DATA STANDARDIZATION

MICHAL S. GAL† AND DANIEL L. RUBINFELD‡

With data rapidly becoming the lifeblood of the global economy, the ability to improve its use significantly affects both social and private welfare. Data standardization is key to facilitating and improving the use of data when data portability and interoperability are needed. Absent data standardization, a "Tower of Babel" of different databases may be created, limiting synergetic knowledge production. Based on interviews with data scientists, this Article identifies three main technological obstacles to data portability and interoperability: metadata uncertainties, data transfer obstacles, and missing data. It then explains how data standardization can remove at least some of these obstacles and lead to smoother data flows and better machine learning. The Article then identifies and analyzes additional effects of data standardization. As shown, data standardization has the potential to support a competitive and distributed data collection ecosystem and lead to easier policing in cases where rights are infringed or unjustified harms are created by data-fed algorithms. At the same time, increasing the scale and scope of data analysis can create negative externalities in the form of better profiling, increased harms to privacy, and cybersecurity harms. Standardization also has implications for investment and innovation, especially if lock-in to an inefficient standard occurs. The Article then explores whether market-led standardization initiatives can be relied upon to increase welfare, and the role governmental-facilitated data standardization should play, if at all.

INTRODUCTION 738
I. DATA: ANALYSIS AND MARKETS 742
A. *Data and Data Analysis* 742
B. *Data Markets* 746
C. *Technological Obstacles to the Use of Data by Others* 747
II. THE STANDARDIZATION OF DATA 749
A. *What Is Data Standardization?* 749
B. *The Effects of Data Standardization on Data Use* ... 750
C. *Considerations Relevant to Data Standardization* 751

---

† Professor of Law, University of Haifa Faculty of Law and Chair of the International Society of Competition Law Scholars (ASCOLA).

‡ Robert L. Bridges Professor of Law and Professor of Economics Emeritus, U.C. Berkeley, and Professor of Law, New York University School of Law. The authors thank Daniel Francis, Avigdor Gal, Inge Graef, Scott Hemphill, Yoram Shiftan, Thomas Streinz, Eviatar Matania, and participants at the NYU Symposium on Data Law in the Global Economy and the biannual TILTing Perspectives Conference for wonderful comments on previous drafts; Benedict Kingsbury for helpful discussions; Eviatar Alkobi, Ilana Atron, Saar Ben Zeev, Ran Chaplin, Lior Frank, and Tamar Shtub for excellent research assistance; and the Center for Cyber Law and Policy at the University of Haifa for funding. Any mistakes or omissions remain the authors'.

1. *The Welfare Effects of Expanding the Potential Uses of Data* 754
2. *Effects on Competition in Data and Data-Based Markets* 757
3. *The International Dimension* 759
III. GOVERNMENT-FACILITATED DATA STANDARDIZATION . 761
A. *Potential Market Failures* 762
B. *What Role for Government?* 764
CONCLUSION 769

INTRODUCTION

Two decades ago, Sir Tim Berners-Lee, the founder of the World Wide Web, attempted to create what he called the semantic web. The idea was that all webpages should link their terminology to a common ontological standard.[1] The idea failed, partly because encoding such metadata—the data describing the data included in the dataset—was too complex, time-consuming, and prone to error.[2] But the idea behind this endeavor—setting standards with regard to data in order to facilitate its understanding and use by others—is still relevant today.

With data "rapidly becoming the lifeblood of the global economy,"[3] the efficiency of its use can significantly affect both social and private welfare. Data are an essential raw material in the data-driven economy.[4] Predictions based on patterns and correlations identified in the data affect numerous aspects of our lives, including health, education, transportation, and sustainability.[5]

For many applications, the quality of predictions is correlated with the volume of the data used in the analysis, as well as the diver-

---

[1] Tim Berners-Lee et al., *The Semantic Web*, SCI. AM. (May 17, 2001), https://www.researchgate.net/profile/James_Hendler/publication/12011854_Publishing_on_the_Semantic_Web/links/0deec527a72ca0bd4a000000/Publishing-on-the-Semantic-Web.pdf; Philippe Fournier-Viger, *The Semantic Web and Why It Failed*, DATA MINING BLOG (May 13, 2018), http://data-mining.philippe-fournier-viger.com/lessons-from-the-past-the-semantic-web-ontologies-and-why-it-failed.

[2] Fournier-Viger, *supra* note 1.

[3] European Political Strategy Ctr., European Comm'n, *Enter the Data Economy: EU Policies for a Thriving Data Ecosystem*, 21 EPSC STRATEGIC NOTES 1 (Jan. 11, 2017), https://ec.europa.eu/epsc/sites/epsc/files/strategic_note_issue_21.pdf.

[4] *See generally* OECD, DATA-DRIVEN INNOVATION: BIG DATA FOR GROWTH AND WELL-BEING (2015) (describing how data now drives all aspects of innovation in the economy and society).

[5] *See, e.g.*, COUNCIL OF ECON. ADVISORS, EXEC. OFFICE OF THE PRESIDENT, BIG DATA AND DIFFERENTIAL PRICING (2015), https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/docs/Big_Data_Report_Nonembargo_v2.pdf (examining the ability of companies to charge different prices to different consumers based on predictions gathered from big data).

sity of its sources, its accuracy, and its freshness. The volume of data collected has grown exponentially. By 2020, data collected worldwide are destined to reach forty thousand exabytes,[6] which the World Bank likened to a stack of reports reaching from the Earth to beyond Pluto.[7] Yet much of these data are collected in a system that is largely modular and distributed.[8] To illustrate, it is predicted that by 2020, thirty billion Internet of Things devices, controlled by numerous market players, will be hooked to the internet, collecting and using data.[9]

In such a system, data portability (the ability to transfer data without affecting its content) and interoperability (the ability to integrate two or more datasets) significantly affect the efficient use of data and, resultantly, public and private welfare.[10] Indeed, data portability may facilitate more and better data exchanges, thereby enabling more entities to use the data. Data interoperability can create data synergies: Combining data from different sources improves the knowledge that can be mined from them.[11] To illustrate, consider medical data on patients' responses to a treatment for a rare disease. Unless data were shared among its collectors and combined into a coherent dataset, it would be difficult to reach a better understanding of how to treat the disease. Or consider the operation of a smart city: Data from various sources (traffic lights, public transportation, pollution sensors, police reports, etc.) need to be integrated to enable synchronized and efficient operation.[12] It is thus not surprising that barriers to data porta-

---

[6] JOHN GANTZ & DAVID REINSEL, IDC, THE DIGITAL UNIVERSE IN 2020: BIG DATA, BIGGER DIGITAL SHADOWS, AND BIGGEST GROWTH IN THE FAR EAST 1 (2012), https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf. An exabyte is 10^{18} bytes.

[7] WORLD BANK GRP., WORLD DEVELOPMENT REPORT 2016: DIGITAL DIVIDENDS 244 (2016).

[8] *See* GREG ALLEN & TANIEL CHAN, BELFER CTR., HARVARD KENNEDY SCH., ARTIFICIAL INTELLIGENCE AND NATIONAL SECURITY 27 (2017), https://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf.

[9] *Internet of Things (IoT) Connected Devices Installed Base Worldwide from 2015 to 2025 (in Billions)*, STATISTA, https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide (last visited Mar. 20, 2019). Some firms enjoy comparative advantages in data collection. *See infra* Section II.B. The term Internet of Things relates to the digitization of the physical world through the creation of a network of devices (e.g., cars and refrigerators) that contain electronics, software, sensors, actuators, and connectivity which allows them to connect, interact, and exchange data.

[10] For the context of data-based deep learning, see Iain M. Cockburn et al., *The Impact of Artificial Intelligence on Innovation*, *in* THE ECONOMICS OF ARTIFICIAL INTELLIGENCE: AN AGENDA 115, 125–28, 139–43 (Ajay K. Agrawal et al. eds., 2019).

[11] *See, e.g.*, OECD, *supra* note 4, at 193 (describing how the value of data "increases when the data can be linked with and integrated into other data sets").

[12] DIRECTORATE-GEN. OF COMMC'NS NETWORKS, CONTENT & TECH., EUR. COMM'N, STUDY ON EMERGING ISSUES OF DATA OWNERSHIP, INTEROPERABILITY, (RE-)USABILITY

bility and interoperability have been identified as major barriers to the efficient operation of our data-intensive economy.[13]

Data standardization may be key to facilitating and improving the use of data,[14] by increasing data portability and interoperability. Indeed, standardization is a precondition for the operation of industries in which cross-firm and cross-industry data exchanges are critical.[15] Standardization can also create substantial benefits when data synergies carry high value. Healthcare offers a prime example: The use of similar indicators to record patients' responses to a treatment facilitates the integration of data that can inform clinical care, public health, and biomedical research. It can also improve healthcare delivery—for example, by offering personnel in different locations faster, interoperable access to patient records.[16] Accordingly, the U.S. Health Information Technology for Economic and Clinical Health (HITECH) Act[17] offers incentive payments to healthcare providers that meet a set of standards for electronic health records, designed to create a nationwide, interoperable health infrastructure.[18]

Understanding and evaluating the benefits and costs of data standardization is thus vitally important. Absent standardization, a "Tower of Babel" of different databases might be created, limiting potential data uses. Indeed, a recent survey found that more than half of data users in Europe identified lack of interoperability and technical standards as a blocking factor, preventing them from deploying new business models.[19] Lack of data standards was also found to constitute an important driver of costs, especially for small and medium

---

AND ACCESS TO DATA, AND LIABILITY 292–93 (2018) [hereinafter EUR. COMM'N], https://www.wik.org/fileadmin/Studien/2018/EU_Data_ownership_en.pdf.

[13] *Id.* at 15, 88; Oscar Borgogno & Giuseppe Colangelo, *Data Sharing and Interoperability: Fostering Innovation and Competition Through APIs*, COMPUTER L. & SECURITY REV. (forthcoming 2019) (manuscript at 2) (on file with authors).

[14] This Article defines standardization broadly, as a set of technical specifications designed to create a common design for a product or process. *See* Mark A. Lemley, *Intellectual Property Rights and Standard-Setting Organizations*, 90 CALIF. L. REV. 1889, 1896 (2002).

[15] *See, e.g.*, POLICY DEP'T A: ECON. & SCI. POLICY, DIRECTORATE-GEN. FOR INTERNAL POLICIES, EUROPEAN PARLIAMENT, INDUSTRY 4.0, 24 (Feb. 2016), http://www.europarl.europa.eu/RegData/etudes/STUD/2016/570007/IPOL_STU(2016)570007_EN.pdf (listing standardization as a precondition for the implementation of "Industry 4.0").

[16] JASON, AGENCY FOR HEALTHCARE RESEARCH & QUALITY, NO. 14-0041-EF, A ROBUST HEALTH DATA INFRASTRUCTURE 10 (2013), https://www.healthit.gov/sites/default/files/ptp13-700hhs_white.pdf.

[17] Health Information Technology for Economic and Clinical Health (HITECH) Act, Pub. L. No. 111-5, 123 Stat. 226 (2009) (codified at 42 U.S.C. §§ 300jj to 300jj-52 (2012)).

[18] John D. Halamka & Mickey Tripathi, *The HITECH Era in Retrospect*, 377 NEW ENG. J. MED. 907, 908 (2017).

[19] EUR. COMM'N, *supra* note 12, at 89.

enterprises, but also for large firms willing to share their data,[20] thereby potentially jeopardizing competition as well as innovation.[21] Furthermore, as Iain Cockburn and others emphasize, barriers to data sharing could result in the balkanization of data within particular sectors or even firms, thereby not only impeding innovation within markets, but also reducing spillovers to the improvement of analytical tools and to other markets.[22] Accordingly, broadening and improving the use of data through data standardization—whether of data semantics, attributes, structure, formats, or interfaces[23]—is likely to affect the competitive advantages of firms and nations. While data standards do not, by themselves, mandate actual data sharing, they open the door to it by creating the technological infrastructure that supports the development and diffusion of data and data analysis.[24] The need for evaluating data standards is further increased by the fact that other jurisdictions are currently in the process of setting data standards, which are likely to affect at least some domestic industries.[25]

Despite the importance of data standardization, the role of the government in such standardization is rarely examined. This Article analyzes the justifications for and the limitations of data standardization in light of data's special characteristics. In Part I, the Article examines the relevant characteristics of data and data markets. This Part discusses the technological obstacles to widening the use of data and explains how data standardization affects these obstacles. Part II analyzes the benefits and costs of data standardization. This Part shows that data raise new considerations about the appropriate interventionist role for regulators that are generally absent from debates about the standardization of other products and adds a novel dimension to some of the traditional considerations raised in more

---

[20] *Id.*

[21] Borgogno & Colangelo, *supra* note 13, at 3.

[22] Cockburn et al., *supra* note 10, at 15.

[23] *See also* Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV. J.L. & TECH. 1 (2011) (arguing in favor of anonymized research data and discrediting the alleged risks associated with such anonymized data).

[24] This Article does not argue for mandatory data sharing, a subject which deserves an analysis of its own. Nonetheless, as elaborated in Section II.C.2, *infra*, creating conditions for data sharing might strengthen incentives for data owners to voluntarily do so. Data sharing is also affected by other legal instruments, such as intellectual property and privacy law, which are beyond the scope of this Article. *See, e.g.*, Jorge L. Contreras & Jerome H. Reichman, *Sharing by Design: Data and Decentralized Commons*, 350 SCI. 1312, 1312 (2015) (arguing that to realize the promise of widespread sharing of data, intellectual property, data privacy, national security, and other legal and policy obstacles must be overcome); Nicolo Zingales, *Of Coffee Pods, Videogames, and Missed Interoperability: Reflections for EU Governance of the Internet of Things* (Dec. 2015), https://ssrn.com/abstract=2707570.

[25] *See infra* Section II.C.3.

typical cases. For example, data standardization can lead to smoother data flows, better machine learning, and easier policing in cases where rights are infringed upon or unjustified harms are created by data-fed algorithms. It might also help support a more competitive and distributed data collection ecosystem. At the same time, increasing the scale and scope of data analysis can create negative externalities in the form of better profiling, increased harms to privacy, and cybersecurity harms. This Part explores repercussions for investment and innovation in data collection and analysis.

Part III explores whether market-led standardization initiatives can be relied upon to increase welfare and evaluates the role governmental-facilitated standardization should play, if any. It is shown that the need for reviewing and possibly facilitating data standards can be especially strong where potential data synergies are cross-industry or intertemporal. This Part also explores the appropriateness of different regulatory methods for achieving these tasks. As shown, while governmental facilitation of data standardization may be justified only in limited scenarios, the current situation in which data standardization is rarely considered carries a high price tag.

I
DATA: ANALYSIS AND MARKETS

To understand the effects of data standardization, this Part explores the relevant characteristics of data, data analysis, and data markets, as well as some technological obstacles to the use of data and to data integration.

A. Data and Data Analysis

Not all data are alike. Different datasets may contain different variables, such as place, time, and subject.[26] For example, one dataset might include data on a patient's temperature, while another might include blood sugar levels. Dataset attributes are determined by the technological capabilities of sensors and communication devices, as well as by the preferences of the data collector.[27]

While some types of data are not fungible,[28] other datasets can be relevant for multiple users, sometimes operating in diverse markets.[29]

---

[26] *See, e.g.*, Helen Nissenbaum, *Must Privacy Give Way to Use Regulation?*, *in* DIGITAL MEDIA AND DEMOCRATIC FUTURES (Michael X. Delli Carpini ed., forthcoming 2019).

[27] *Id.*

[28] MAURICE E. STUCKE & ALLEN P. GRUNES, BIG DATA AND COMPETITION POLICY 79 (2016).

[29] *See* FED. TRADE COMM'N, DATA BROKERS: A CALL FOR TRANSPARENCY AND ACCOUNTABILITY 14 (2014) (describing how "data brokers" obtain information from

For example, data-based personal digital profiles can be relevant not only for the creation of personalized products and services,[30] but also for identifying voting patterns.[31] Furthermore, some types of data have relevance well beyond the data subjects themselves and their geographic spheres. Thus, correlations found in data describing the reaction of patients to a combination of pharmaceutical drugs in one locale may inform the treatment of patients elsewhere.

What gives large datasets ("big data") value is the potency of the insights that can be gleaned from their analysis.[32] Advancements in data science, including machine learning and deep learning,[33] have increased the ability of algorithms to reveal interesting relationships between attributes and to mine valuable knowledge for descriptive as well as predictive functions. Accordingly, data analysis allows for regularized customization of decisionmaking, thereby reducing risk and improving performance.[34] It also underlies new products and technologies that are changing fundamental features of contemporary life, as in the case of smart cities and autonomous cars.[35]

---

different sources, such as telephone companies and automobile dealers, for use in a variety of markets); ANJA LAMBRECHT & CATHERINE E. TUCKER, COMPETITION POLICY INT'L, CAN BIG DATA PROTECT A FIRM FROM COMPETITION? 1–2 (2017), https://www.competitionpolicyinternational.com/wp-content/uploads/2017/01/CPI-Lambrecht-Tucker.pdf (describing how "big data is non-rivalrous, meaning consumption of the good does not decrease its availability to others" and how "big data has near-zero marginal cost of production and distribution even over long distances," which leads "to a thriving industry where consumer big data is resold").

[30] Shoshana Zuboff, *Big Other: Surveillance Capitalism and the Prospects of an Information Civilization*, 30 J. INFO. TECH. 75, 83 (2015) (describing the personalization of search results and ads from Google).

[31] *See, e.g.*, Carole Cadwalladr & Emma Graham-Harrison, *Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach*, GUARDIAN (Mar. 17, 2018, 6:03 PM), https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election (discussing how Cambridge Analytica used personal digital profiles to identify voters' preferences and influence their votes).

[32] McKinsey estimates that data mining by firms increases operating margins by more than sixty percent. JAMES MANYIKA ET AL., MCKINSEY GLOB. INST., BIG DATA: THE NEXT FRONTIER FOR INNOVATION, COMPETITION, AND PRODUCTIVITY 2 (2011).

[33] Both constitute methods of data analysis. Machine learning uses statistical techniques to give computer systems the ability to "learn" (i.e., progressively improve performance) from data, without being explicitly programmed. *See generally* TREVOR HASTIE ET AL., THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION (2d ed. 2017) (providing a detailed description of the statistical frameworks for data learning and analysis). Deep learning offers an alternative paradigm for predicting complex multi-causal phenomena which is based on learning data representations, as opposed to task-specific algorithms. Yann LeCun et al., *Deep Learning*, 521 NATURE 436, 436 (2015).

[34] MANYIKA ET AL., *supra* note 32, at 5.

[35] OECD, *supra* note 4, at 382.

The quality of the knowledge mined from data is affected by data's four main characteristics: volume, velocity, variety,[36] and veracity. *Volume* relates to the quantity of data points in the dataset. *Velocity* relates to the "freshness" of the data. *Variety* concerns the number of different sources from which the data are gathered, and *veracity* the accuracy of the data. The relative importance of each of these characteristics may differ among uses. For example, where old data can serve as a sufficiently effective input, velocity is unimportant.[37]

Data are often characterized by economies of scale and scope, at least up to a point.[38] This implies that the more available and more varied the data, the better the knowledge that can be mined from it. As Mayer-Schönberger and Padova observe, "the value of data can be greatly enhanced . . . by combining it with other data sources. It is like a single puzzle piece that taken by itself offers little value, but when combined with others to complete an image is turned into something precious."[39]

The volume, variety, velocity, and veracity of the data may also affect the quality of the algorithm used for its analysis, due to the algorithm's feedback loop, with the algorithm evolving from learning based on an analysis of past predictions.[40] Accordingly, the better the data, the better the algorithm and the better its predictions.

The qualities of a dataset can also create positive externalities with respect to other datasets. This is because an algorithm can "learn" from a high-value dataset to perform tasks that can then be performed on different datasets—i.e., "transfer learning."[41] For example, Facebook was able to develop a better face-recognition algorithm because its algorithm could "learn" from a vast dataset that had a high level of accuracy, based on the abundant photos uploaded

---

[36] OECD, SUPPORTING INVESTMENT IN KNOWLEDGE CAPITAL, GROWTH AND INNOVATION 325 (2013); PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., EXEC. OFFICE OF THE PRESIDENT, BIG DATA AND PRIVACY: A TECHNOLOGICAL PERSPECTIVE 2 (2014); Mark Lycett, *'Datafication': Making Sense of (Big) Data in a Complex World*, 22 EUR. J. INFO. SYS. 381, 381 (2013).

