# *Chapter* 4

# Demystifying Data

The knowledge economy relies fundamentally upon the ubiquitous surveillance of people, objects and their environments, based on the idea that the more detailed data amassed, including personal data, the more value accrued. In fact, anything less than total surveillance is seen as a deviation from the logic of the market since data that is not observed and measured cannot be monetized. (The same logic holds for states that pursue ever-greater surveillance in the name of security.) Corporate actors have commercial interests in extracting insights from data that they perceive may have economic value. Governments collect and interpret data from state bodies, including statistical organizations, security intelligence agencies and health departments, and purchase data insights from companies that are intended to facilitate the delivery or management of government programmes. Civil-society groups also accord value to data, for example, undertaking campaigns to crowd-source data to identify government corruption, organizing population counts of wildlife or using sensors to measure pollution or industrial noise levels. The economic, social and political value that governments and non-state actors, both companies and civil society, accord to data is emblematic of the information-imperium state, for which control over knowledge is central to the exercise of power. Those who wish to lay claim to this power, however, must possess the resources and capacity to collect and interpret data, which typically requires technical expertise to deal with large volumes of data.

In our knowledge-driven society, the word 'data' is almost talismanic, often evoking fear and awe more than understanding. It doesn't help that there remains great confusion about how data should be treated. As political scientist Dan Breznitz remarks,

> The reality is that we do not even have a decent understanding of how data should be used, who should use it, what technologies it might spawn, who

> should regulate it, who it should be regulated for, or how it should be regulated. (Breznitz 2021, 175)

This chapter is designed to clarify some of the foundational concepts needed to think through Breznitz's questions and to understand the data-driven economy and society that we explore in the remainder of the book. Much of this chapter follows from what we discussed in chapter 1, since data is merely a particular form of knowledge. However, that the concept of data has begun to assume almost mystical powers makes it necessary to describe and define data directly. In particular, we highlight how control over data is a key means of exerting power in a knowledge-driven society.

The economic and social importance that the information-imperium state accords to interpreting data and companies' search for commercially valuable data are evident, to take one example, in the wide variety of fitness and health apps and data-collecting wearables that measure nutritional intake, exercise, sleep and heart rate. Commercial actors that design hardware and software to capture and quantify bodily data promise users accurate, reliable and, crucially, actionable health knowledge that users may employ to address current medical conditions, as well as detecting and perhaps deterring future health problems. Digital sociologist Deborah Lupton (2016) and others refer to this phenomenon as the 'quantified self'. With the real-time monitoring of bodily data, the thinking goes, people will be able to better understand and manage their health. Users are encouraged to take charge of their bodies, changing diet, exercise and health management based on tips from the apps or wearables. However, these individual-level choices of modifying diet or stress levels are often wholly inadequate for people facing complex or chronic health problems, those without access to health professionals or those who face structural obstacles of poverty, racism and discrimination (see, e.g., Lupton 2017).

The commodification of bodily data is particularly evident in the 'femtech' market, a broad array of apps and services devoted to monitoring fertility, menstruation and pregnancy, as well as nutrition, fitness and sexual wellness (see, e.g., Thomas and Lupton 2016; Corbin 2019). Menstrual-tracking apps, such as the popular Flo, Glow and Clue apps, ask users to record their sex drive, diet, moods, the state of their skin, workouts, constipation, cervical mucus quality, masturbation frequency and basal body temperature to identify ovulation. If users become pregnant, they are encouraged to enter details of their sleep, diet, emotional state, weight, the appearance and colour of their cervical fluid, and even when and in what positions they have sex. Information collected on birth includes birth type, length of labour, birthing complications like haemorrhage and in the case of pregnancy loss, the date and type of loss, like whether the baby was stillborn (Harwell 2019). Data collection intensifies after birth as parents can monitor babies and children throughout their

childhood, including using sensor-embedded clothing to monitor infants' respiration, pulse rate and blood oxygen levels and set alarms to detect any irregularities to heartbeats or breathing (Bonafide et al. 2017).

Health technologies like fitness or menstruation apps can be useful tools, but our point is that the purposes and processes of data-collection matters: not all data-collection efforts are necessarily beneficial. In the case of data-driven health technologies, for instance, the shift towards quantifying health has introduced 'a host of new challenges and limitations, such as new selection and other types of biases' (Sharon 2018, 2). Wearables, for example, have had difficulty measuring heart rates in people with darker skin as the optical sensors work better for paler skin (Hailu 2019). Apps and wearables may not be able to capture measurements precisely or universally, and ordinary users may not appreciate the difference between advice from qualified medical professionals and app-derived health advice. There are also critical questions of privacy and individual consent, as some companies like Amazon are requiring employees to use wearables to track worker productivity in warehouses or, in the case of the trucking and construction industries, worker safety. The data economy's imperative is to exploit data, even sensitive health data. This practice is evident in the US Federal Trade Commission's finding in 2021 that the fertility app Flo misled users about its disclosure of users' health data to Facebook (Federal Trade Commission 2021b). The commodification of health data poses additional security risks in the wake of the US Supreme Court's overturning of *Roe v. Wade* and subsequent criminalization of abortion in many states as privacy experts warn that law enforcement could use app data to identify users within or even transiting through the United States whose pregnancy starts and then stops (Hu 2022).

Data is a core constituent element of 'smartness', whether for health technologies, the algorithm-driven gig economy or smart cities. In the case of smart cities, data is integral to delivering the seamless integration of digital and physical infrastructure and the responsive delivery of services like transit, energy, waste disposal and communications. To integrate infrastructure and provide essential services, data-collecting sensors enable 'ubiquitous trackability' of people and objects within the urban environment (Koops 2014; cited in Edwards 2016, 39). For those with a technological solutionist mindset, data is also regarded as an essential component to solving even intractable complex social problems such as unaffordable housing or deteriorating infrastructure (see Kitchin 2014b; Morozov and Bria 2018).

Technology vendors tend to portray smart cities as a way to 'rationalise the planning and management of cities' (Shelton et al. 2015, 13) through the pervasive accumulation and application of data (see Kitchin 2014b; Sadowski and Bendor 2019). In the designs for the ill-fated Quayside neighbourhood that was pitched as the most advanced form of the smart city, for example, Sidewalk Labs proposed heated sidewalks, autonomous vehicles, self-driving

garbage bins and package-delivering robots. These plans were data intensive. Sensors in sidewalks would be responsive to weather, activating heaters when cold wet weather is detected. Self-driving garbage bins' volume sensors would detect when bins should be emptied, while their optical sensors would enable them to move to disposal centres to empty themselves (Sidewalk Labs 2019d, 79). Another set of sensors, designed to regulate mobility within the neighbourhood, would collect data on the presence of pedestrians and cyclists, vehicle and bicycle volume and speed, with real-time monitoring of the locations of app-connected taxis, ride-hail vehicles, bicycles and electric scooters to optimize transit usage and provide real-time information on weather and traffic conditions (Sidewalk Labs 2019d, 50).

Data collection and even surveillance are natural human activities necessary for any functioning society, whether in addressing transit problems or improving maternal health. Problems arise, however, depending upon how surveillance is undertaken and by whom, how the data is treated and who benefits from the surveillance activities and who bears the risks. Public health surveillance undertaken during the Covid-19 pandemic, for instance, again demonstrated that racialized people and those who are marginalized often experience disproportionate levels of state surveillance in comparison with other populations with similar behaviour. In the United States, studies of those arrested for violating Covid protocols, such as social distancing requirements, found that those arrested were disproportionately Black or Latinx (Sundquist 2021).

Data is necessary for sound policymaking, but accessing data in a usable form can be challenging, not just technically but also politically. This challenge was evident in a legal battle over health data between the provincial government of British Columbia, on Canada's west coast, and a coalition of Indigenous Tribal Councils. Indigenous leaders from these councils demanded access to Covid-19 datasets collected by provincial health authorities pertinent to their territories so that they could determine the necessity of stay-at-home orders and resource sharing with other Indigenous nations, arguing that without detailed case counts and locations they 'are working blindfolded' (Slett and Sayers 2020). The province repeatedly denied these requests, stating that sharing the data would violate privacy laws (The Canadian Press 2020). This case clearly shows the power of being able to control, interpret and make decisions using data. In this case, Indigenous leaders claimed that the BC government's actions reflected 'a colonial refusal to share information' (Slett and Sayers 2020).

Data, as these examples show, has emerged as a flashpoint for widespread concern over governmental and corporate power. To explore how data has become a means of exerting power in a knowledge-driven society, this chapter first offers a definition of data as an entirely human-constructed form of knowledge. It then briefly considers two different types of data, personal and

non-personal. Then, it offers a sketch of our current data-driven economy and society. Building on the eight ground rules for understanding knowledge that we explored in chapter 1, it highlights eight key characteristics, and one inconvenient truth, of the data-based society as it currently exists.

## DEFINING DATA

As chapter 1 laid out, data is a form of knowledge. Often, data is used interchangeably with 'information' or is treated as a building block for knowledge, which is seen as involving a deeper, more complex understanding of the world. Our decision to equate data with knowledge is designed to highlight the fact that data itself is created by human action. It involves an interpretation of an underlying or not-wholly-accessible reality, which we term information, the real 'raw material' from which data is created.

Data can never give us a full picture of reality. It is always and everywhere shaped by our necessarily limited modes of perception and our decisions about what aspects of a particular phenomenon to observe, capture (as data) and interpret.

More precisely, data is the knowledge that data collectors perceive as somehow valuable, interesting or worthy of collecting and using. Makers of fitness wearables, for example, decided that measuring users' sleep patterns and daily activity levels provides useful data about users' health. However, many wearables and fitness app companies initially did not capture data about pregnancy or nursing, an oversight that likely reflects the male-dominated software development industry, but also aptly highlights how data creation is a partial representation of reality (Conditt 2019). Because human decisions about the value of certain information result in the generation of data, there is no such thing as 'raw data' (Gitelman 2013, 2). Data must 'be imagined as data to exist and function as such' (Gitelman 2013, 3). In other words, data does not exist independently from human actions, and once collected, data must be interpreted for it to have meaning and value.

