5

Distributed Data Economics

David Shrier

Part I: Foundations for Distributed Data Economics

Distributed ledgers offer new horizons of opportunity for the monetization of data, and new models whereby individual consumers gain more control over and benefit from their personal data, versus the predominant model of today that awards the greatest economic gains to the platform marketing companies such as Facebook and Google, which generally leverage the economics of data extraction. With the market cap of digitally traded tokens of distributed ledger companies exceeding US$240 billion, significant investment in the private sector is supporting the creation of the new distributed data ecology.[1]

Before we can explore distributed data economics, we need to understand where the data is derived, how it is manufactured and how it has been monetized in the past.

Blockchain systems, or distributed ledgers, are fundamentally databases. While a great deal of attention has been paid to the design, architecture,

[1] Coinmarketcap.com accessed March 1, 2020.

D. Shrier (✉)
Imperial College London, London, UK
e-mail: david@visionaryfuture.co

E. Kaili and D. Psarrakis (eds.), *Disintermediation Economics*,
https://doi.org/10.1007/978-3-030-65781-9_5

support, distribution and fundraising surrounding these databases, insufficient attention has been paid to the nature and quality of the data going into these systems, and how that data is being monetized. To paraphrase the chief innovation officer of one of the top banks in Europe, given how problematic consumer data often is in terms of its quality, we run the risk of creating immutable problems.[2]

The Potential of Distributed Data

Imagine a world where consumers dictate how their personal data is used, not a handful of corporate conglomerates. Imagine a world where there is vibrant competition, and choice among service providers, for everything from personal banking to healthcare to energy services. Imagine a world where companies pay consumers directly, instead of marketing platforms, in order to acquire their business. Imagine a world where artists get paid royalties for their work directly instead of most of the profits of small artists disappearing into the coffers of the corporations which run the recording labels and distribution systems. Imagine a world, even, where a public health crisis can be resolved through a nearly automatic collective action by a community to contain the spread of an infectious disease. Virus epidemiology information and gene sequencing could be automatically propagated through a distributed data network in seconds or minutes, instead of the current system which requires layers of human approvals and sometimes sees political intervention at the expense of public health, as with the COVID-19 coronavirus outbreak.[3]

These are all possibilities in a distributed data economy, but there are many obstacles to overcome—not the least of which the historical legacy that surrounds personal data.

The Data Aggregators

Data aggregators grew out of an opportunity to monetize the voluminous data that began to emerge out of the connected world, such as data from payments systems and telecommunications systems that offer rich sources of information about human behaviour.

---

[2] Rutter, K. Panel Discussion. February 2019. London Blockchain Foundation. London, England.

[3] Global Biodefense (2020) "Lab That First Shared Novel Coronavirus Genome Still Shut Down by Chinese Government" February 28, 2020 [online], https://globalbiodefense.com/headlines/chinese-lab-that-first-shared-novel-coronavirus-genome-shut-down/.

These data sources have been generating ever-more-greater volumes of data from billions of individuals at ever-faster rates. Consultancy IDC projects that there will be more than 44 zettabytes of data generated around the world, up from 4.4 zettabytes in 2013.[4] One zettabyte is 2 to the 70th power bytes. If this book were in printed form, and filled with a zettabyte of data, you would have 10 volumes each tall enough to reach the sun.[5]

Who are the titans of this first generation of data aggregation? Acxiom (the relevant division now owned by marketing conglomerate Interpublic Group) was the undisputed Zeus on Mount Olympus of data aggregation. Credit bureaus such as Equifax, Experian and TransUnion join them on this lofty vantage. Not unlike the Gods of Olympus, these data aggregators are extraordinarily difficult to reach, for example if you have a dispute about bad data that entered your record through fraud or identity theft. Yes, there may be a web form that you can eventually puzzle through, but oversight is weak and recourse limited. In some cases, the credit bureaus have purchased collection agencies, which enforce action based on…credit bureau data. Consumers are caught in a self-contained universe if they attempt to dispute a claim. Companies like Plaid, CreditKarma, Mint and MyLife.com now assemble and derive insights around consumer data. Dozens of vendors sell aggregated "anonymized" mobility data, information about how blocks of consumers move around a city, neighborhood or specific location.

Insight into consumers enables one to quickly pierce the anonymity of the crowd. With a few demographic dimensions (age, approximate income, city), an individual can be traced to their home address. Other, more indirect privacy penetrations are possible. For example, researchers discovered that four points of shopping data (such as date and location of purchase) could uniquely re-identify an individual out of millions of records.[6]

[4] Kugler L (2018) "The War Over the Value of Personal Data." *Communications of the ACM* February 2018; 61(2): 17–19, https://cacm.acm.org/magazines/2018/2/224626-the-war-over-the-value-of-personal-data/abstract.

[5] Berkan R (2012) "Big Data: A Blessing and a Curse." *SearchEngine Journal* [online], https://www.searchenginejournal.com/big-data-blessing/53528/.

[6] de Montjoye, Hidalgo, Verleysen, Blondel (2013) "Unique in the Crowd." *Nature Scientific Reports* 3: 1376 [online], https://www.nature.com/articles/srep01376.

The Emergence of Fine-Grained Human Behavioural Data

Fine-grained human behavioural insights can be extracted by understanding the digital traces, or "breadcrumbs", that people leave on ubiquitous electronic networks that pervade every aspect of modern society.[7]

The first modern payment card was issued in 1950.[8] Adoption was slow initially, but began picking up steam as data communications services improved in the 1970s and 1980s. Credit cards are expected to carry more than 850 billion purchase transactions by 2028, up from a current level of 369 billion.[9] In Europe, 2018 alone saw more than US$ 3 trillion of purchase volume.[10] With this growth in payments systems have come insights into consumer purchasing behaviours, derived from the purchasing data, that has proven highly valuable to marketers.

Mobile phones, likewise, have emerged as rich source of human factors data within the past ten to fifteen years.[11] Other data sources began emerging—for example, Catalina Marketing harvested the "scan" data from checkout registers retail stores, generating a fine-grained map of consumer shopping behaviours (albeit one that struggled with the consumer migration to e-commerce).[12] Loyalty programs (earning "points" or "miles") have further generated actionable data on consumers that merchants have used to fine tune marketing.[13]

With the World Wide Web (popularly referred to as the "internet") exploding into widespread adoption in the late 1990s and beyond, a new vehicle was created for the acquisition of personal consumer data.

---

[7] Pentland A (2013) "The Data-Driven Society." *Scientific American* October 2013; 309(4): 78–83.

[8] Steele J (2018) "The History of Credit Cards." Experian Blog March 16, 2018 [online], https://www.experian.com/blogs/ask-experian/the-history-of-credit-cards/.

[9] Nilson (2020) *The Nilson Report* January 2020: 1167 [online], https://nilsonreport.com/upload/Cover_Chart_1167.jpg.

[10] Nilson (2019) *The Nilson Report* June 2019: 1156 [online], https://nilsonreport.com/upload/Cover_Chart_1156.jpg.

[11] Kostas Konsolakis, Hermie Hermens, Claudia Villalonga, Miriam Vollenbroek-Hutten and Oresti Banos Human Behaviour Analysis through Smartphones (2018). *Proceedings* 2, 1243 [online], https://doi.org/10.3390/proceedings2191243.

[12] Springer J (2018) "How the Digital Shift Checked Catalina Into Chapter 11: The 'Big Data' Marketing Pioneer Seeks a Speedy Restructuring." Winsight Grocery Business December 17, 2018 [online], https://www.winsightgrocerybusiness.com/industry-partners/how-digital-shift-checked-catalina-chapter-11.

[13] Wise Marketer Staff (2019) "How Data Analytics Is Transforming Loyalty Rewards Programs." The Wise Marketer March 28, 2019 [online], https://www.thewisemarketer.com/infographic/how-data-analytics-is-transforming-loyalty-rewards-programs/.

This proliferation of consumer data created a virtual feast of digital information for the data aggregators to gorge themselves on. Initially, consumer brand companies such as Proctor & Gamble and Nestlé, hungry themselves for smarter and better ways to market their products, supported this nascent industry with billions in revenue. Over time, other consumer-facing sectors such as financial services and auto embraced this approach to identifying and targeting relevant audiences and individuals.

Oligopoly Platform Companies

Increasingly, oligopoly platform companies such as the BATs (Baidu, Ali Baba, Tencent) and the FANGs (Facebook, Amazon, Netflix, Google) are themselves aggregating and tying together data from disparate sources and offering marketing analytics services to their corporate customers. Continuous location streams from mobile operating systems such as Android and messing apps such as WeChat and WhatsApp enable a very fine-grained understanding of behaviour—and ability to identify not only an individual, but their preferences and even predictions on future behaviours.[14]

Part II: Legacy Data Economics

Data Depletion

A decade ago, the World Economic Forum published a white paper "Personal data: The Emergence of a New Asset Class",[15] coincident with the emergence of the expression "Data is the New Oil". Like oil and gas, data systems represent a long-term asset class with the long-cycle investment required to harvest them and maintenance investment is also required on an ongoing basis. Databases also have an analogous concept to oil and gas reserves: depletion. In the data world, this is commonly referred to as "decay" or "data decay", namely the rate at which information in a database becomes obsolete. As the world of personal data economics has become more complex and interconnected, the World Economic Forum and others are looking at new approaches for creating and apportioning value from using data in new ways.

[14] Bogomolov A (2018) Andrey Bogomolov "Predictive Modeling of Human Behavior: Supervised Learning from Telecom Metadata." Ph.D. Thesis, University of Trento, Italy 2018 [online], https://pdfs.semanticscholar.org/1423/704d2ca219ad657838a6086d34c1cc6030ee.pdf.

[15] World Economic Forum (2011) "Personal Data: The Emergence of a New Asset Class" [online], https://www.weforum.org/reports/personal-data-emergence-new-asset-class.

Approaches such as using federated data to uncover value in latent health information (in turn creating economic incentives to cure rare diseases, for example)[16] as we will describe later in this chapter.

For example, in parts of Europe, as many as 23% of the population has moved within the past 5 years—comparable with one of the most mobile societies, the USA, with a rate of 24%.[17] Factors ranging from employment-driven movement (e.g. Polish workers in Paris) to humanitarian crisis (e.g. Syria) have further accelerated these trends. This means that name-and-address information becomes obsolete.

This leads us to a world where consumer data sets can decay 30% per year or more. For business data in certain markets (e.g. tech job contact details in San Francisco), that decay rate can exceed 70%.[18] If you are seeking to understand society or understand customers, you therefore need to invest not in *data sets* but in *data systems*, that enable you to keep pacing with the rapidly degrading data. Data is a river, not a rock, and should be viewed as a rapidly moving resource rather than a fixed object in space and time. The systems supporting that data should incorporate a mechanism for improving *recency*.

The Value of Personal Data

Once a data system is architected, and means of acquiring and compiling information (let us say, about consumers), what is the value of that data?

Industry Valuation of Consumers

How much, fundamentally, is a person valued economically, from the perspective of data economics?

The answer is of course in the manner in which it is consumed, how it is monetized. As of this writing, an individual is worth on average, globally, $359 per year to Google but $1,793 to Amazon (in terms of revenue).[19] The

[16] World Economic Forum (2020) "Global Data Access for Solving Rare Disease: A Health Economics Value Framework".

[17] Chandler A (2016) "Why Do Americans Move So Much More Than Europeans? How the National Mythos and U.S. Labor Laws Influence Geographic Mobility." *The Atlantic* October 21, 2016 [online], https://www.theatlantic.com/business/archive/2016/10/us-geographic-mobility/504968/.

[18] Brence T (2016) "Customer Data Decay: Why Your Contact Data Is Rotten." Informatica Blog August 3, 2016 [online], https://blogs.informatica.com/2016/08/03/customer-data-decay-why-your-contact-data-is-rotten/#fbid=g6xcTDA0Uu4.

[19] Ngo S (2018) "Here's How Much Google and Facebook Really Think You're Worth." *Showbiz Cheatsheet* [online], https://www.cheatsheet.com/money-career/heres-much-google-facebook-really-think-youre-worth.html/.

average American Facebook user is worth about $220 per year, but EU users are worth only about ¼ as much, perhaps due to stricter advertising regulations—one would expect that it would be much closer to the US revenue, given that per capita income in EU member states like Germany and Norway are comparable to or even greater than the USA on a PPP basis.[20]

Amazon is an interesting case study. While it delivers more revenue volume through its shopping services, nearly two-thirds of its operating profit for 2019 came from Amazon Web Services (AWS), which also grew 25% faster than Amazon's core products business.[21] International business segments are still operating at a loss.[22] And AWS is very high margin revenue—23% operating margin versus 5% operating margin overall for Amazon.[23] What this means is that user data generates a large volume of low-margin revenue for Amazon, while corporate revenue tied to cloud services now comprises the majority of Amazon's profits.

Facebook, on the other hand, runs at a 34% operating margin as of 2019, even after a rise in expenses over 2018.[24] They have been able to successfully monetize user data 790+ % better than Amazon. Facebook's margins are more than double the typical media company.[25] There are some who believe they should be regulated like an oligopolistic media company and not simply a "technology provider" as they would like to be classified.[26]

Indeed, oligopoly platforms like Facebook and LinkedIn demonstrate an interesting aspect of personal data monetization—the "network effect". The more people use these platforms and the more connected they are to each other, the more valuable the platform experience is to users (making them "stickier" and spending more time interacting with the platform) and the more marketers are willing to pay to access these audiences. Two-sided

[20] Dazeinfo (2020) "Facebook ARPU by Region: Q2 2010–Q4 2019" updated January 31, 2020 [online], https://dazeinfo.com/2018/08/23/facebook-average-revenue-per-user-by-region-dgraph/.

[21] Condon S (2020) "AWS Brings in Nearly $10b in Sales for Amazon in Q4, Hits $40b Annual Run Rate," https://www.zdnet.com/article/aws-brings-in-nearly-10b-in-sales-for-amazon-in-q4/.

[22] Ibid.

[23] The Motley Fool Staff (2019) "How Amazon Actually Makes Money." *The Motley Fool* February 19, 2019 updated April 10, 2019 [online], https://www.fool.com/investing/2019/02/19/how-amazon-actually-makes-money.aspx.

[24] Rodriguez S (2020) "Facebook Stock Falls After Showing 51% Rise in Expenses." CNBC.com, https://www.cnbc.com/2020/01/29/facebook-fb-earnings-q4-2019.html.

[25] CSI Market "Broadcasting and Cable Profitability," accessed March 1, 2020, https://csimarket.com/Industry/industry_Profitability_Ratios.php?ind=902

[26] Bell C (2018) "Facebook: We're Not a Media Company. Also Facebook: Watch Our News Shows." Mashable.com [online], https://mashable.com/2018/06/08/facebook-media-company-news-shows/.

networks such as Airbnb or Lyft or Ola have an indirect network effect but still see this power-law value creation curve.[27]

For Facebook, at least, its consumption of user data profits may be reaching the dregs of the bottle. New data privacy laws, repeated cyberhacks, and growing awareness about the relatively weak responses Facebook has given with respect to the use of its platform to promote misinformation, are beginning to shift Facebook's interaction with regulators and policymakers, and may put pressure on its ability to monetize user data.[28] The government backlash against Facebook-sponsored Libra Project,[29] an overt attempt to acquire even more consumer data off its network (this time in the payments arena), illustrates the dangers of a data monetization policy that fails to transparently and rigorously address data ethics. Indeed, the announcement of Libra stimulated a number of governments to accelerate their Central Bank Digital Currency (CBDC) projects with the express purpose of competing with or suppressing Libra.[30] The Reserve Bank of Canada went further and said they would only launch a CBDC if Libra were successful.[31] Companies like Apple, for example, have not stimulated government response to such a degree, perhaps through more astute government affairs efforts coupled with data privacy actions perceived as beneficial to consumers.[32]

Consumer Self-Worth

The converse system is instructive to explore: how much value do consumers attribute to their various personal data elements? Someone who publishes articles on LinkedIn might not place tremendous value on their own name, since it can be found attached to the article they published. Other data elements about an individual are much more sensitive. According to research

[27] Flint P (2018) "70 Percent of Value in Tech Is Driven by Network Effects" [online], https://www.linkedin.com/pulse/70-percent-value-tech-driven-network-effects-pete-flint/.

[28] Guy E (2018) "Inside the Two Years That Shook Facebook—and the World." Wired.com February 12, 2018, https://www.wired.com/story/inside-facebook-mark-zuckerberg-2-years-of-hell/.

[29] Shrier D (2019) "The Future of Money Isn't Libra or Chinacoin, It's Federated" [online], https://www.linkedin.com/pulse/future-money-isnt-libra-chinacoin-its-federated-david-shrier/.

[30] Baydakova A (2020) "Central Bankers From Canada, Netherlands, Ukraine Call Blockchain Unnecessary for Digital Fiat." Coindesk.com February 24, 2020 [online], https://www.coindesk.com/central-bankers-from-canada-netherlands-ukraine-call-blockchain-unnecessary-for-digital-fiat.

[31] Baydakova A (2020) "Bank of Canada Won't Issue Its Own Crypto Unless Libra Succeeds: Deputy Governor." Coindesk.com February 25, 2020 [online], https://www.coindesk.com/bank-of-canada-wont-issue-its-own-crypto-unless-libra-succeeds-deputy-governor.

[32] O'Flaherty K (2019) "Apple Issues New Blow to Facebook and Google with This Bold Privacy Move." Forbes.com November 6, 2019 [online], https://www.forbes.com/sites/kateoflahertyuk/2019/11/06/apple-issues-new-blow-to-facebook-and-google-with-this-privacy-move/#1d9685fc481d.

conducted by the University of Trento, "where I am right now" (a user's location in time and space) is the most "valuable" personal data. Media consumption, at the other end of the spectrum (where you read news or information), is valued little or not at all, and which apps you use falls somewhere in between. This landmark "Money Walks" study also determined that people value their own data more on days, which are outliers, where unusual events or activities are occurring, versus ordinary days.[33]

Generalizations are slippery in the world of personal data values. One has only to look at the disparity among personal data protection laws in Germany, the USA, and Nigeria, to pick three countries, to see consumer sensitivities or lack thereof. With that said, consumers have generally shown a willingness to share information around activities if it will help improve their experience with a product or service. Age and gender also play into how much or little individuals value their personal data.[34] Perhaps unsurprisingly, the Millennial generation of digital natives is more prone to data sharing without remuneration.[35] Some business models have been constructed around tangible benefits for tangible personal data sharing: Waze works so well because Waze users share traffic and other road condition data with each other; Netflix's recommendation engine, a core component of its value proposition, requires that users allow for the cross-fertilization of viewing preferences ("Other viewers like you watched…")—sometimes getting the service into trouble, even when publishing "anonymized" insights into this data.[36]

---

[33] Staiano J, Oliver N, Lepri B, de Oliveira R, Caraviello M, Sebe N (2014) "Money Walks: A Human-Centric Study on the Economics of Personal Mobile Data." *UbiComp '14 Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 583–594. New York, NY: ACM.

[34] Liem C, Petropoulos G (2016) "The Economic Value of Personal Data for Online Platforms, Firms and Consumers." Bruegel.com blog post [online], https://bruegel.org/2016/01/the-economic-value-of-personal-data-for-online-platforms-firms-and-consumers/.

[35] Christofides E, Muise A, Desmarais S (2012) "Hey Mom, What's on Your Facebook? Comparing Facebook Disclosure and Privacy in Adolescents and Adults." *Social Psychological and Personality Science* 3(1) January 2012.

[36] Saltzman M (2018) "How to See Everything Netflix Knows About You." *USA Today* April 17, 2018 updated May 14, 2018 [online], https://www.usatoday.com/story/tech/columnist/saltzman/2018/04/17/you-can-see-what-netflix-knows-you-but-you-cant-download/510782002/.

The Societal Cost of Legacy Data Models

The Economics of Privacy

Economists have, for decades, been exploring the economic cost of privacy. Not unlike filtration systems for water, or public safety patrols by law enforcement, or security gates on buildings, data privacy has associated costs. For example, if a job applicant, for a position working at a pharmacy that dispenses prescription pharmaceuticals, conceals that he or she has been arrested for selling drugs illegally, his or her personal privacy is protected, but the business in question assumes much greater economic risk than it intended (business, regulatory, and reputational). If a private citizen is in a traffic collision, but his or her medical records are locked in a secure system that the first responders cannot access, there may be a direct and deleterious impact on health care if, for example, the individual is allergic to certain medications or has a Do Not Resuscitate (DNR) order on file. If an individual wishes to apply for a loan, certain efficiencies are introduced if a trusted third party such as a credit bureau can provide assurances to the lender (the bank, for example) that the individual has adequate creditworthiness.[37]

For each of these circumstances, there are counter-arguments that suggest an equivalent economic burden. The job seeker may have been falsely accused, and lacked the funds to adequately defend against state prosecution. The person in the traffic collision might have their medical information improperly accessed another time if it is too readily available, and suffer other harm as a result. The loan applicant might be disputing information on the credit file, but be unsuccessful in having incorrect information removed. For example, at least one of the major credit bureaus has purchased a loan collection agency where they buy debts. Even if one wishes to dispute a bureau report, the counterparty who is making false claims…is the same bureau. Or the bureau's algorithm might be discriminatory such as to disfavor the class of people to which the individual belongs.[38] Shoshanna Zuboff and others have railed against the rise of "surveillance capitalism" and the society decay

[37] Acquisiti A, College H (2010) "The Economics of Personal Data and the Economics of Privacy." OECD Joint WPISP-WPIE Roundtable, Background Paper [online], https://www.oecd.org/sti/ieconomy/46968784.pdf.

[38] Acquisti A, Taylor C, Wagman L (2017) "The Economics of Privacy" [online], https://www.ftc.gov/system/files/documents/public_comments/2017/10/00006-141501.pdf.

accompanied by the growth of the oligopolistic data platform companies like Facebook and Google.[39]

Generally speaking, the attitude about data privacy is weighed against public good. Someone's right to medical privacy is typically not superseded by the public's right to be aware of an infectious disease crisis; even if Patient Zero's identity is protected, the public needs to be aware that the disease is spreading from X location, so that steps may be taken to contain the infection. In some domiciles, the identities including photos of sex offenders are published; in other domiciles, this is viewed as punitive rather than rehabilitative and other measures are taken. Yet, where is the line drawn with respect to digital data? If someone threatens on a Facebook posting to blow up a school, that is certainly a safety concern; should public officials take steps around the individual, the school, or both? What if an individual simply states a preference for a political candidate who holds extremist views? In one instance, most domiciles err on the side of caution for society, in the second instance, most domiciles would protect the identity of the individual. But is a posting on a Facebook page truly private? The affirmative statement that protects classes of personal data is one that only recently has come to be codified in statute and regulation, as we will discuss later in the chapter.

Cyber (In)security

Hackers have noticed the value of personal data and have engaged in large-scale data theft over the past few years. Unfortunately, as we point out in our book *New Solutions for Cybersecurity*, these massive data stores have been inadequately secured. The Aadhaar biometric and demographic database of 1.2 billion Indians was breached with an individual record for sale for a reported Rs 500 (about €6.50).[40] Equifax saw its entire US database stolen, about 148 million Americans.[41] As more and more personal data is acquired and analyzed and stored, it creates ever-more-tempting targets for cybercriminals, both those acting purely from a profit motive as well as a growing array

[39] Naughton J (2019) "'The Goal Is to Automate Us': Welcome to the Age of Surveillance Capitalism." *The Guardian* January 20, 2019 [online], https://www.theguardian.com/technology/2019/jan/20/shoshana-zuboff-age-of-surveillance-capitalism-google-facebook.

[40] Tech2 News Staff (2018) "Aadhaar Security Breaches: Here Are the Major Untowards Incidents That Have Happened with Aadhaar and What Was Actually Affected." Firstpost. September 25, 2018 [online], https://www.firstpost.com/tech/news-analysis/aadhaar-security-breaches-here-are-the-major-untoward-incidents-that-have-happened-with-aadhaar-and-what-was-actually-affected-4300349.html.

[41] Electronic Privacy Information Center (2018) "Equifax Data Breach" updated February 13, 2020 [online], https://epic.org/privacy/data-breach/equifax/.

of state-sponsored data thieves. Economic analysis can reveal the trade-offs in terms of the cost of implementing better data security versus the benefits of mitigating societal, business or individual harm.[42]

Part III: New Generation Data Ecologies

We have now established the value of personal data, considered the costs of personal data privacy and impacts of poor cybersecurity, and hinted at some of the opportunity embedded in rich data streams. In this section, we are going to investigate powerful computational tools that assist with positive social change, and the necessary privacy and personal data governance regulations that accompany them—laying the foundation for the distributed data economy. Without appropriate protections, these richer data streams offer a significant challenge with respect to protecting consumers from exploitation.

Social Physics of Personal Data

More than a decade of research by Prof. Alex Pentland at the Massachusetts Institute of Technology and its collaborating institutions has derived a new computational social science of "*social physics*".[43] Placing machine-learning rigor behind Adam Smith's musings on communal good,[44] social physics has uncovered new transparency and insights into society, and has enabled interventions at scale such as mapping how vaccines could reduce the spread of malaria in sub-Saharan Africa[45] and helping a region in central Europe reduce energy usage by 17% to "go green" (enabling them to only use renewables and not rely on fossil fuels to power their homes) at a fraction of the economic cost of conventional methods.[46] These insights have been derived from analyzing anonymized, aggregated datasets consisting of the tiny digital

[42] Garcia ME 2013) "The Economics of Data Breach: Asymmetric Information and Policy Interventions." PhD dissertation. Ohio State University, Columbus, OH [online], https://etd.ohiolink.edu/!etd.send_file%3Faccession%3Dosu1365784884%26disposition%3Dinline.

[43] Pentland A (2015) *Social Physics*: *How Social Networks Can Make Us Smarter* (2nd ed.). New York: Penguin Press.

[44] Smith, A (1759) The Theory of Moral Sentiments. London: Printed for A. Millar, and A. Kincaid and J. Bell.

[45] Wesolowski A, Eagle N, Tatem A, Smith DL, Noor AM, Snow RW, Buckee CO (2012) "Quantifying the Impact of Human Mobility on Malaria." *Science* October 12, 2012; 338(6104): 267–270.

[46] Mani A, Rahwan I, Pentland A (2013) "Inducing Peer Pressure to Promote Cooperation." *Nature Scientific Reports* 3, Article number: 1735.

traces people leave throughout the day by using their credit cards and mobile phones.[47]

This naturally has lead Prof. Pentland and his collaborators to the necessary twin of social physics insights, the domain of personal data privacy. Pentland chaired the World Economic Forum privacy working group that evolved a set of principles he termed the "New Deal on Data".[48] This thinking offers direct lineage to the emergence of new privacy regulations in Europe and elsewhere.

Emergent Regulation: GDPR, PSD2, Open Banking, and the California Consumer Privacy Act

The European Union has been highly sensitive to, and progressive on, the topic of personal data and personal data monetization. Cognizant of the data privacy issues and the impacts apparent from the exploitation of consumer data by private sector interests, the EU has promulgated a body of law and regulation to change the frame.

The first major legislation, GDPR, helps establish basic rights for the individual against the corporate conglomerate, around ideas like ownership of personal data, governance (control) over personal data and "the right to be forgotten". It also introduces penalties for data breaches, a significant step towards helping consumers understand the hidden economic cost of the shadowy realm of data brokers and data aggregators. An interesting artifact of GDPR is that it is enforceable for the rights of European citizens even if they are not physically present in Europe i.e. a vacationing French family in Florida using local internet or telecom services would enjoy the same protections, in theory, as if they were in Toulouse. Lesser known is the fact that GDPR creates a de facto open banking mandate around data portability.[49] In the USA, California passed a similar regulation (the California Consumer Privacy Act).

Its sister regulation, the Second Payment Services Directive (PSD2), enables personal portability of critical data, in this case, bank data. The UK passed a very similar regulation, Open Banking. In each case, the goal is to move away from a model where an oligopoly of large corporations create

[47] Waldawsky-Berger I (2018) "Social Physics: Reinventing Analytics to Better Predict Human Behaviors." Wall St. Journal CIO Blog, https://blogs.wsj.com/cio/2018/09/14/social-physics-reinventing-analytics-to-better-predict-human-behaviors/.

[48] Pentland A (2009) "Reality Mining of Mobile Communications: Toward a New Deal on Data." *The Global Information Technology Report 2008–2009*, S Dutta (ed.). New York, NY: World Economic Forum.

[49] European Commission "What Are My Rights?" accessed March 1, 2020 [online], https://ec.europa.eu/info/law/law-topic/data-protection/reform/rights-citizens/my-rights/what-are-my-rights_en.

insurmountable switching costs for consumers, and towards a model where there is increased competition (accompanied by hopefully lower prices and/or better service) for consumers because their personal financial data becomes portable. An added benefit that delivers second-order economic cost improvement both for consumers and banks is better cybersecurity; API's offering a more robust cyber protocol than the previous market of "screen scrapers" that would pretend to be users and log into different banking websites to collect personal financial data on behalf of consumers. Quite harmonious in principle with GDPR, the open banking mandates that requires personal consumer financial data to become transparent and portable subject to the individual consumer's desires.[50]

In practice, compliance with these two regulations is proving to be challenging for companies.[51] New solutions that incorporate distributed ledger and artificial intelligence may enable not only compliance with GDPR and PSD2, but also enable personal monetization of distributed data for the benefit of the individual, rather than the corporate actor such as one of the FAMGAs (Facebook-Amazon-Microsoft-Google-Ali baba). We will discuss this in the section on the distributed data economy.

Part IV: The Distributed Data Economy

In this final section, we will discuss the potential and the enablers surrounding distributed data. Notice the shift in language as we progress through this chapter, from economics to ecologies to economy. We are self-consciously envisioning an evolution to a higher order of societal organization.

Laying the Groundwork for Distributed Data

We are not quite at a place in the progression of both our technologies and of our legal and business frameworks to harness the full potential of distributed data. Too, we need to improve data literacy if we hope for individuals to be able to appreciate, fully understand and take advantage of the benefits that can arise from distributed data. Before these actions can take place, we have

[50] Manthorpe R (2018) "What Is Open Banking and PSD2? WIRED Explains." Wired.com [online], https://www.wired.co.uk/article/open-banking-cma-psd2-explained.

[51] Mikkelsen D, Soller H, Strandell-Jansson M, Wahlers M (2019) "GDPR Compliance Since May 2018: A Continuing Challenge." *CPO Magazine* [online], https://www.cpomagazine.com/data-protection/gdpr-compliance-since-may-2018-a-continuing-challenge/.

core infrastructure challenges to address to create the necessary conditions for a distributed data future:

A. **First, we need to source more and better data**. Data quality remains one of the unspoken tragedies of the big data revolution. "Water, water, everywhere / Nor any drop to drink".[52] All of those zettabytes of data being produced, much of it personal data, and yet very little attention is paid to the actual quality of the information. Vicki Raeburn, who was Chief Data Quality Officer, shared that she would have to go lie down if she spent too much time considering the current quality of the "big data" that is being proudly discussed[53]… made immutable, thanks to blockchain. To bowdlerize Kirsty Rutter, former Chief Innovation Officer of Barclays UK, when she spoke on a panel about DLTs in the spring of 2019, "immutable [garbage]". Another input into the data equation is the characteristic of data transience (as stated earlier, "data is a river, not a rock"). Particularly as we start to explore social physics, we find that dynamic data flows need new management and technology processes to ensure accuracy of predictions. We need to move data systems to the point of near-real-time analysis, which perhaps can be partially enabled by systems such as OPAL that we will discuss below. Measurement and performance management of data systems has an imperative of managing and mitigating data senescence.
B. **Second, we need better analytics conducted on that data**. At this writing, we are still in the very infancy of big data analytics. Let us recall that advances such as Google-incubated TensorFlow are not even four years old.[54] Newer systems like Endor's artificial intelligence prediction engine are just barely beginning to scale.[55] Hybridized systems that combine human intuition and synthesis with machine analysis are emerging, yet need substantially more research and development before

[52] Coleridge ST (1798) "The Rime of the Ancient Mariner." *Lyrical Ballads, with a Few Other Poems*, 1st ed. London.

[53] In a personal conversation with the author, August 2019.

[54] Dean J (2015) "TensorFlow—Google's Latest Machine Learning System, Open Sourced for Everyone." Google AI Blog.

[55] BusinessWire (2019) "Endor Launches Predictions Protocol to Democratize Access to AI and Data Science," https://venturebeat.com/2019/04/04/endor-launches-predictions-protocol-to-democratize-access-to-ai-and-data-science/.

they become industry standard. New maths are emerging and new capabilities will arise as breakthroughs like quantum computing (today very much a laboratory technology) become commercially available.[56]

C. **Then, we can evolve into fully distributed data networks**. While DLTs are computationally expensive, this is purchased with the coin of economic benefit that can be derived from better sharing of the right information at the right time under the right controls (data governance). IPFS and other hybridized schemes that have "on chain" and "off chain" data storage enable the transparency and resiliency of a distributed ledger with the scalability of massive data sets.

Distributed Data Ethics

Increasingly attention is being paid to the ethics of big data and artificial intelligence, and the complexity of addressing these issues will only increase in a distributed data world. Research ethicists, for example, are raising questions about how conventional university approaches to protecting individuals break down in the face of big data[57]; ethicists have not yet begun to explore seriously what this means in the distributed data world.

Data systems become geometrically more useful with the application of artificial intelligence, and advanced artificial intelligence (AI) systems such as machine learning and deep learning are powered by large volumes of data. Accordingly, the data discussion and the AI discussion quickly converge. Technology scholars such as Luciano Floridi of the Oxford Internet Institute have proposed a framework approach to AI ethics, based on reviewing numerous ethical frameworks and models and converging on 5 pillars of ethical AI[58] summarized below:

(1) *Beneficence*: AI should be doing good for society.
(2) *Non-maleficence*: AI needs to go further than just doing good, it also needs to make sure that it doesn't create harm (a consumer might have a shorter commute thanks to a driving map application, but what if the

[56] Fan S (2019) "Quantum Computing, Now and in the (Not Too Distant) Future." *Singularity Hub* February 26, 2019 [online], https://singularityhub.com/2019/02/26/quantum-computing-now-and-in-the-not-too-distant-future/.

[57] Raymond N (2019) "Safeguards for Human Studies Can't Cope with Big Data." *Nature* 568: 277, https://www.nature.com/articles/d41586-019-01164-z.

[58] Floridi, L, Cowls, J (2019) "A Unified Framework of Five Principles for AI in Society." *Harvard Data Science Review* 1(1) [online], https://doi.org/10.1162/99608f92.8cd550d1.

data about their commute patterns is then exploited to malignant ends?). This is where issues like data privacy and data security come into play.

(3) *Autonomy*: in creating systems of autonomous machines (which become even more difficult to manage in a distributed data economy), we need to ensure that *human autonomy* isn't compromised. People still need to make decisions about things that affect them.

(4) *Justice*: the AI systems should both be fair, and promote fairness. Algorithmic discrimination by AI has been written about extensively,[59] and societal values needed to be embedded into AI-driven systems to ensure that technology aligns with law and morality.

(5) *Explicability*: the decisions made by the AI need to be understandable by humans, which helps ensure accountability for them.

A similar model might be conceived of for a distributed data economy. Indeed, as distributed systems become more widely adopted, an *ethical framework for the distributed data economy* becomes imperative, since distributed systems are intrinsically more difficult to control from a central source. How can you audit the code and activities of thousands of servers in a complex encrypted network, given the difficulties in doing so with a handful? How do you monitor the content of secure, encrypted communications streams that may be making decisions that are adverse to society?

Governance for the New Data Order

The OPAL Project (www.OpalProject.org) provides a federated approach for managing the mechanics of a highly distributed data environment that is nonetheless useful, and allows for better control of the ethical dimensions discussed in the previous section. The basic principles of OPAL focus on improving data security and data governance. They are highly congenial with a distributed data future.

The old style of handling data, which leads to a number of the data insecurity challenges seen with Equifax or Aadhaar, entails accumulating a large volume of data in a single repository, where analytics can then be conducted on it. The information theory of centralization posits that it is more efficient from a computer systems management perspective both to maintain the database and to perform analytics. Economic analysis has shown that when

[59] Eubanks V (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* St. New York: Martin's Press.

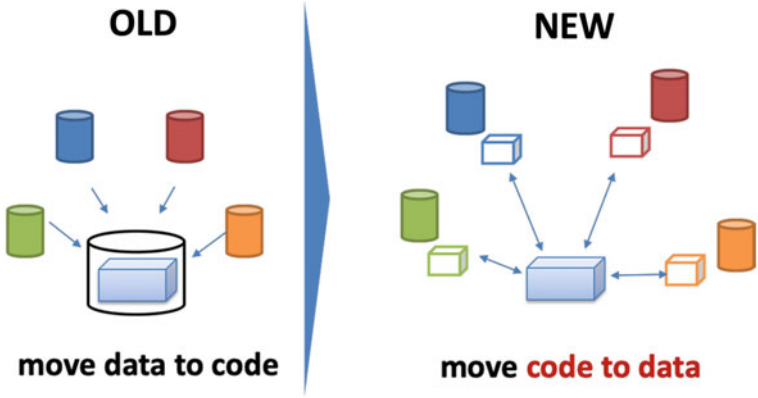

Fig. 5.1 The OPAL method to protect data (*Source* The author)

there is low uncertainty, it is more economically beneficial for a company to centralize, albeit with reduced flexibility.[60]

By bringing the code to the data, instead of the data to the code, we have an opportunity to dramatically improve information security. Contemporary information management systems, notably blockchain or current-generation distributed ledger technologies, provide for a robust code architecture to manage these more complex information flows. Next generation systems that implement near-homomorphic encryption, such as Enigma, offer the potential for exponential improvements on information security and encryption while maintaining sufficient flexibility and accessibility for the data system to be useful in a number of applications (Fig. 5.1).[61]

Distributed ledger technologies appear tailor-made for these types of distributed data systems, and standards like OPAL create a body of coherency around how algorithms and data governance are managed. Bringing these systems together pose a viable platform on which GDPR and PSD2 compliance can be maintained, while extending more control to the users and better security over the data.[62]

[60] Velu C, Madnick S, Van Alstyne M (2013) "Centralizing Data Management with Considerations of Uncertainty and Information-Based Flexibility." Composite Information Systems Laboratory (CISL) Working Paper, http://web.mit.edu/smadnick/www/wp/2013-02.pdf.

[61] Hardjono T, Shrier D, Pentland A, eds. (2019) *Trusted Data, Revised and Expanded Edition: A New Framework for Identity and Data Sharing*. Cambridge, MA: MIT Press.

[62] IBM Security "Blockchain and GDPR" (2018) Cambridge, MA: IBM [online]. https://www.ibm.com/downloads/cas/2EXR2XYP.

Data for the People

A distributed data economy is only possible with sufficient data literacy of consumers to understand, and take action around, the monetization (and protection) of personal data. Market demand to support data aggregators who act on behalf of consumers, a new kind of "data co-op", will only emerge at sufficient levels to support a distributed data economy if consumers attain greater levels of sophistication around how much various types of data are worth, and how those consumers can govern and manage their personal data economics.

The idea of consumer-powered personal data markets isn't new; companies like Datacoup have been trying for years to get critical mass.[63] A common expression in data-aware circles is "If you're not paying for the product – you are the product" (attributed to various individuals).[64] Too few consumers are conscious of this dynamic. The issues of consumer data illiteracy have been (1) the lack of a mandate that empowers consumers around their data (now being solved with regulations like GDPR) and (2) sophistication among users. The UN has highlighted this as a priority within its World Data Forum, with participants stating that "improving data literacy was needed".[65]

The models already exist for providing greater information literacy. Frameworks have been proposed for large-scale community engagement powered by data, posing questions about how data can be used for public good and asserting that control mechanisms built into distributed data, like OPAL, can manage the tension between personal data privacy and benefits to society.[66] Through a concerted set of actions, data literacy can be introduced to different segments of society, and create the necessary ingredients to promote the distributed data economy.

Yet data literacy by itself is insufficient. It's not only data, but *metadata* (abstractions drawn from a collection of or interpretation of data) that powers many of the artificial intelligence models today, and will even more so in the future. Informed consent by users, and therefore user data literacy, needs to

---

[63] Simonite T (2014) "Sell Your Personal Data for $8 a Month." MIT Technology Review February 12, 2014 [online]. https://www.technologyreview.com/s/524621/sell-your-personal-data-for-8-a-month/.

[64] O'Reilly T (2017) "You're Not the Customer; You're the Product." *Quote Investigator* [online], https://quoteinvestigator.com/2017/07/16/product/.

[65] United Nations (2018) "World Data Forum Wraps Up with a Declaration to Boost Financing for Data and Statistics" [online], https://www.un.org/development/desa/en/news/statistics/2018-world-data-forum-wraps-up.html.

[66] Letouze E, Oliver N (2019) "Sharing Is Caring Four Key Requirements for Sustainable Private Data Sharing and Use for Public Good" [online], http://datapopalliance.org/wp-content/uploads/2019/11/DPA_VFI-SHARING-IS-CARING.pdf.

take into account not only the specific data elements an individual might be exposed to a company (for example, the person's location to enable the use of a map application) but also the inferential insights derived from that (wealth, income, and credit score can be estimated based on pattern analysis of a collection of data points related to location[67]).

Distributed Data Policy

Policymakers globally are actively grappling with questions of how to engage with distributed data and how to regulate it, while also managing the potential for innovation and new enterprise formation it contains. Governments around the world, for example, are contemplating CBDC projects, bringing government coffers directly in line with distributed data opportunity. More than 120 national data privacy laws have been put in place,[68] risking a Tower of Babel in the absence of harmonization as data moves cross-border but is regulated locally.

Areas that the European Union, for example, could pursue with respect to data policy include:

Robust Governance. Encouragement by regulators of the private sector use of distributed ledger-based framework approaches, such as OPAL, would help to harmonize activities, streamline oversight and deliver the benefits of standards in terms of market formation and market growth. These efforts in Europe could be tied to the large-scale funding already allocated for blockchain investment.

Adaptation and Innovation Support. GDPR and PSD2 merit active review and augmentation, as would privacy and data portability regulations more broadly (particularly in light of how GDPR and PSD2 have been used as models by other jurisdictions). Now that Europe has had time to see how corporations are attempting to comply in practice, adjustments can be made to the frameworks and the interpretation guidelines. For example, an unintended consequence of GDPR has been to make it more difficult for smaller companies to comply, and introducing new costs and business and financial risks to start-up ventures, although potentially also have created areas of

[67] Pentland A (2019) "Building a Data-Rich Society." *Trusted Data: A New Framework for Identity and Data Sharing*, 109. Cambridge, MA: MIT Press.

[68] World Economic Forum (2020) "Shaping the Future of Technology Governance: Data Policy" [online], https://www.weforum.org/platforms/shaping-the-future-of-technology-governance-data-policy.

opportunity.[69] Steps that regulators and policymakers can take to modulate the effects of these regulations on innovation include:

(1) *Progressive regulatory models*, similar to how some jurisdictions have addressed financial services licensing (e.g. the Bank of England's e-money, "halfway" and "full" banking licenses in an effort to support challenger entry into the banking sector).
(2) More "*sandboxing*" opportunities for start-ups to engage with regulators in a contained environment, where issues can be candidly raised and addressed, and greater compliance capacity within start-ups developed. "Tech sprints" and hackathons around distributed data, data privacy, and data portability, with regulator and policymaker involvement, become another mechanism to simultaneously build government capacity around new technologies and align private sector activity with areas of opportunity.
(3) *Safe harbor exemptions* for defined activities for small and medium sized enterprises (SMEs), so that the cost of compliance does not drive them out of the market
(4) Encouragement of and potentially funding for private sector *compliance-as-a-service* providers, which could also help reduce the costs for individual SMEs while maintaining quality and rigor.

Coordination. Further coordination among the European Union, the OECD, the UN data agencies and the G20, along with other bodies such as ASEAN, African Union, OAS and Caricom, would help mitigate the risks of distributed data policy disharmonization on the one hand, and "jurisdiction shopping" by large corporate interests on the other (which run the risk of undermining data policy). The EU's Gaia-X project, on the one hand, provides some independence from the purely corporate models of virtualized data and data governance such as those offered by Amazon or Microsoft.[70] On the other hand, it's an EU-only initiative at present. Could it be offered to other nations as well? In a distributed data world, data is global, and government response needs to be global, not just regional.

By aligning these areas of Robust Governance, Adaptation and Innovation Support, and Coordination, the disruption that distributed data represents

[69] Martin, N., Matt, C., Niebel, C. et al. (2019) "How Data Protection Regulation Affects Startup Innovation." *Information Systems Frontiers* 21: 1307–1324. https://doi.org/10.1007/s10796-019-09974-2.

[70] Bedingfield W (2020) "Europe Has a Plan to Break Google and Amazon's Cloud Dominance." Wired UK January 27, 2020 [online], https://www.wired.co.uk/article/europe-gaia-x-cloud-amazon-google.

can be better managed, and the economic benefits realized while mitigating potential harms to society.

Acknowledgements The author wishes to thank William Hoffman of the World Economic Forum and Professor Alex Pentland of the Massachusetts Institute of Technology for their suggestions and insights that contributed to this chapter.