[37] Daniel L. Rubinfeld & Michal S. Gal, *Access Barriers to Big Data*, 59 ARIZ. L. REV. 339, 347 (2017).

[38] *Id.* at 352–55.

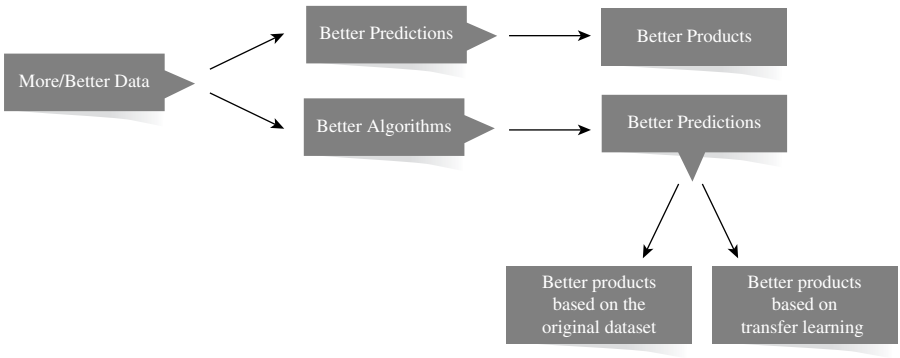
[39] Viktor Mayer-Schönberger & Yann Padova, *Regime Change? Enabling Big Data Through Europe's New Data Protection Regulation*, 17 COLUM. SCI. & TECH. L. REV. 315, 320 (2016).

[40] *See* STUCKE & GRUNES, *supra* note 28, at 170 (describing this feedback loop in the context of Google's search engine algorithm).

[41] Lilyana Mihalkova et al., *Mapping and Revising Markov Logic Networks for Transfer Learning*, *in* PROCEEDINGS OF THE 22ND CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI-07) 608, 608 (2007).

to its website and tagged by users.[42] The same algorithm can then be used for security cameras, surveillance, etc. Diagram 1 summarizes the benefits that flow from more and better data.

DIAGRAM 1: SCALE AND SCOPE ECONOMIES IN DATA ANALYSIS



How big must big data be in order to maximize its benefits? The answer varies with the types and uses of data. In general, tasks such as identifying patterns, generating predictions, and promptly adapting to rapidly changing circumstances require vast datasets of fresh, varied, and accurate data.[43] Furthermore, the increasing use of deep learning as a data analysis tool "implies a shift towards investigative approaches that use large datasets to generate predictions for physical and logical events that have previously resisted systematic empirical scrutiny."[44] Accordingly, large volumes of diversified data have been recognized as central resources for economic markets, for governance (e.g., health hazards and terror threats), and for cybersecurity.[45]

Importantly for this Article's analysis, data are intangible and non-rivalrous; one individual's use of data does not, as a general rule, impact the ability of others to make use of the same data.[46] Moreover, data are often transferable or replicable at very low marginal cost. Finally, they are divisible and can potentially be integrated with other

[42] Tom Simonite, *Facebook Creates Software that Matches Faces Almost as Well as You Do*, MIT TECH. REV. (Mar. 17, 2014), https://www.technologyreview.com/s/525586/facebook-creates-software-that-matches-faces-almost-as-well-as-you-do.

[43] On facial recognition algorithms, see, for example, PATRICK GROTHER ET AL., NAT'L INST. OF STANDARDS & TECH., ONGOING FACE RECOGNITION VENDOR TEST (FRVT) (2018), https://www.nist.gov/sites/default/files/documents/2018/06/21/frvt_report_2018_06_21.pdf, which provides a continuously updated report of the ongoing Face Recognition Vendor Test.

[44] Cockburn et al., *supra* note 10, at 22.

[45] *See, e.g.*, OECD, *supra* note 4.

[46] STUCKE & GRUNES, *supra* note 28, at 44.

data, whether collected by the same entity or by another.[47] Where economies of scale and scope cannot be achieved by a single entity or by a single source of data, data integration may significantly increase data's predictive value. Yet, as elaborated below, the integration of huge amounts of data into one high-quality dataset is complex and raises synchronization and search optimization issues.[48] The challenge is to integrate data that are not necessarily similar in source or structure and to do so quickly and at a reasonable cost.[49]

B. Data Markets

The data value chain consists of five main links: collection, organization, analysis, storage, and use.[50] *Collection* relates to the extraction, recording, and aggregation of data into a form that can be used for data mining. *Organization* involves structuring the database, including synthesis of certain data points, and the addition of headings and explanatory notes. It turns the data into information. *Analysis* relates to the integration and processing of different types of data. It transforms information into knowledge. *Storage* entails archiving data in retrievable forms. *Use* involves utilizing data-based knowledge for prediction and decisionmaking in relevant markets. As noted, this value chain also has a dynamic internal reciprocal dimension: Data regarding the success of an algorithm's past predictions can be used to "teach" the algorithm to make better, more accurate predictions in the future.[51]

Sources of data vary significantly, ranging from individual activity both off and online to locational signals and data collected from sensors embedded in "things."[52] Data collection may be based on different models, including competition over internet users' online attention and mandatory provision by data subjects (such as tax returns).[53] Moreover, some data are collected as a by-product of other activities (such as data on geothermic conditions in oil rigging explorations).[54]

---

47 FED. TRADE COMM'N, *supra* note 29, at 14.

48 *See infra* Section II.C.

49 *The 6 Challenges of Big Data Integration*, FLYDATA, https://www.flydata.com/the-6-challenges-of-big-data-integration (last visited Mar. 20, 2019).

50 This paragraph builds on Niva Elkin-Koren & Michal S. Gal, *The Chilling Effects of Governance-by-Data on Data Markets*, 86 U. CHI. L. REV. 403, 407–10 (2019).

51 *See supra* Section II.A.

52 Rubinfeld & Gal, *supra* note 37, at 350–51.

53 *See, e.g.*, TIM WU, THE ATTENTION MERCHANTS 167, 251 (2016) (describing AOL, Facebook, and Google as "online attention merchants" based on their collection and financial exploitation of users' data).

54 Rubinfeld & Gal, *supra* note 37, at 357.

All links along the data value chain exhibit some entry barriers, which may be technological, behavioral, or legal.[55] Some types of data are collected by numerous firms at low cost (e.g., smartphone users' locational data), and similar data can sometimes be collected from different sources. However, some data collection activities are costly and/or hard to replicate.[56] This may reflect exclusivity of access points, temporal advantages (e.g., aerial maps taken before a natural disaster), network effects in collection or collection-inducing products or services, or legal and behavioral limitations on collection.[57]

Competition for data collection, analysis, and storage, as well as competition in markets for data-based products or services, is shaped by the height of entry barriers at each link of the data value chain and by the interactions between such links.[58] Whatever the market structure, data—or the knowledge mined from it—can be traded. Indeed, the demand for data has created an ecosystem consisting of numerous firms which trade in data.[59] This, in turn, enables firms to use data collected elsewhere to scale up their datasets.

C. Technological Obstacles to the Use of Data by Others

To understand the benefits of data standardization, this Section identifies the three main obstacles to the current use of data collected that would be solved through standardization.

The first involves *metadata uncertainties*.[60] Metadata comprise the data that describe the data included in a dataset. Metadata may relate, for example, to data semantics (attributes of the data, such as the metrics used), or to how accurately the data were recorded. Obstacles to using data collected by others may arise when the relevant metadata are partial or unknown. Metadata uncertainties limit others' ability to understand what different data points signify (e.g., does the label

---

[55] *See id.* at 349–67 (describing the particular technological, legal, and behavioral barriers to the collection, storage, synthesis, analysis, and usage of big data).

[56] *See id.* at 351, 357–63.

[57] *See id.* at 351, 355.

[58] *See* OECD, *supra* note 4, at 391–92 ("New businesses labelled under the 'sharing economy' have overcome high entry barriers and created new competition in established markets, notably in urban mobility and accommodation."); STUCKE & GRUNES, *supra* note 28.

[59] *See* STAFF OF S. COMM. ON COM., SCI. & TRANSP., 113TH CONG., A REVIEW OF THE DATA BROKER INDUSTRY: COLLECTION, USE, AND SALE OF CONSUMER DATA FOR MARKETING PURPOSES 20 (2013) (describing the sources of database marketing company Acxiom's consumer data); *see also* FED. TRADE COMM'N, *supra* note 29, at 8–9 (naming and detailing the activities of various firms in the data brokerage industry).

[60] Avigdor Gal, *Uncertain Schema Matching*, *in* 13 SYNTHESIS LECTURES ON DATA MANAGEMENT 1, 2 (M. Tamer Özsu ed., 2011).

"address" relate to billing or to shipping?).[61] As such, they can increase information asymmetries regarding the content of datasets, thereby reducing incentives to engage in mutually beneficial data sharing. Such uncertainties can also lead to incorrect assumptions that may skew the data analysis.

The importance of metadata can be illustrated by the Mars Climate Orbiter incident.[62] Designed to study the Martian climate, the orbiter was launched in December 1998 following years of preparation and investment of billions of dollars. But in September 1999, as the spacecraft prepared to enter orbit around Mars, its trajectory brought it too close to the planet and it burned up in the atmosphere. A post-mortem analysis found that the failure resulted from the use of two different standards in one database. Specifically, trajectory software created by Lockheed Martin Astronautics produced output in English units instead of the metric units specified in Lockheed's contract with NASA. The combination of data from the two sources led to the erroneous calculations of trajectory.[63]

The second limitation involves *obstacles to data transformation*, which can raise the costs of combining the available data into coherent datasets.[64] One such obstacle results from data granularity, as when similarly attributed data are collected at different temporal intervals that are difficult to integrate. Another obstacle can arise from the need to reorganize data into a new, combined dataset with a different structure or internal organization. While data scientists are developing tools that "translate" data to make it compatible with other data, the costs and complexities of using these tools may be high. To illustrate, Amazon has encountered difficulties in migrating user data from a

---

[61] *See also, e.g.*, OFFICE OF THE NAT'L COORDINATOR FOR HEALTH INFO. TECH., CONNECTING HEALTH AND CARE FOR THE NATION: A SHARED NATIONWIDE INTEROPERABILITY ROADMAP 25 (drft. 2015) [hereinafter ROADMAP], http://www.healthit.gov/sites/default/files/nationwide-interoperability-roadmap-draft-version-1.0.pdf ("[A] health professional would readily understand that 'Tylenol' and 'acetaminophen' are generally used interchangeably. However, two computer systems exchanging those phrases may treat the terms entirely differently if the systems are not bound to a standardized vocabulary or terminology that equates them as synonyms.").

[62] *See* MARS CLIMATE ORBITER MISHAP INVESTIGATION BD., MARS CLIMATE ORBITER MISHAP INVESTIGATION BOARD PHASE I REPORT (1999), https://llis.nasa.gov/llis_lib/pdf/1009464main1_0641-mr.pdf (describing the root of and contributing causes to the Mars Climate Orbiter's failure).

[63] *See id.* at 6–7 ("[T]he fact that the angular momentum (impulse) data was in English, rather than metric, units, resulted in small errors being introduced in the trajectory estimate over the course of the 9-month journey.").

[64] *See* Gal, *supra* note 60, at 9–14.

variety of databases to its web services platform.[65] In a world with fast-changing data, the costs of such transformation are ongoing.

The third obstacle involves *missing data*. This limitation, which is probably the most difficult to correct ex post, arises when some necessary data were not collected, and collecting the missing data at a later time might be impossible, or the costs of ex post collection might be prohibitive. Missing data may result, inter alia, from limited capacity of a database to store the data,[66] or from data collectors' limited foreseeability of the value of data interoperability.[67]

These three limitations reduce users' incentives and ability to extend the use of data and to achieve data synergies. Indeed, a recent study found that "merging different datasets and making them interoperable is one of the most resource-intensive activities for data (re-)users and that, even within the same value chain, datasets are rarely interoperable by default."[68]

II
The Standardization of Data

Part II explores the role of data standardization. It begins by characterizing data standardization and explains how it could resolve some of the obstacles to the use of data. It then shows that data standardization requires a complex balancing of considerations that go beyond those typically regarded as relevant in traditional industries.

A. What Is Data Standardization?

Data standardization involves setting standards that relate to the data value chain. Standards can relate, for example, to the attributes of the data to be collected; to the terminology, structure, and organization of the dataset; to aspects of data storage (location, etc.); or to its use (including protocols for data portability). The first known data standard was created at the end of World War II, in response to the logistical complexity of the 1948 Berlin Airlift. Air traffic was slowed

---

[65] *See* Silvia Doomra, *Challenges when Migrating from Oracle to PostgreSQL—and how to Overcome Them*, Amazon Web Services: AWS Database Blog (Feb. 1, 2018), https://aws.amazon.com/blogs/database/challenges-when-migrating-from-oracle-to-postgresql-and-how-to-overcome-them.

[66] For an in-depth analysis of the problems involved in collecting and storing data, see Blue Ribbon Task Force on Sustainable Dig. Pres. & Access, Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information (2010), http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

[67] *Cf.* Mayer-Schönberger & Padova, *supra* note 39, at 319–20 ("[T]he value of data can be greatly enhanced not only by having and analyzing more of it, but by combining it with other data sources.").

[68] Eur. Comm'n, *supra* note 12, at 89.

by bottlenecks created at unloading, with ground crews having to check long lists of goods brought by each plane. To resolve this problem, a standardized system of codes was created, allowing shipment notices to be electronically reported before the plane landed.[69]

The most commonly used data standards are Application Programming Interfaces (APIs). These are computer protocols that define how software components communicate with one another.[70] In particular, APIs ease the flow of data by describing the kinds of data that can be retrieved, how to retrieve it, and the format in which data will be shared. They may also include the associated metadata, which describe the data's attributes or semantics and enable users to interpret the meaning and significance of different data points.[71] Yet while in some industries a consensus exists with regard to what APIs to use, in many others there is no consensus on how APIs should be determined, and whether open, standardized ones are preferable.[72] Furthermore, APIs do not necessarily solve data transformation and missing data problems.

B. The Effects of Data Standardization on Data Use

Data standardization can potentially reduce all of the obstacles to data use by others. First, it can reduce metadata uncertainties by requiring that data semantics follow certain norms and rules. One example is the standardized dot matrix font set for Chinese ideograms for information exchange. The standard enables different datasets to interface with one another by ensuring that the tens of thousands of Chinese characters used in all datasets are similar.[73]

Data standards can also reduce obstacles to data transformation, for example by standardizing the structure and organization of datasets.[74] To illustrate, many public transport firms have adopted a standard format for reporting public transportation schedules and associated geographic information.[75] This standard facilitates data

---

[69] ASHUTOSH DESHMUKH, DIGITAL ACCOUNTING: THE EFFECTS OF THE INTERNET AND ERP ON ACCOUNTING 2 (2006).

[70] Borgogno & Colangelo, *supra* note 13, at 8.

[71] *See id.* at 6, 8 (identifying the data-sharing benefits of providing APIs along with the associated metadata).

[72] *See id.* at 8–10.

[73] Frederick R. Burke, Note, *The Administrative Law of Standardization in the PRC*, 1 J. CHINESE L. 271, 273 (1987) (citation omitted).

[74] *See* EUR. COMM'N, *supra* note 12, at 88–89 (describing how a lack of standardization and interoperability prevent effective data sharing and use).

[75] *See* GENERAL TRANSIT FEED SPECIFICATION, http://gtfs.org (last visited Mar. 20, 2019) ("The General Transit Feed Specification (GTFS) defines an open standard format for exchanging public transportation schedule, geographic and fare information.").

interoperability by introducing a modular set of compatible data, as well as common data models and schemas. It has made possible the creation of multi-modal synchronized journey planner applications such as Moovit and Google Maps.[76] Finally, data standardization can reduce the missing data problem. The U.S. Office of the National Coordinator for Health Information Technology, for example, sets standards for the collection of observations regarding patients' allergies.[77] Of course, data standardization is only one way to address obstacles to sharing or integrating data, and it may carry high costs—considerations, which are analyzed below.[78]

In addition, data standards affect what one can learn from data, regardless of whether the data have been shared. Structured Query Language (SQL), a database query language introduced in the 1970s and still in use today, offers a case in point.[79] SQL accesses databases using simple syntax. The relational model that underlies this standard relies on tables of data, and significantly curtails users' ability to articulate complex queries.[80] Competing object-oriented database models suggested a way to overcome such limitations.[81] However, despite their potential, market players did not invest in their development—in part because the main database vendors, Oracle and IBM, had significant sunk costs in SQL.[82] Instead, vendors made easy, low-cost changes and renamed their databases object-relational databases.[83] As a result, the data industry never fully adopted this better standard.[84]

C. Considerations Relevant to Data Standardization

The use of standards is driven by powerful demand-side and supply-side forces. Indeed, several studies have established a clear

---

[76] *See GTFS Realtime Overview*, Google Developers: Google Transit APIs, https://developers.google.com/transit/gtfs-realtime (last visited Mar. 20, 2019).

[77] *See* Office of the Nat'l Coordinator for Health Info. Tech., 2019 Interoperability Standards Advisory 2 (2019), https://www.healthit.gov/isa/sites/isa/files/inline-files/2019ISAReferenceEdition.pdf.

[78] *See infra* Section II.C.

[79] *SQL*, Encyclopædia Britannica, https://www.britannica.com/technology/SQL (last visited Mar. 30, 2019).

[80] For some criticisms of SQL, see Donald D. Chamberlin, *Early History of SQL*, IEEE Annals Hist. Computing, Oct.–Dec. 2012, at 78, 80–81.

[81] *See* Won Kim, Introduction to Object-Oriented Databases 3–4 (1990) (explaining how object-oriented database technology attempted to incorporate solutions to many of the problems of conventional database technology).

[82] *See* Neal Leavitt, *Whatever Happened to Object-Oriented Databases?*, IEEE Computer, Aug. 2000, at 16, 19 (noting that the lack of vendors backing object-oriented databases was a significant obstacle to their use despite predicted popularity).

[83] *See id.* at 17–19.

[84] *See id.* at 19 (describing object-oriented databases as only used in niche markets, with no greater popularity likely to come).

connection between standardization, productivity growth, and overall economic growth in some industries.[85] While many of the common rationales for standardization apply to data, data standardization adds another layer of complexity to such rationales. What follows is a brief survey of the common arguments for standardization.

Compatibility standards offer users the guarantee of interoperability: Whenever and wherever they use a product or service, they can be assured that the product will be functional.[86] To illustrate, similar electric plugs and sockets ensure that products operated by electricity can be utilized anywhere. The benefits of interoperability may be especially great when network externalities exist (i.e., when the value that individuals place on products or services depends on the use of those products by others).[87] Compatibility standards also lower switching costs, by ensuring the retention of functionality.[88] This has the potential to facilitate competition in standard-based and interconnected standard-compatible markets, thereby reducing costs and increasing innovation, product quality, and choice.[89]

Minimum quality standards may correct market failures.[90] This may be the case where quality cannot be easily measured and/or observed and information asymmetries prevent consumers from comparing product quality, or where not all consumers internalize externalities, as with the case of health, environmental, and security standards.[91] Setting minimum standards may also promote consumer trust.[92]

---

[85] *See, e.g.*, G.M. PETER SWANN, THE ECONOMICS OF STANDARDIZATION: AN UPDATE 4–16 (2010); Knut Blind & Andre Jungmittag, *The Impact of Patents and Standards on Macroeconomic Growth: A Panel Approach Covering Four Countries and 12 Sectors*, 29 J. PRODUCTIVITY ANALYSIS 51, 51 (2008).

[86] *See* Paul A. David & Shane Greenstein, *The Economics of Compatibility Standards: An Introduction to Recent Research*, 1 ECON. INNOVATION & NEW TECH. 3, 4 (1990) (defining compatibility standards as those that assure users that a product will successfully be incorporated into various systems provided by various suppliers); Michael L. Katz & Carl Shapiro, *Network Externalities, Competition, and Compatibility*, 75 AM. ECON. REV. 424, 439 (1985) (noting the role of public policy and economic incentives in generating increased compatibility standards for consumer benefit).

[87] J. Gregory Sidak, *The Value of a Standard Versus the Value of Standardization*, 68 BAYLOR L. REV. 59, 61–62 (2016).

[88] *See, e.g.*, Joseph Farrell & Garth Saloner, *Standardization, Compatibility, and Innovation*, 16 RAND J. ECON. 70, 71–72 (1985) (describing switching to new technology as theoretically low-cost and beneficial but burdened with issues of excess inertia).

[89] *See, e.g.*, U.S. DEP'T OF JUSTICE & FED. TRADE COMM'N, ANTITRUST ENFORCEMENT AND INTELLECTUAL PROPERTY RIGHTS: PROMOTING INNOVATION AND COMPETITION 33 (2007).

[90] SWANN, *supra* note 85, at 15.

[91] KNUT BLIND, STANDARDISATION: A CATALYST FOR INNOVATION 30 (2009).

[92] *Id.*

Of course, standardization can be costly. With standardization comes a reduced variety of choices available to users.[93] In addition, when the standard is proprietary, the exercise of market power could lead to reduced competition and higher prices.[94] Moreover, compliance may be costly.[95] Standards can also negatively affect innovation if the industry, or part thereof, is locked into an inferior standard.[96]

These considerations apply equally to data. Data standardization can increase interoperability (of datasets), lower switching costs for consumers (from one data collector to another), and limit duplication (of data collection, storage, and analysis). Potential harms are also relevant. Most important is the risk of lock-in to an inefficient standard. To illustrate, assume that a data standard requires all medical data collectors to gather certain types of data at specified intervals, but these intervals are too far apart to provide meaningful data. While data could be collected at shorter intervals, the standard might send a wrong signal as to the appropriate interval. In addition, data standards can impose high compliance costs on all market players, potentially leading to higher prices and reduced competition.[97] Data standards can also negatively affect competition by raising some competitors' costs.[98] Finally, they might make coordination and collusion easier.[99]

Data standardization also raises considerations that are generally less relevant to non-data standards. The discussion that follows identifies three such considerations, which arise from the increased use of

---

[93] SWANN, *supra* note 85, at 24–25.

[94] *See generally* Lemley, *supra* note 14 (discussing the importance of standard-setting organizations (SSOs) in the context of intellectual property rules, including how antitrust law plays a role in protecting both the durability of competition and against efforts to raise costs).

[95] SWANN, *supra* note 85, at 15.

[96] *See* Farrell & Saloner, *supra* note 88, at 71 ("[I]t is plausible that [an] industry, once firmly bound together by the benefits of compatibility or standardization, will be inclined to move extremely reluctantly to a new and better standard because of the coordination problems involved.").

[97] *See* Orla Lynskey, *Aligning Data Protection Rights with Competition Law Remedies? The GDPR Right to Data Portability*, 42 EUR. L. REV. 793, 808 (2017) (identifying the cost of compliance associated with data portability as higher than the relevant literature currently suggests); Peter Swire & Yianni Lagos, *Why the Right to Data Portability Likely Reduces Consumer Welfare: Antitrust and Privacy Critique*, 72 MD. L. REV. 335, 352 (2013) (explaining that the Right to Data Portability could burden large and small companies alike).

[98] *See Developments in the Law—More Data, More Problems*, 13 HARV. L. REV. 1715, 1722, 1733–34 (2018) (suggesting that the requirements for data storage applied by the Second Circuit create a comparative advantage for Microsoft relative to its competitors).

[99] *See* ARIEL EZRACHI & MAURICE E. STUCKE, VIRTUAL COMPETITION 29–33 (2016) (advocating for increased regulation of the data-driven economy in order to address potential opportunities for collusion); Michal S. Gal, *Algorithms as Illegal Agreements*, 34 BERKELEY TECH. L.J. (forthcoming 2019) (manuscript at 2) (on file with authors).

data, facilitated by data standardization. Some of these considerations derive from the fact that data add a unique technological dimension to standardization: While standardization generally creates interoperability between the standardized product and other products (e.g., plugs and sockets create compatibility between electricity-powered products and electricity service providers), data standardization can also create portability and interoperability between the standardized products (datasets) themselves. In the analysis that follows, it is assumed that the chosen standards are otherwise economically efficient. This assumption is relaxed later.

1. The Welfare Effects of Expanding the Potential Uses of Data

Data standardization can potentially increase incentives for data collection, organization, and storage, thereby generating ever-larger amounts of more accessible data.[100] This is because expanding the potential uses of data increases its value, and the interoperability of different data sources reduces investment risks associated with data collection, organization, and storage. By reducing data portability costs and enabling more market players to utilize the data, data standardization may also increase incentives for data sharing.

The effects of this potential increase in both the amount and versatility of available data are essential variables in any analysis of the welfare effects of data standardization. As elaborated above, increased use of data may facilitate cumulative and synergetic knowledge production, which may stimulate new and better products or services.[101] Indeed, the positive effects of better knowledge may also extend beyond the market for which the data are immediately relevant, due to transfer learning.[102] The importance of widening the use of data cannot be overstated. Data serve as a foundational input in the information-based economy. Where firms compete over data-based advantages, inefficient use of data can be costly. This is especially true where aggregation of data from several sources is essential for the operation of markets, such as in the case of smart cities and smart

[100] Data standards also affect the types of data to be collected. By reducing the costs of integration of external data, standards may reduce incentives to collect duplicative data while increasing incentives to collect unique data.

[101] This was recognized by the European Commission. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A Digital Single Market Strategy for Europe*, at 14–15, COM (2015) 192 final (May 6, 2015). Realizing potential data synergies also depends on the information market participants possess regarding relevant datasets. Barbara Engels, *Data Portability Among Online Platforms*, INTERNET POL'Y REV., June 2016, at 1, 9.

[102] *See supra* notes 29–32, 42–43 and accompanying text.

homes, or where data synergies are substantial and cannot be easily achieved by one entity. Indeed, research has shown that access to data can shape both the level and direction of innovative activity, thereby affecting both private as well as social welfare.[103]

Yet the foundational role of data in our economy, and its diverse uses, add complexity to the welfare analysis, raising issues that go well beyond economics to the social, political, and legal spheres. Take, for example, the creation of a digital profile.[104] The more data, the more accurate the profile, and the more personalized the treatment of the data subject. In the economic sphere, an individual may receive offers for products that better fit her preferences but possibly at higher, discriminatory prices and/or lower quality.[105] In the social sphere, she may receive more tailored suggestions for connections (e.g., via LinkedIn) and content that cater to her prior interests, but that might also potentially limit her viewpoints.[106] In the political sphere, her personalized digital feed could be designed to strengthen certain opinions and affect political choices.[107] In the legal sphere, digital profiles could inform decisions made by law enforcement or judicial bodies (e.g., a suspect's flight risk), and even lead to the creation of personalized laws.[108] This discussion also illustrates that the same data can be used in both welfare-enhancing and welfare-reducing ways.

---

[103] *See* Jeffrey L. Furman & Scott Stern, *Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research*, 101 AM. ECON. REV. 1933, 1936 (2011); Heidi L. Williams, *Intellectual Property Rights and Innovation: Evidence from the Human Genome*, 121 J. POL. ECON. 1, 1–2 (2013).

[104] *See* EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES 7, 44 (2014) [hereinafter SEIZING OPPORTUNITIES] (outlining the type of data contained within a profile and the method by which profiles are created); *see generally* EXEC. OFFICE OF THE PRESIDENT, BIG DATA: A REPORT ON ALGORITHMIC SYSTEMS, OPPORTUNITY, AND CIVIL RIGHTS (2016) (addressing harms that can result from supplying data to algorithmic profiling software).

[105] *See* FED. TRADE COMM'N, BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? 9–11 (2016) (outlining ways in which the use of big data can generate harmful consequences for low-income groups).

[106] *See* Christoph B. Graber, *The Future of Online Content Personalisation: Technology, Law and Digital Freedoms* 6–8 (Univ. of Zurich, I-Call Working Paper No. 2016/01, 2016) (discussing online content personalization and popular criticisms of it).

[107] *See, e.g.*, Emma Graham-Harrison, Carole Cadwalladr & Hilary Osborne, *Cambridge Analytica Boasts of Dirty Tricks to Swing Elections*, GUARDIAN (Mar. 19, 2018, 3:00 PM), https://www.theguardian.com/uk-news/2018/mar/19/cambridge-analytica-execs-boast-dirty-tricks-honey-traps-elections.

[108] *See generally* Omri Ben-Shahar & Ariel Porat, *Personalizing Negligence Law*, 91 N.Y.U. L. REV. 627 (2016) (arguing that courts can and should use data to create personalized reasonable person standards for negligence inquiries); Ariel Porat & Lior Jacob Strahilevitz, *Personalizing Default Rules and Disclosure with Big Data*, 112 MICH. L. REV. 1417 (2014) (discussing personalized law and advocating for increased personalization in the context of default rules and disclosure).

Data standardization also raises privacy concerns. The easier it is to share data, the greater the concern that private data will fall into more hands.[109] Furthermore, increases in the size and quality of a dataset can create negative privacy externalities, due to the fact that missing information about a data subject can be indirectly learned by observing other data subjects with similar attributes.[110] For example, a person's tastes in literature and fashion can be deduced indirectly from data on her social and workplace circles. A similar process may reduce the ability of individuals to hide their identities and could also harm data anonymization efforts.[111] This could significantly impair privacy, leaving users exposed in the digital environment.[112] It could also reduce the willingness of potential data subjects to allow their private data to be collected, thereby potentially affecting data collection and innovation.[113]

Data standardization can also affect cybersecurity.[114] Integration of databases may enable security systems to more efficiently detect patterns of suspicious activity, and the scale of data may allow algorithms to more rapidly learn from past patterns to detect future attacks.[115] Yet these benefits come with tradeoffs. The more standardized the data, the easier it might be for hackers to access and use it. The potential harm becomes even greater to the extent that data standardization enables the creation of larger, less dispersed databases, as the size of the dataset may be positively correlated with the potential harm from security breaches.[116] Furthermore, an ineffi-

---

[109] *See* Swire & Lagos, *supra* note 97, at 339.

[110] Yoan Hermstrüwer, *Contracting Around Privacy: The (Behavioral) Law and Economics of Consent and Big Data*, 8 J. INTELL. PROP. INFO. TECH. & ELECTRONIC COM. L. 9, 12 (2017).

[111] *See* Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, *in* 2008 IEEE SYMPOSIUM ON SECURITY AND PRIVACY 111, 111 (2008) ("[T]he adversary with a small amount of background knowledge about an individual can use it to identify, with high probability, this individual's record in [an] anonymized dataset and to learn all anonymously released information about him or her, including sensitive attributes.").

[112] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1748 (2010).

[113] Elkin-Koren & Gal, *supra* note 50 (manuscript at 18–23).

[114] Security harms do not involve privacy alone but can also engender economic harms, for example through the loss of financial data and identity theft. *See, e.g.*, CLARE SULLIVAN, DIGITAL IDENTITY: AN EMERGENT LEGAL CONCEPT 113–16 (2011) (providing a definition of identity theft).

[115] *See* ALLEN & CHAN, *supra* note 8, at 27 (noting the volume of new data "stored across unintegrated databases"); TATIANA TROPINA & CORMAC CALLANAN, SELF- AND CO-REGULATION IN CYBERCRIME, CYBERSECURITY AND NATIONAL SECURITY 14 (2015) (explaining the difficulties for cybersecurity from a decentralized cyber environment).

[116] Wolfgang Kerber & Heike Schweitzer, *Interoperability in the Digital Economy*, 8 J. INTELL. PROP. INFO. TECH. & ELECTRONIC COM. L. 39, 54 (2017).

cient standard can reduce organizations' ability to detect cyber threats. A case in point is the Common Criteria for Information Technology Security Evaluation standard, created to ease the process of specifying, implementing, and evaluating security products.[117] Yet its stringent requirements (e.g., for verifying that a system is not under attack) make its implementation costly and slow down the detection of threats.[118] While its application may be justified with regard to key infrastructure, it may not be efficient for other markets.[119]

The above considerations should be taken into account when analyzing the effects of data standardization. Our analysis indicates that the costs and benefits of data standardization might differ among different types of data or its uses. It also suggests that often the beneficial and harmful effects of increased use of data cannot be easily separated. Accordingly, it may generally be better to prevent certain uses of data, including its sharing under certain circumstances, than to prevent data standardization. At the same time, in some settings encouraging data standardization must be accompanied by some safeguards—legal, technological, or even cultural—that ensure that its overall effects on social welfare are positive. For example, instead of preventing data standardization, which may lead to the creation of larger datasets, the government could facilitate their better protection.

2. *Effects on Competition in Data and Data-Based Markets*

Data standardization can help support a competitive and distributed data collection ecosystem. As suggested above, data standardization can increase the incentives of firms to collect and to share data. As a result, the market for data may become more competitive. Furthermore, the increased ability of firms to integrate different datasets may reduce the need to rely on one source for data, either internal or external.[120]

This is no small accomplishment, as it may carry the seeds for reducing one of the main concerns raised in digital markets: the fact that some firms may accrue significant and durable market power

---

[117] COMMON CRITERIA, https://www.commoncriteriaportal.org (last visited Mar. 14, 2019).

[118] Dieter Ernst & Sheri Martin, *The Common Criteria for Information Technology Security Evaluation—Implications for China's Policy on Information Security Standards* 5 (E.-W. Ctr., Working Paper No. 108, 2010), https://www.eastwestcenter.org/system/tdf/private/econwp108.pdf?file=1.

[119] *Id.*

[120] For the tradeoffs between internal and external relationships in interconnections, see R.H. Coase, *The Nature of the Firm*, 4 ECONOMICA 386, 394–95 (1937).

which is based in part on their control of vast amounts of data.[121] The comparative advantages enjoyed by such firms are partly based on economies of scale and scope in data collection and analysis as well as on network effects.[122] Of significance are "feedback" network effects, which arise when the quality of data—which is a function of its volume, variety, veracity, and velocity—enables the supplier to accelerate its learning and increase product quality.[123] Such network effects were recognized, inter alia, in the Microsoft–Yahoo case, in which the Department of Justice stated that "access to a larger set of queries [by different users] . . . should accelerate the automated learning of Microsoft's search . . . algorithms."[124] As Maurice Stucke and Allen Grunes argue, this feedback loop may be accelerated when a firm enjoys a variety of data sources, which can then be combined to yield more accurate predictions.[125] For example, Google may combine data regarding a user's email, geo-location, and browser history to better predict his preferences. Other firms, which lack such a variety of data sources, may find it difficult to match these capabilities.[126] The quality gap created by such network effects carries the potential to entrench or strengthen the dominance of some firms.[127] As a result, data-based markets could exhibit highly concentrated structures with a single dominant firm possessing a massive share.[128] Benefits arising from data collection and analysis that are not the result of artificial entry barriers are not prohibited by antitrust legislation.[129] Accordingly, alternative solutions become more important.

---

121 *E.g.*, OECD, *supra* note 4, at 107; John M. Newman, Antitrust in Digital Markets 14–16 (Mar. 15, 2019) (unpublished manuscript), https://ssrn.com/abstract=3201004.

122 STUCKE & GRUNES, *supra* note 28, at 43–44 (quoting OECD, EXPLORING THE ECONOMICS OF PERSONAL DATA: A SURVEY OF METHODOLOGIES FOR MEASURING MONETARY VALUE 34 (2013)); Howard A. Shelanski, *Information, Innovation, and Competition Policy for the Internet*, 161 U. PA. L. REV. 1663, 1679 (2013).

123 *See* CARL SHAPIRO & HAL R. VARIAN, INFORMATION RULES: A STRATEGIC GUIDE TO THE NETWORK ECONOMY 175–84 (1999) (explaining the concepts of feedback and network externalities).

124 Press Release, U.S. Dep't of Justice, Statement of the Department of Justice Antitrust Division on Its Decision to Close Its Investigation of the Internet Search and Paid Search Advertising Agreement Between Microsoft Corporation and Yahoo! Inc. (Feb. 18, 2010), https://www.justice.gov/opa/pr/statement-department-justice-antitrust-division-its-decision-close-its-investigation-internet.

125 STUCKE & GRUNES, *supra* note 28, at 186.

126 *Id.* at 201.

127 *Id*. at 183, 201.

128 *Id.* at 201–04; OECD, *supra* note 4, at 107.

129 An important question focuses on what should be considered monopolization and what should be considered competition on the merits. *See, e.g.*, STUCKE & GRUNES, *supra* note 28, at 279 (discussing this issue and comparing U.S. and EU competition laws).

It may be significantly difficult for any competitor to overcome such comparative advantages by collecting the data itself, especially where first-mover advantages and switching costs are high. Yet this difficulty may be overcome if the competitor could combine data collected by numerous sources. Thus, the lower the costs and obstacles to data portability and interoperability, the stronger the potential competitive pressures on large data collectors. And since data are nonrivalrous and often easily replicable, data collectors can share their data with many potential users, thereby potentially strengthening competition even further. The size of this potential competitive pressure is dependent, of course, on the willingness of data collectors to cooperate and/or share their proprietary data even if standardization is required.[130] Furthermore, other barriers may still exist, such as switching costs where past data are important for the user.[131]

Data standardization can also increase competition in markets that are connected by data. Assume, for example, that to create a smart home, data regarding the operation of different appliances must be integrated. An increased ability to integrate such data will facilitate entry into markets for home appliances and smart homes. Of course, a more dispersed market structure might come with its own costs. In particular, intermediary platforms that connect the data gathered from different players could themselves possess market power.[132]

At the same time, it is important to ensure that the standard—as well as other legal tools that facilitate data sharing—do not strengthen those market players that already enjoy scale and scope comparative advantages in data collection and analysis.[133]

3. The International Dimension

To this point, this Article's analysis has been (implicitly) restricted to a closed, domestic market. Adding the international dimension may change some welfare implications of data standardization. This is because domestic and foreign data standards are likely to

---

[130] *See supra* Section II.C.1 for an analysis of the effects of standardization on incentives to share data. Incentives to share may also be affected by the strength of intellectual property rights in data collected. For the importance of creating a consistent data regime which ensures an efficient interplay between data sharing and intellectual property legal regimes, see Inge Graef et al., *Data Portability and Data Control: Lessons for an Emerging Concept in EU Law*, 19 GERMAN L.J. 1359 (2018).

[131] Graef et al., *supra* note 130, at 166.

[132] Michal S. Gal & Niva Elkin-Koren, *Algorithmic Consumers*, 30 HARV. J.L. & TECH. 309, 338 (2017).

[133] Miguel de la Mano & Jorge Padilla, *Big Tech Banking*, 14 J. COMPETITION L. & ECON. 494 (2019) (arguing that dominant platforms are best placed to leverage the explosion of big data on individuals and firms).

affect international competitiveness, which, in turn, changes the domestic welfare calculus.

Localized standards can create entry barriers for foreign competitors and can reduce the interoperability of domestic and foreign datasets. To illustrate, information interface software systems in China must adhere to the standard dot matrix format authorized by the Chinese government.[134] The height of standardization-created barriers is affected by the compatibility of foreign standards with local ones and the ease of switching between them. These barriers can, in turn, affect the value of the data and the industries that are based on it.

More importantly, data standards can affect the international competitiveness of domestic firms, either facilitating or restraining it. Of particular concern is the possibility that inefficient standards will impose high compliance costs on domestic firms and/or limit data integration, thereby constraining domestic firms' ability to compete in international markets.[135] Observe that even when data collected in one country are not relevant to consumers in another, limitations on the use of data could reduce the ability of domestic firms to create better algorithms and could impede transfer learning and its application to foreign data or to data in other markets.

The effects of standards on international competitiveness should not be disregarded. Indeed, the battle over data-driven comparative advantages is no longer confined to private firms. Governments are beginning to realize how data-driven advantages for production, investment, employment, and trade patterns could significantly affect the welfare and well-being of their citizens. Russian President Vladimir Putin put this bluntly, proclaiming that whoever leads in artificial intelligence "will be the ruler of the world."[136] Leading in artificial intelligence requires more and higher-quality data from which the algorithms can learn. While obviously simplistic, there is much truth to Putin's statement. Indeed, recognizing the value of data, governments have started to invest in creating ecosystems for data-driven advantages. China, in particular, has begun seeking to create such advantages for its domestic firms by, inter alia, motivating the creation of huge, comprehensive databases in areas where big data is considered to be of utmost importance (e.g., medical devices, autonomous cars,

134 Burke, *supra* note 73, at 273.

135 *See* SEIZING OPPORTUNITIES, *supra* note 104, at 20 ("The Internet's complexity, global reach, and constant evolution require timely, scalable, and innovation-enabling policies.").

136 *Putin: Leader in Artificial Intelligence Will Rule World*, CNBC (Sept. 4, 2017, 2:33 AM), https://www.cnbc.com/2017/09/04/putin-leader-in-artificial-intelligence-will-rule-world.html.

smart cities).[137] Accordingly, it can no longer be assumed that data races will be mostly played out between U.S. digital giants such as Facebook and Google. Rather, the battlefield is fast changing, with the assistance or the backing of foreign governments.

Furthermore, should the United States not take an active role in examining and, in some cases, possibly even facilitating data standards, American firms might find themselves bound by foreign standards.[138] The reason is that many firms do business in foreign jurisdictions, which may require that data portability and interoperability—even with regard to common commercial acts such as placing an order—be performed in accordance with foreign standards. Such standards might also trickle down to firms that only operate in the United States if other businesses with which they interact employ the foreign standard in all of their relationships. Given that the European Union has recently acknowledged the importance of data standards for ensuring a comprehensive data sharing environment,[139] and its market players are currently in the process of setting such standards in order to comply with portability requirements,[140] there is no time to lose to ensure that domestic data standardization considerations are not disregarded. In some cases, this may be accomplished by encouraging international cooperation in setting pro-competitive requirements for data standardization. Overcoming the tendency of many countries to focus narrowly on short-run domestic interests is a difficult, but not impossible, exercise. Yet in not all cases will international standards serve domestic interests. In some cases, it might be better to focus on domestic standards, while taking into account the externalities created by the interplay among different standards.

III
Government-Facilitated Data Standardization

As the analysis shows, determining whether to set data standards to enable or ease data use and interoperability requires a more varied and complex analysis than the analysis of standards generally. Fur-

---

137 Cheng Yu & Ma Si, *Industrial Internet to Boost Smart Manufacturing*, China Daily (Dec. 1, 2017, 7:35 AM), http://www.chinadaily.com.cn/business/tech/2017-12/01/content_35148829.htm.

138 *See generally* Anu Bradford, *The Brussels Effect*, 107 Nw. U. L. Rev. 1 (2012) (arguing that even "[w]ithout the need to use international institutions or seek other nations' cooperation, the EU has a strong and growing ability to promulgate regulations that become entrenched in the legal frameworks of [other countries]").

139 Article 29 Data Prot. Working Party, Guidelines on the Right to Data Portability 3 (2017), http://ec.europa.eu/newsroom/document.cfm?doc_id=44099.

140 *See, e.g.*, Borgogno & Colangelo, *supra* note 13, at 15, 22 (discussing such efforts in the technology and banking sectors).

thermore, many of the relevant considerations involve externalities imposed by data collectors and users on others. Our analysis also shows that the stakes are high and that considering data standardization is timely, if not urgent. The question thus becomes what role, if any, the government should play in facilitating the adoption of data standards.

A. Potential Market Failures

Can one rely on the market to create and implement efficient data standards? In many settings the answer is in the affirmative, given the large benefits to be had from data standardization. Private endeavors have focused mainly on data portability rather than on data interoperability.[141] Yet, in some settings significant market failures may prevent socially beneficial data standardization. Indeed, as noted, in a recent study firms across many industries perceived the lack of data standardization to constitute a major obstacle to business activity and to the development of innovative products.[142] Accordingly, it seems that at least some data standards are in high demand but in low supply in the data economy. This Section explores some reasons for this market failure.

First, the incentives of different market players may differ and affect the ability to create an efficient standard. Some market participants may favor "a status quo characterized by high costs to switch products and services, greater lock-in and reduced data portability."[143] This may characterize large incumbents who enjoy data-based comparative advantages that cannot be easily matched by others, absent data standardization. By preventing the creation of the standard, incumbents essentially raise their rivals' costs relative to their own. Firms might also not agree on which standard to apply. Indeed, the banking industry exemplifies how "complex and troublesome it could be to ensure a sound and effective adoption" of a data standard across an industry.[144] With the advent of the Internet of Things, the increasing number and heterogeneity of market players is likely to lead to increased conflicts of interest.[145]

Second, even if a standard is voluntarily created, its content may serve the interests of some market players. Concerns arise from the

---

[141] *See, e.g.*, Lynskey, *supra* note 97, at 797–98 (discussing private sector data portability efforts); DATAPORTABILITY PROJECT, http://www.dataportability.org (last visited Mar. 14, 2019); OPEN DATA INST., https://theodi.org (last visited Mar. 14, 2019).

[142] EUR. COMM'N, *supra* note 12, at 88–91.

[143] ROADMAP, *supra* note 61, at 38.

[144] Borgogno & Colangelo, *supra* note 13, at 27–28.

[145] *See id.* at 3–4 (describing the data demands of the Internet of Things).

private interests of those involved in setting the standard, which may lead to strategic conduct resulting from their sunk costs, the relative costs and benefits that the standard imposes on their rivals, or their property rights in the standard.[146] Observe that even subtle standardization choices might make it expensive and difficult for some market players to make use of the data.[147]

Third, collective action problems might lead market players not to create data standards even when it is beneficial for all of them to do so.[148] In the absence of an arbiter, "[t]he market may be sufficiently fragmented that no one approach gains a critical mass of support," leading to a patchwork of inconsistent data standards that slow data flows that might especially disadvantage small data collectors.[149] Furthermore, there might be insufficient time for deliberation before the market sets on its course. Most importantly, the uncertainty resulting from the fact that users cannot be assured that others will follow their move to the new standard creates a coordination problem.[150] Coordination incentives could also be limited by lack of knowledge among data collectors about the data's potential uses and users or concerning the obstacles to integrating it with other types of data. Antitrust concerns, too, could limit incentives to standardize.[151] The creation of efficient data standards might also be inhibited by internal constraints, short-term strategic conduct, or historical legacies.

Even if the standardization that serves the interests of all market players is achieved, the standard might not reflect the social optimum. The problem arises when private standard setters disregard the positive as well as the negative spillovers they create on data subjects, on firms in other markets, and on social welfare. An inherent tension also exists between temporal beneficiaries of data analysis: While

---

146 Stanley M. Besen & Joseph Farrell, *Choosing How to Compete: Strategies and Tactics in Standardization*, 8 J. ECON. PERSP. 117, 128 (1994).

147 *See generally id.* (discussing the financial and strategic difficulties of standardization that extend naturally into discussions of data standardization).

148 *See, e.g.*, ROADMAP, *supra* note 61, at 37 (explaining that, "[d]espite strong agreement" on the benefits of interoperability, the health industry has not yet achieved that goal).

149 Kevin Werbach, *Higher Standards: Regulation in the Network Age*, 23 HARV. J.L. & TECH. 179, 201 (2009); Borgogno & Colangelo, *supra* note 13, at 14–15.

150 *See* Joseph Farrell & Garth Saloner, *Coordination Through Committees and Markets*, 19 RAND J. ECON. 235, 236 (1988) (describing how, absent coordination, different users might adopt different standards that might be difficult to change once adopted and how such a diversity of standards may be sub-optimal for all users).

151 *See, e.g.*, James J. Anton & Dennis A. Yao, *Standard-Setting Consortia, Antitrust, and High-Technology Industries*, 64 ANTITRUST L.J. 247 (1995) (discussing antitrust issues in standard setting); Sean P. Gates, *Standards, Innovation, and Antitrust: Integrating Innovation Concerns into the Analysis of Collaborative Standard Setting*, 47 EMORY L.J. 583, 645 (1998) (examining antitrust enforcement in standard setting).

tomorrow's users may benefit from past data collection, their gains are not always easily shared with the collectors of such data.[152] This may be especially true with regard to transfer learning. Accordingly, private and public incentives to reach efficient data standards are not necessarily aligned.

Market failures may also arise with regard to the implementation of an acceptable standard. The Linked Open Data project[153] provides an interesting example. To solve the metadata uncertainty problem, firms were encouraged to create a "virtual Rosetta Stone" by mapping the attributes of their datasets to other datasets in an attempt to create an ever-expanding "data translator" that would allow transitivity between datasets.[154] The project largely failed, due to its high compliance costs and its voluntary nature: Each firm had a limited incentive to invest the necessary resources unless many other firms did so as well.[155] Moving away from an inferior standard may also be challenging, as the SQL example, elaborated above, illustrates.[156] Accordingly, the market alone cannot always be relied upon to create and implement efficient data standards, even when the benefits to be had are significant.

B. What Role for Government?

These potential market failures strengthen the case for reevaluating the role of the government in data standardization. There is a regulatory role in the acknowledgment, evaluation, and—in the right cases—possible facilitation of data standardization as part of a toolbox for implementing a positive agenda for better use of data. The potential benefits from increased uses of data, as well as the costs accruing from the potential loss of international competitiveness and from the continuing use of a patchwork of (inefficient) standards,

---

[152] Part of the problem involves determining the value that can be gained from the data shared, before all its future uses are known and before other datasets that will be combined with it are known. In a related context, see Cockburn et al., *supra* note 10, at 8, which discusses innovation incentives in the context of "intertemporal spillovers."

[153] Mark Fischetti, *The Web Turns 20: Linked Data Gives People Power, Part 1 of 4*, SCI. AM. (Nov. 23, 2010), https://www.scientificamerican.com/article/berners-lee-linked-data.

[154] Tim Berners-Lee, *Linked Data*, W3 (July 27, 2006), https://www.w3.org/DesignIssues/LinkedData.html.

[155] Peter Neish, *Linked Data: What Is It and Why Should You Care?*, 64 AUSTL. LIBR. J. 3, 4 (2015) ("One of the main promises of linked data is the increased value that the links between entities can provide; however, if there are few other data-sets with which the data can link, then this benefit is not realised.").

[156] *See supra* text accompanying notes 79–85; *see also* Daniel L. Rubinfeld, *Antitrust Enforcement in Dynamic Network Industries*, 43 ANTITRUST BULL. 859, 863 (1998) ("When the dominant firm's product becomes the standard for the industry, firms that are developing alternative standards may find it difficult to compete effectively.").

should act as a catalyst for government agencies to seriously consider data standardization issues. Also, to the extent that data standardization is likely to result from market forces, it behooves government to ensure that it is shaped in a way that serves social welfare. This role is strengthened by the externalities and wide social considerations that come into play. As Kevin Werbach suggests, "[s]tandardization is regulation," given that it "integrate[s] public policy considerations into the technical 'code' of the industry."[157]

Of course, this does not imply that governmental intervention should be lightly considered or that the government should be the one setting the standards. Yet in some market settings, the government may have an important role to play as a direct or indirect facilitator of data standardization. Below, several aspects of this role are explored.

As an initial first-stage effort, regulatory authorities should carefully study market dynamics and characteristics to identify where data standards may create significant benefits that outweigh their costs. Such costs include the costs of standard setting, implementation, and oversight; of compliance with the standard; of lock-in to an inefficient standard; of limited diversity; and of the negative effects of the increased use of some data on privacy and security.[158] It is important to look beyond specific industries to analyze the potential data synergies and externalities in markets in which cross-industry data integration is essential, as in the case of smart cities. The fact that many industries operate in such ecosystems could pose serious obstacles to realizing their potential benefits. The need for study is strengthened by the fact that the current situation is characterized by a patchwork of inconsistent legacy data collection and organizational methods, developed over time by various market players, which are not conducive to data integration. The fact that "[s]witching costs and lock-in are ubiquitous in information systems,"[159] and thus ex post changes might be costly to apply, further strengthens the need for a timely review of such issues.

As part of this initial stage, the government should also analyze alternative solutions to ensure that data standardization is indeed the most efficient tool to increase data interoperability and use. For example, where metadata uncertainties are significant, the regulator

---

157 Werbach, *supra* note 149, at 179, 204.

158 Concerns may include the possibility that regulation will be inefficient, ineffective, insufficiently tailored to differing contexts and needs, or premature. Alexander Macgillivray, *Summary of Comments Received Regarding Data Portability*, WHITE HOUSE (Jan. 10, 2017, 9:19 AM), https://obamawhitehouse.archives.gov/blog/2017/01/10/summary-comments-received-regarding-data-portability.

159 SHAPIRO & VARIAN, *supra* note 123, at 104.

should evaluate whether an efficient "data translator" exists or may be relatively easily developed. Such algorithms, which relate the data attributes of one dataset to those of another dataset, can (partly) solve some of the data integration problems outlined above while significantly reducing intervention in the choices of market players. Another potential partial market solution involves the development of algorithms that reduce the size and quality of data needed.[160] Accordingly, the regulator must evaluate—and even possibly promote—the adoption of such solutions before suggesting data standardization.[161] Finally, where incentives for data sharing are bound to be weak, or the benefits from increased use are bound to be small, the justifications for data standardization are also reduced.

To perform these tasks, governmental agencies with relevant responsibilities must acquire the appropriate technical expertise. They must be able to understand the implications of their decisions on all market players, to evaluate whether industry standards are economically efficient, and to assess whether the market could and would develop timely and efficient standards without governmental intervention.[162] Creating an ecosystem of standards that can work in different contexts, and that can interoperate where required, is likely also to require consultation with industry or even a coordinated governance process that includes the participation of market players.[163] Both suggestions build on the fact that market players often have "substantial knowledge and understanding about both existing technical needs as well as the merits of different proposed solutions."[164] The particular governmental agency that takes the lead in doing so, and the specific agenda it pursues, may vary across industries. Nonetheless, the

[160] To illustrate, transfer learning can potentially reduce the amount of data needed to perform a new task.

[161] For the use of converters to reduce the need for standardization, see Joseph Farrell & Timothy Simcoe, *Four Paths to Compatibility*, *in* THE OXFORD HANDBOOK OF THE DIGITAL ECONOMY 34, 46–47 (Martin Peitz & Joel Waldfogel eds., 2012).

[162] Interestingly, the European Union is engaged in such an experiment. While the law mandates that shared data be "interoperable," it merely encourages market players to develop such interoperable formats and standards. Regulation EU 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1, 13; ARTICLE 29 DATA PROT. WORKING PARTY, *supra* note 139, at 18.

[163] For the implementation of such a tool in the healthcare context, see ROADMAP, *supra* note 61.

[164] Letter from Marina Lao et al., Dirs., Fed. Trade Comm'n, to Karen B. DeSalvo, Nat'l Coordinator for Health Info. Tech., U.S. Dep't of Health & Human Servs. 12 (Apr. 3, 2015), https://www.ftc.gov/system/files/documents/advocacy_documents/ftc-staff-comment-office-national-coordinator-health-information-technology-regarding-its-draft/1504-roadmaphealth.pdf.

National Institute of Standards and Technology (NIST), which is actively engaged in promoting scientific standards in numerous industries,[165] may be best placed to explore the need for cross-industry standards.

Once it is established that standardization will likely increase social welfare, the second stage involves facilitating the creation of efficient data standards. Regulators face a range of options with regard to how standards can be set, each with its own costs and benefits. These include adopting private solutions, establishing standard-setting organizations (SSOs) or facilitating their actions, or suggesting or determining standards themselves.[166] The preferred regulatory model may differ among industries and types of data, depending, inter alia, on the relative competence of different standard setters,[167] the extent of divergence between private and social interests, and the way such a divergence might shape the decisions of the standard setter. Yet it seems that in most cases a supervised delegation to an industry-based SSO, comprised of professional data scientists, will be more advantageous than performing the task by a governmental entity. While regulators play an important role in determining when market failures prevent the creation of welfare-enhancing data standards, they generally have less competence in evaluating the standards that will work best in a given market setting. Where private SSOs are preferred, the regulator may need to set and enforce some basic rules for their operation in order to increase the organizations' ability to create efficient standards.[168] In addition, the regulator should ensure that the broader social implications of data standardization are given sufficient weight.

Once a data standard is agreed upon, the regulator should decide how to facilitate its adoption. Options include setting best practices, mandating the adoption of data standards, and creating soft incentives for their adoption.[169] The HITECH Act, for example, provides finan-

---

165 *Voluntary Product Standards Program*, NAT'L INST. OF STANDARDS & TECH., U.S. DEP'T OF COM., https://www.nist.gov/standardsgov/voluntary-product-standards-program (last updated Apr. 2, 2009).

166 *See generally* Farrell & Saloner, *supra* note 150 (explaining different ways to bring about coordination).

167 On the different organizational forms of SSOs and their comparative competencies, see, for example, Kerber & Schweitzer, *supra* note 116, at 44–48; Lemley, *supra* note 14, at 1898–99.

168 *See, e.g.*, Lemley, *supra* note 14, at 1895 (suggesting an appropriate role for the government with respect to SSOs dealing with intellectual property rules).

169 The Office of the National Coordinator for Health Information Technology, for example, releases an annual list of best available standards, "to be used by technology developers and to inform coordinated governance efforts." ROADMAP, *supra* note 61, at 84;

cial inducements to entities that adopt the new standards.[170] Such payments may be especially important where the costs involved in creating standard-compatible databases are significant. Interestingly, even an indirect threat of regulation may nudge private firms to adopt a standard. It might come as no surprise that the Data Transfer Project undertaken in June 2018 by Microsoft, Google, Facebook, and Twitter, which sets a standard to enable user-initiated data portability among project participants,[171] was initiated amidst increased calls for the government to reign in the power of large digital firms resulting from the control of data.[172]

The analysis also leads to several observations with regard to the content of data standards. First, standards may be important where missing data are a core problem. This is because, while metadata uncertainties and data transformation issues can be (partially) resolved ex post by using better transformation algorithms, the missing data problem cannot be solved so easily. Second, in most cases, measurement, identification, and semantic standards (such as those relating to measurement units, product codes, and terminology) can be easier to determine than those relating to the structure or organization of datasets. Yet such standards can go a long way towards facilitating the efficient use of data, as exemplified by the Mars Climate Orbiter example.[173] Third, the appropriate scope of any standard is likely to vary across industries. It might even be appropriate to impose standards only on some market players in a given industry (e.g., those that collect more than a minimum amount of data). Furthermore, standards should be flexible, enabling them to change with learning, and may vary across industries. Finally, policy solutions should be comprehensive. They should, for example, potentially include limitations on some uses of data as well as stronger cybersecurity protection solutions in order to alleviate both privacy and cybersecurity threats.

The suggestions that have been made are a world apart from the current situation. While some sector-specific regulators, in the healthcare and transportation sectors in particular, have recognized the importance of data standardization, currently no governmental body

see also OFFICE OF THE NAT'L COORDINATOR FOR HEALTH IT, 2015 INTEROPERABILITY STANDARDS ADVISORY 1 (2015).

[170] *See* Health Information Technology for Economic and Clinical Health (HITECH) Act, Pub. L. No. 111-5, §§ 3011–18, 123 Stat. 226, 246–58 (2009) (codified at 42 U.S.C. §§ 300jj–31 to –38 (2012)).

[171] DATA TRANSFER PROJECT, https://datatransferproject.dev (last visited Mar. 15, 2019).

[172] *See, e.g.*, Newman, *supra* note 121, at 6.

[173] *See supra* notes 62–64 and accompanying text.

is exploring the need for a general data standardization agenda or trying to identify those industries in which it might be highly beneficial.

Finally, it is noteworthy that in some situations the government may have no choice but to set data standards. This might be the case where the government collects and organizes data internally (such as meteorological, demographic, or legal data) or where it contracts with others to provide it with certain types of data. Data standards might be necessary for its internal sharing of data, but also for enabling the reuse of certain types of governmental data by other undertakings, in order to enable further exploitation of its economic potential. Setting such standards is likely to raise some of the considerations that are also relevant to the government's broader role as a potential standard facilitator. Moreover, given that governmental data are likely to be shared with numerous industries, standards will need to be constructed to fit across industries. Such standards might then indirectly affect those adopted by the market. Accordingly, the government cannot shy away from this role.

Conclusion

Turning the volume and variety of data amassed by numerous data collectors into valuable assets requires overcoming technological and technical obstacles to creating data synergies and to facilitating increased use of data. One major obstacle involves the use of different standards by different data collectors, creating a "Tower of Babel" that could significantly harm the welfare of individuals, firms, and nations. The rise of the Internet of Things, of artificial intelligence techniques that require vast amounts of data, and of technological environments that must combine data from different sources further intensifies the need for tools that enable cumulative or synergetic knowledge production.[174]

This set of technological issues creates new regulatory challenges that must be recognized and addressed if society is to benefit from the information economy. Accordingly, this Article encourages development of a regulatory environment that recognizes the potential effects of data standards and that is open to taking a more proactive approach towards their facilitation in appropriate cases. Of course, given the costs and risks involved in altering private choices, the government should adopt a cautious approach before intervention. Yet given the high stakes at issue for both private and social welfare, disregarding the concerns raised here or always relying on the market to

[174] *See* Cockburn et al., *supra* note 10, at 24.

create and implement social welfare-enhancing standards should not be an option.