### Two Types of Data: Personal and Non-Personal

Personal data generally attracts the greatest media and policymaker attention, as evidenced by the recent spate of data-protection laws in countries worldwide in the last several years. This attention is unsurprising considering the harm that can result from the leaking or theft of people's sensitive personal data, but personal data is only one category of data, the other being non-personal data. Non-personal data covers things such as data observed from industrial processes like the manufacture of pharmaceuticals or commercial

buildings' tracking of energy and water consumption and presents its own set of policy challenges.

Personal data relates directly or indirectly to an identifiable individual. The EU's General Data Protection Regulation (GDPR), which came into force in May 2018 and which is generally seen as the world's most developed form of data regulation, defines personal data as 'a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person' (European Parliament 2018 Art 4(1)).

Personal data is a category that includes data from a variety of sources. A primary source is user-submitted data (often referred to as *volunteered* data) that people generate when they use fitness wearables, gig economy applications like Airbnb or Uber, or social media platforms, or when they access online government services. 'Volunteered' is something of a misnomer, as people may have few options but to provide personal data to access necessary products and services.

Personal data also includes *observed* data, which captures individuals' actions and behaviour, such as the collection of geolocational data from public transit use or from cell phones. Even more indirectly, personal data can be *inferred* by analysing other data to create inferred data. Credit scores are a common – and highly consequential – form of inferred data. Financial institutions construct credit scores by analysing an individual's income, spending habits, debts and other information to build a profile of their creditworthiness (Lauer 2017).

In a society in which data is a commodity that can be bought and sold, it is very difficult – and in the absence of regulation practically impossible – for individuals to understand how the data that they volunteer, or that is collected via observation, is used. Data freely given for one purpose – for example, data collected in the course of a job application or a DNA test to trace one's family tree – can become inferred data used for different or previously undisclosed purposes, such as whether an individual (or others deemed via data analysis to be like them) qualify for life insurance.

In contrast to personal data, and as noted earlier, non-personal data is a category that covers a vast range of information, including weather and environmental conditions, as well as industrial activities. The gas and oil industry, for example, relies upon sensors to detect pipeline leaks, and shipping companies track vehicles and packages in real time. Similarly, medical systems may rely upon internet-connected monitoring, diagnostic and treatment devices to share data among healthcare providers and insurers (DeNardis and Raymond 2017). From an industry perspective, data-driven tools like the sensors discussed here are perceived to be essential to making businesses more effective and productive as they can capture data that may be

used to reduce waste in production processes or identify possible new product lines (see Srnicek 2017).

While personal data raises privacy concerns, non-personal data presents a separate set of policy challenges, primarily related to data control and ownership. Entities, usually corporations, with the technical and commercial infrastructure to collect and extract meaning from data can leverage the skills to dominate industry sectors or to make a private corporation that collects this data indispensable to, say, a municipal government that wants to understand traffic patterns for planning purposes. For example, as chapters 6 and 7 explore, companies may establish data monopolies designed to crowd out other actors that depend on access to data for their own activities. A growing business for big agricultural firms like John Deere is capturing farm data – from sensor-studded tractors driven by farmers – and then selling to farmers insights such as soil or harvesting conditions. Traditionally, farmers painstakingly collected this information themselves, viewed it as a form of traditional knowledge about their lands, crops and livestock, and treated it as proprietary property important to farming as a business. In the data-driven economy, however, companies, not farmers, largely control agricultural data at the cost of farmers' autonomy and in a manner that increases the structural power of big agri-data firms in the agricultural industry over other industry actors, as chapter 7 explores.

Addressing these challenges, including whether to promote such monopolies for favoured domestic industries and how to limit the harms resulting from such control, is a key challenge for the information-imperium state.

## The Datafication of Everything

It's readily apparent in everyday life that more types of data are being collected from people, objects and the built environment than at any time in the past. Two key drivers of this phenomenon are digitization and datafication. Digitization is the conversion of information into binary code readable by computers. Datafication, a term popularized by scholars Victor Mayer-Schönberger and Kenneth Cukier (2013; see also van Dijck 2014), entails capturing a phenomenon in a quantified data format so that it can be recorded, analysed and accorded value. Datafication 'necessitates a desire to quantify and to record' (Mayer-Schönenberger and Cukier 2013, 78) a wide range of phenomena that formerly weren't captured or measured as data. Locations of people and objects are now routinely tracked: for example, internet-connected thermostats gather data on the detected motion within a residence, ambient light levels, temperature, humidity, heating and cooling usage, and carbon monoxide and smoke levels. Human experiences and interactions are a particular focus, as datafication also involves the process of

quantifying social interactions into data to make inferences about behaviour, largely for commercial purposes (see van Dijck 2014).

Just as more phenomena are being datafied, there is a growing array of actors involved in the collection, storage and use of data. Intensive data collection used to be the exclusive purview of states through tools like the census. Now, however, civil society, for example, can work with governments or industry or act alone to collect data on any number of issues, from pollution levels, populations of migrating birds and incidents of government corruption to the number of children travelling to school by bicycle.

Some data may be shared between governments and the private sector, but there are often legal restrictions on sharing specific state-collected data, especially data in sensitive areas like health, taxation or security. This, however, is not always the case. As we discuss in chapters 6 and 8, companies and governments have become increasingly interdependent in their data practices. This interdependence is captured by the concept of the information-imperium state which involves both state and non-state actors as key decision-makers who not only compete but also cooperate in the exercise of structural power. With different actors involved in the collection, processing and use of data, it can be difficult to distinguish public-sector data from private-sector data, further complicating data governance.

The growing centrality of data to the global economy is evident in what scholars alternatively term 'data capitalism' (West 2019), 'surveillance capitalism' (Foster and McChesney 2014; Zuboff 2015), the 'information-industrial complex' (Powers and Jablonski 2015), 'platform capitalism' (Srnicek 2017), 'informational capitalism' (Cohen 2019) and the 'sensor society' (Andrejevic and Burdon 2015). Common to these concepts is the identification of a massive expansion of surveillance systems and data-collection practices, as well as a focus on control over data. For corporate actors, datafication typically results in business models built upon the commodification of data, undertaken through contractual terms-of-service agreements and under the protection of intellectual property (IP) laws. Companies prefer to treat the data they collect as proprietary, from which they will extract value, even when the data originates in the public realm. Companies' practices of capturing the lion's share of accrued value over data can be understood as what Science and Technology Studies scholar Kean Birch terms 'data rentiership', which entails the transformation of data into an asset, that is, the 'assetization' of data to extract value from data (Birch 2020). Concepts like data rentiership and assetization are part of a broader scholarship that emphasizes the proprietary control over the accumulation, ownership and interpretation of knowledge, including IP (see, e.g., Drahos and Braithwaite 2002). Before being monetized, however, actors must identify and capture as data the information that they believe is of potential economic or social value, as we describe further next.

### Data Is Political

Data does not exist independently of people. The processes by which information is conceptualized as data to be collected, stored, processed and used are inherently political. A full understanding of data requires a focus on human actors and the power relationships at play in how we understand and use data. Decisions about the production and use of data are subject to power struggles. Equally importantly, 'data is generative of new forms of power relations' (Bigo et al. 2019, 4).

Decisions to collect and use data are undertaken within specific thought systems that set out what data is determined to be valuable and what devices and technologies will capture that data (Kitchin 2014b, 9). In other words, how we understand data is 'framed technically, economically, ethically, temporally, spatially and philosophically' (Kitchin 2014a, 3). As Science and Technology Studies scholar Yanni Alexander Loukissas (2019, 14) notes, data is not universal. Data does not necessarily 'travel' well: data practices and data themselves differ from one context to the next. Tech companies that operate transnationally, for example, may also transfer or store data outside the country in which it was collected. Governments, however, may want to have data stored and governed within the jurisdiction of collection for reasons of national security or to boost the domestic data economy, thereby conflicting with big tech companies. In response, tech companies like Tencent, Alibaba, Amazon, Microsoft, Google and Facebook have heavily lobbied countries, including India or Indonesia that were considering rules that would require data to be stored within the country of collection (i.e., data localization rules) that would conflict with companies' preferences on transnational data flows (see, e.g., Basu et al. 2019). As debates over data localization (a topic we explore in more depth in chapter 9) show, local context matters. There are always a politics and a culture at play.

## DATA'S EIGHT CHARACTERISTICS

Data requires human deliberation to conceptualize the collection of particular information as valuable or important, such as people's gaits, facial expressions or real-time locations of public transit vehicles. In other words, data is subject to politics, and the laws and norms shaping the identification, collection and use of data reflect the historical, social, political and economic influences of the era. Our current data-intensive economy and society are no different, reflecting specific state- and market-based interests and rationalities driving and shaping the mass accumulation and the use of data. With this in mind, we now turn to data's eight characteristics and one inconvenient truth.

## Characteristic 1: Data Is Not Neutral

One of the most pernicious assumptions at play in policy circles is the idea that data is objective, untainted by human norms or bias. Relatedly, technicians and engineers who design and build data-collection and data-intensive technologies often portray themselves as somehow separate from how their creations are used and from the ensuing consequences.

As we explained in chapter 1 and as scholars from critical data studies and Science and Technology Studies have long pointed out, data is not neutral (see e.g., Kitchin 2014b). How we understand and treat data, including decisions to monetize personal data, reflect specific social, economic, legal and technological ideas within particular societies (see Kitchin 2014a). As the following chapters discuss, states and private actors, particularly large corporations, understand and treat data in ways that reflect specific mindsets.

Bias can be entrenched within the design and operation of technology, thereby affecting what information is considered data and how it is used and valued, as well as what populations are deemed more necessary for intensive monitoring. Scholars and activists have long highlighted bias and discriminatory features designed into software, particularly anti-Black racism (Daniels 2013; Noble 2018). For example, automated speech recognition systems developed by companies like Amazon, Apple and Google have been found to be more accurate in identifying voice commands from native English speakers in the United States than speakers with non-native English accents, and the assistants also have a racial bias in understanding African American speakers compared with white speakers (Koenecke et al. 2020). This 'accent gap' (Harwell 2018) highlights a lack of diverse voice data in training datasets. More broadly, software accuracy problems and lack of training dataset diversity reflect institutional decisions about what data, populations and technologies are considered more commercially important than others.

## Characteristic 2: Data Is a Product

The collection and use of data are fundamental to the proper functioning of software-facilitated products and services. Automatic thermostats, for example, can only work if they can measure the temperature in your house; sensors designed to measure soil moisture need to detect moisture levels. Such data is valuable, and not only because it allows the thermostat to regulate the temperature. Data has become valuable, in and of itself, as a product separate from its purely instrumental purpose. Business models built upon data extraction have become increasingly common, collecting and parsing vast amounts of data from their users. The 'platform' – companies such as Uber, Google, Facebook and even industrial companies like Rolls-Royce, which embed sensors in their

products to track their usage – are designed explicitly around the imperative of collecting as much data as possible (Srnicek 2017).

From the user's perspective, utilizing 'Google maps or hitting the "like"-button on Facebook . . . are not motivated by the intention to produce data, but rather to get directions and to signal approval respectively' (Grabher and König 2020, 105). This is not how companies see things. As Andrew Ng, founder of Google Brain project and former chief scientist at China's Baidu, explains, in a data-driven economy, tech companies 'launch products not for the revenue but for the data' and then 'monetize the data through a different product' (Lynch 2017).

Data, for them, is a fictitious commodity, to use Polanyi's term. Data is not 'produced for sale', but is 'brought to market' (that is, commodified) by companies. The problem with data as a fictitious commodity is not the fact of data collection – you need to provide your location to get Google Maps to get you to your destination, after all, and we want our sensor-operated thermostat to turn the furnace off when our room reaches a certain temperature – but rather when it is repurposed away from the reason for which it was produced. This repurposing is done not in the interest of the individual, but of the actor employing the data for another product or service. By commodifying their users' personal data, companies produce 'surveillance assets' to generate revenue with the goal of influencing and predicting consumer behaviour (Zuboff 2015, 81; see also West 2019). Seen in this way, Google Maps is not a map app but a data-collection mechanism that looks like a map (Zuboff 2019). The purpose of the app is to collect data; the service delivered is a means to an end. Music-streaming services like Spotify deliver 'listening as a service' in which the listening audience is commodified. Nor is this data-based platform economic model limited to the online space. For example, an executive with Vizio, a California-based television manufacturer, said that customers can opt out of data collection, but that if they did so, companies 'would have to charge higher prices for hardware if they didn't run content, advertising, and data businesses' (Patel 2019). Commodifying data, as this statement makes clear, is at the heart of the Vizio business model.

## Characteristic 3: The Centrality of the Proprietary Control of Data

Closely aligned with the treatment of data as a commodity is the impetus to retain proprietary control over data – to keep it within the organization, so that the organization can extract the maximum amount of value. Simply put: 'Whoever controls data, controls the world', an oft-quoted statement popularly attributed either to Jack Ma – former chair and one of the founders of

Alibaba Group, a tech giant in China – or Masayoshi Son, CEO of the Japanese internet, energy and financial conglomerate SoftBank Group (Pfluger 2019). The business models of traditional manufacturers are shifting to emphasize monetizing data. John Deere, for example, is not only one of the largest manufacturers of agricultural equipment. It is also a data analytics company that sells access to data on soil and crop conditions.

Typically, proprietary control over data is contrasted with open-data frameworks, in which data is publicly accessible for anyone to use. However, open-access policies are not a panacea when it comes to issues of control. It takes skill and resources to process and use data, no matter the sources. Larger companies, with human resources and advanced technical infrastructure, including data analytics capacity, possess advantages over start-ups lacking these capacities. Google, Facebook, Amazon, Tencent and Alibaba are amongst a new generation of actors whose business models focus on the accumulation and monetization of data. Those who 'are able to collect data from multiple sources, aggregate it, and do innovative things with it' (Mayer-Schönenberger and Cukier 2013, 135) benefit economically from data and, equally importantly, the authority to create rules regarding the use of the data.

Beyond companies, states have long sought to monopolize data collection, analysis and use relating to the populations within their territories, linking the control over data with state sovereignty and, thus, control over their territory (Kitchin 2014a; Ruppert et al. 2017). State monopoly on data production, however, has been increasingly challenged by companies active in data collection and analytics. As chapter 8 explores, states may work with private actors who provide the hardware, software or data expertise to monitor populations or deliver services such as social assistance or protection of at-risk children. In other situations, governments may have interests in maintaining a 'monopoly of interpretation' (Baack 2015, 4), in areas of strategic interest to the state, such as national security.

## Characteristic 4: The Surveillance Imperative

Data must be observed to be created and collected. The rising importance of knowledge in the form of data to the economy and all facets of social life necessitates constant surveillance of people, objects and their environments, whether to maximize state or personal security or to maximize profits. While individuals, companies and states have always engaged in data collection, the ubiquitous nature of surveillance has changed its goals and effects. Traditionally, surveillance has been understood as 'purposeful, routine, systematic and focused attention' intended to control or manage specific individuals or populations, such as who pose a risk to public safety (Lyon 2015). Now, however, surveillance is increasingly being broadened from focused attention on targeted individuals to systems of pervasive continuous surveillance.

The objective of surveillance in such a 'sensor society' is to capture 'a comprehensive portrait of a particular population, environment, or ecosystem (broadly construed)' (Andrejevic and Burdon 2015, 23) to enable the identification of patterns in data to understand and, more importantly, anticipate actions to predict consumer behaviour and, for state actors, to monitor and control populations.

Both states and corporate actors in the knowledge society are driven by the ubiquitous surveillance imperative. From social media platforms and smart cities to software-enabled Internet of Things (IoT) products, companies have increasingly adopted business models reliant upon the normalization of pervasive, continuous surveillance of consumers, as chapters 6 and 7 explore.

With respect to national security, while China is the paradigmatic example of state surveillance with its systems of online and real-world surveillance, all states have interests in surveilling and controlling their populations, as chapter 8 argues. As the US global surveillance system revealed by Edward Snowden demonstrated (Schneier 2015; Lyon 2015; Greenwald 2014), such surveillance is not unique to authoritarian countries but is evident in all countries that define security in terms of the amount of data to which one has access – that is, countries that embrace the logic of the information-imperium state.

## Characteristic 5: Data Collection Is Speculative

The drive towards total surveillance is complemented and reinforced by the assumption, or belief, that the value or use of some data may only become clear in the future. Such data is seen as useful not only for the development of new products or services but also in terms of safeguarding national security. This perspective introduces, in turn, a bias towards data overcollection, lest you miss out on data that later turns out to be valuable. Or, worse, that someone else collected that now-useful data.

As a result, data-intensive companies tend to operate with a 'collect-it-all' mentality, with the goal of generating 'new patterns of correlation' that can be repurposed indefinitely (Andrejevic and Burdon 2015, 23–24). This data-maximalist attitude is complemented by a drive to maximize surveillance, to minimize privacy and to engage in expansive data-collection practices that amass more data than required for the effective operation of current products and services, often without the knowledge or consent of customers. Data-maximalist mindsets are evident in Silicon Valley's 'move fast and break things' ethos, which condones, among other aggressive business practices, the all-encompassing collection of data even without users' permission with the idea that specific uses will be determined later.

Google's attempts to map the world offer a particularly egregious case of this collect-it-all (no matter the legality) mentality. Between 2007 and 2010,

Google deployed Street View mapping vehicles around cities worldwide, capturing panoramic digital images of neighbourhoods and collecting Wi-Fi network data to provide location-based services like mapping (Federal Communications Commission 2012). Google also illegally captured the content of internet communications, including email and text messages, and passwords. Over a dozen countries investigated Google for violation of their privacy laws. Google belatedly admitted to the US Federal Communications Commission that the illegal data collection was a 'deliberate software-design decision' made by Google engineers working on the Street View project (Federal Communications Commission 2012, 2). Illegal data collection in this case was not a bug; it was a deliberate decision to collect potentially valuable information to create new products.

As chapter 8 examines, states also exhibit data collection–maximalist tendencies typical of the information-imperium state. Recall that the information-imperium state is characterized by an overarching emphasis on the capture and control of knowledge, in this case data. National security is a particular focus of states' speculative, future-oriented, data-driven surveillance. US national intelligence agencies, for example, call upon the private sector for ways to improve facial-recognition technology, especially by strengthening identification with other technologies, including 'whole-body identification, gait recognition and/or anthropomorphic classification (e.g., height, gender)' (Kimery 2019). The drive towards ubiquitous data collection is not just a characteristic of state security services but of the state as a whole, in the name of delivering services like health, immigration and social assistance programmes.

The speculative, data-maximalist approach characteristic of the information-imperium state and the data-driven society stands in stark contrast to calls for a 'data-minimization' approach to commercial and state activities. Such an approach calls on organizations to collect, use or share only the personal information that is necessary for the purpose at hand and not to collect and use personally identifiable information if other information could serve the same purpose (Cavoukian and El Emam 2014, 4). While the data-minimization approach is intuitively appealing because it is designed to maximize user privacy, its implementation, like the exhortation to reduce IP protections to encourage innovation and cultural creation and consumption, is a hard sell in a world in which the control over data is a key element of political, economic and social power.

## Characteristic 6: The Presence of Asymmetries of Knowledge

Anyone who has been surprised by the Instagram ad that appeared in your feed advertising a TV show that you'd been talking about with your friends

or felt queasy about the data profiles that data-hungry social media platforms have created about you intuitively understands the chasm between ordinary users and data collectors. This gap, which some scholars term 'asymmetries of knowledge' (West 2019; Zuboff 2015, 2019), refers to the difference between what data companies know about their users and how little people know about how these companies use their data. 'Data-poor' actors have little understanding of the inner workings of data actors' data-collection capabilities, how or where data is stored and used, and the short- and long-term consequences of data commodification and monetization (Andrejevic and Burdon 2015). Even when data may be freely available, such as when a city provides open data on public transit, data-poor actors often do not have the expertise or resources to make sense of or use such large volumes of data.

Here, it's important to understand that bits of data on their own – say, data collected on an individual – have little value. It's only when that data is collated with many other data points into large datasets that it becomes valuable.[1]

'Data rich' actors, in contrast, are large commercial, academic and government bodies, including security and military agencies, with the resources to exploit the opportunities afforded by big volumes of data, notably to operate costly data infrastructures, especially the development and application of machine learning technologies to deal with large datasets (Andrejevic and Burdon 2015, 21). In short, these actors have the necessary infrastructure, expertise and technologies to analyse large swaths of data, including open data (Andrejevic 2014). Those who can control data are understood to wield 'new kinds of informational power' (West 2019, 22), equivalent to Strange's concept of structural power in the knowledge structure: the ability to set the rules under which others – that is, data-poor actors – operate.

As a quick example, consider the ride-hailing company Uber. Uber's control over its drivers exemplifies the knowledge – and power – asymmetries in the gig economy. The gig economy can be thought of as digital piecemeal work. Lacking the long-term stability, protection and benefits offered by traditional employment, gig workers get paid depending on how many tasks they complete – in this case, taxi rides. Meanwhile, ride-hailing companies use data-driven algorithms to control the working conditions and pay of drivers, often in exploitative unfair ways (Calo and Rosenblat 2017). Uber's algorithms, for example, sometimes conceal from their drivers their fares per trip, thereby pushing drivers into working longer hours for less pay (see also Rosenblat 2018). That drivers don't have access to this data or the algorithm that shapes their working lives marks them as data-poor and solidifies their structural disadvantage when dealing with Uber, their de facto employer.

## Characteristic 7: Claims of Predictive Accuracy Are Overstated

Another foundational faulty assumption underlying the data economy is that human behaviour can be objectively and accurately quantified, understood and predicted through data, an ideology termed 'dataism' (van Dijck 2014), which we will discuss in greater detail in the next chapter. A core claim of dataism is 'veracity through volume' (Crawford et al. 2014, 1667), meaning that mass amounts of data ('big data') are understood to produce valuable expert knowledge.

A core ideology of the information-imperium state, dataism holds that data-intensive processes, including regulation via algorithms, are perceived to be more effective, accurate and efficient than non-big-data human-centred ways of doing things. Even in light of data-driven debacles such as algorithms that unfairly deny people public services to which they are entitled, including housing and child protection (see Eubanks 2018; Hintz et al. 2018), the legitimacy and predictive accuracy that industry accords to algorithms can be 'seductive' for policymakers (Crawford et al. 2014, 1667).[2] Algorithms, in other words, promise straightforward technological fixes to complex social problems. However, not only are these promises faulty because algorithms cannot achieve their designers' lofty goals, but adopting data-driven processes to deliver government programmes can further entrench biased and discriminatory practices.

A dataist mindset typically assumes that data collection is comprehensive and reliable and that the gathered data is accurate and fully represents the phenomenon being examined. Not all information, however, can be translated into data, as aspects of the original phenomenon can become lost or be untranslatable (Loukissas 2019). Dataism also tends to also overlook the reality that datasets can be incomplete. Design anthropologist Sarah Pink and colleagues contend that data can be 'broken', necessitating 'repair and maintenance' work before data analysis can take place, meaning that actors may manipulate and process data in certain ways to make it 'useful' or valuable for certain purposes (Pink et al. 2018, 3).

The assumption that data can speak for itself also ignores a key insight of sociologists of knowledge: because data itself is a human product, it will necessarily never be objective. The concept of broken data, meanwhile, reminds us that instead of assuming data completeness and accuracy, we should be attentive to the ways that data collection, storage and analysis are partial and can be faulty or disrupted, while also recognizing the human labour involved in repairing data to render it valuable.

Algorithms, which are a set of instructions designed to generate a specific desired outcome, are central to efforts to monitor and predict behaviour and events, typically through automated decision-making. Similar to the

economic, political and social power we have accorded data, algorithms are commonly framed as having significant power and legitimacy, offering the 'promise of algorithmic objectivity' (Gillespie 2014, 179). Communications scholar and Microsoft researcher Tarleton Gillespie notes that this objectivity is a 'carefully crafted fiction' intended to portray algorithms' outcomes as 'fair and accurate, and free from subjectivity, error, or attempted influence' (Gillespie 2014, 179).

## Characteristic 8: Individual Consent Legitimizes the Data-Driven Society

The data-driven economy is founded upon the myth of individual informed consent. The idea of voluntary informed consent holds that personally identifiable information should only be collected, stored and used once individual consent is secured, namely with 'the consent being specific, freely given and based on full and adequate information' (Taylor et al. 2017b, 6). Much of the Quayside debate turned on the question of how Sidewalk Labs could (or should) get individual consent for the surveillance throughout the urban landscape that would be necessary to make their plans work.

As we will see in chapter 9, there are two key issues with using individual consent as a regulating principle when it comes to data governance. First, it is problematic to assume that individuals can provide any form of meaningful consent for the collection of their personal data. Individuals are usually deemed to have provided consent through the terms-of-service that pop up whenever one uses software or an online service. In the United States, the dominant perspective of privacy since the late 1990s assumes people act as rational consumers who read (notice) and then give an informed consent (choice) to privacy policies (Cranor 2012, 304). This notice-and-consent approach has been exported globally through US-based internet companies in their terms-of-service agreements that are the legal authority for their data-intensive business models.

Anyone who has ever come across one of these terms-of-service agreements will understand the problem immediately: most people neither read nor understand these often-massive and often-impenetrable documents (see Bakos et al. 2014; Obar and Oeldorf-Hirsch 2020; Tene and Polonetsky 2013). In fact, the standard phrase, 'I agree to these terms and conditions' has been called, without exaggeration, 'the biggest lie on the internet' (Obar and Oeldorf-Hirsch 2020, 130). Even if people painstakingly poured through their terms-of-service agreements, they would need 'ubiquitous omnicompetence' in order to understand how their data may be collected, used and shared, particularly how it may be 'repurposed and sold by every application,

commercial organization, non-commercial organization, government agency, data broker and third-party' (Obar 2015, 4).

Second, it is problematic to view consent solely (or even primarily) as an individual responsibility. The idea of voluntary informed consent is deeply embedded in Anglo-Saxon conceptions of privacy as an individual right (see, e.g., Taylor et al. 2017a). In this understanding of privacy, one individual's disclosure of personal data to an entity does not affect the privacy of another.

This is not always, or even usually, the case. With the rapid growth of social media platforms and the expansion of corporate databases, disclosure of personal data by one individual may result in knowledge of the personal data of others linked to this person. For example, as law enforcement increasingly turns to consumer DNA ancestry sites as an investigative tool, genetic data shared by one person for a specific purpose – to trace a family tree – may be used for other purposes not intended or likely anticipated by the donor. What's more, most individuals' data only has value when combined with others' data, for example, in constructing credit risk standards against which others are judged (and possibly denied access to credit). In those situations, one person's individual consent, even if fully informed, can end up harming other people.

In both cases, it is clear that individual consent–based privacy is too narrow a conceptual lens to use when setting policy. Instead, as we argue in chapter 9, a broader, more collective human rights–based approach to privacy is necessary (see Dencik et al. 2016; Taylor 2017a). Human rights–based approaches, which argue for the importance of protecting individual rights while also establishing or expanding collective rights in the data economy, tend to favour measures that restrict some types of data collection and limit data commodification. We discuss this decommodification approach as an alternative to the information-imperium state in the conclusion.

## AN INCONVENIENT TRUTH: THE FALSE PROMISE OF ANONYMIZATION

Concerns about individual privacy (to say nothing of collective privacy) present probably the most significant roadblock towards the construction of an efficient data-based economy. That ubiquitous surveillance and privacy fit poorly together has not stopped industry, government and policy entrepreneurs from attempting to find privacy workarounds that would allow the data-driven economy to flourish. Technical infrastructures based in part on greater individual control over their data, such as data trusts, which we discuss in chapter 9, are efforts in this vein.

De-identified (or anonymized) data represents a similar attempt to maximize data collection while minimizing privacy risks and concerns. Data

de-identification is a technical process that 'strips' or 'scrubs' personally identifiable information from a dataset, such as names, addresses or birthdates (see, e.g., Lubarsky 2017). By stripping personal identifiers in a robust fashion, the idea is that the data can no longer be traced back to identifiable individuals, and therefore can be broadly used, stored and shared without typically being subject to the same privacy regulations as personally identifiable data.

Data de-identification treats privacy as something that is only relevant to individuals. While many debates on privacy and surveillance in the data economy focus on individuals being tracked, amassing and processing data is often about groups (Taylor et al. 2017a). Aggregate data refers to group-level data that has been created by combining individual-level data, often in anonymized form, for example, to predict trends in energy consumption or health. Governments and companies are interested at the level of the group in terms of forecasting, tracking and influencing behaviour, which is typically undertaken using automated data tools. As such, data de-identification schemes do not do much to address the harms from collecting group-level data.

Putting aside these group-level concerns, data de-identification is often portrayed by companies as a solution to address public or regulator concerns about data security, privacy or the misuse of data, as well as the possible sharing or sale of personal data with third parties. Data de-identification, however, is not a foolproof solution. Over the last decade, a growing body of scholarly research by computer scientists and mathematicians demonstrates that it is increasingly possible to re-identify, or, put it another way, to de-anonymize data (see, e.g., de Montjoye et al. 2013; