2015). Available from: https://www.whitehouse.gov/sites/default/files/omb/legislative/letters/cpbrGactGofG 2015GdiscussionGdraft.pdf.

40. Polonetsky J, Tene O and Jerome J (2015) Beyond the Common Rule: Ethical Structures for Data Research in Non-Academic Settings. Colorado Technology Law Journal 13.

41. National Commission for the Protection of Human Subjects, of Biomedical and Behavioral Research and The National Commission for the Protection of Human Subjects (1979) The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Available from: http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html (accessed 21 October 2015).

42. Dove ES, Townend D, Meslin EM, et al. (2016) Ethics review for international data-intensive research. Science 351(6280): 1399–1400.

43. Jackman M and Kanerva L (2016) Evolving the IRB: Building Robust Review for Industry Research. Washington and Lee Law Review Online 72(3): 442.

44. boyd danah and Crawford K (2012) Critical Questions for Big Data. Information, Communication & Society 15(5): 662–679.

45. Metcalf J (2015) Getting the formula right: Social trust, A/B testing and research ethics. Ethical Resolve. Available from: http://www.ethicalresolve.com/ (accessed 11 December 2015).

46. Havens JC (2015) Creating a code of ethics for artificial intelligence. Mashable. Available from: http://mashable.com/2015/10/03/ethics-artificial-intelligence/ (accessed 11 December 2015).

47. Google (2014) The Advisory Council to Google on the Right to be Forgotten. Available from: https://www.google.com/advisorycouncil/ (accessed 11 December 2015).

48. Sellinger E and Lin P (2014) Inside Google's Mysterious Ethics Board. Forbes. Available from: http://www.forbes.com/sites/privacynotice/2014/02/03/inside-googles-mysterious-ethics-board/ (accessed 11 December 2015).

49. Grant J (2012) Announcing the Palantir Council on Privacy and Civil Liberties. Palantir. Available from: https://palantir.com/2012/11/announcing-the-palantir-council-on-privacy-and-civil-liberties (accessed 11 December 2015).

50. Schroepfer M (2014) Research at Facebook. Facebook Newsroom. Available from: http://newsroom.fb.com/news/2014/10/research-at-facebook/ (accessed 11 December 2015).

Contact Us

Steven C. Tiell
Senior Principal—Digital Ethics
Accenture Labs
steven.c.tiell@accenture.com

Jacob Metcalf
Ethical Resolve
jake@ethicalresolve.com

Data Ethics Research Initiative

Launched by Accenture's Technology Vision team, the Data Ethics Research Initiative brings together leading thinkers and researchers from Accenture Labs and over a dozen external organizations to explore the most pertinent issues of data ethics in the digital economy. The goal of this research initiative is to outline strategic guidelines and tactical actions businesses, government agencies, and NGOs can take to adopt ethical practices throughout their data supply chains.

About Accenture Labs

Accenture Labs invents the future for Accenture, our clients and the market. Focused on solving critical business problems with advanced technology, Accenture Labs brings fresh insights and innovations to our clients, helping them capitalize on dramatic changes in technology, business and society. Our dedicated team of technologists and researchers work with leaders across the company to invest in, incubate and deliver breakthrough ideas and solutions that help our clients create new sources of business advantage.

Accenture Labs is located in seven key research hubs around the world: Silicon Valley, CA; Sophia Antipolis, France; Arlington, Virginia; Beijing, China; Bangalore, India, Dublin, Ireland, and Tel Aviv, Israel. The Labs collaborates extensively with Accenture's network of nearly 400 innovation centers, studios and centers of excellence located in 92 cities and 35 countries globally to deliver cutting-edge research, insights and solutions to clients where they operate and live. For more information, please visit www.accenture.com/labs

About Accenture

Accenture is a leading global professional services company, providing a broad range of services and solutions in strategy, consulting, digital, technology and operations. Combining unmatched experience and specialized skills across more than 40 industries and all business functions – underpinned by the world's largest delivery network – Accenture works at the intersection of business and technology to help clients improve their performance and create sustainable value for their stakeholders. With approximately 384,000 people serving clients in more than 120 countries, Accenture drives innovation to improve the way the world works and lives. Visit us at www.accenture.com.

Learn more: www.accenture.com/DataEthics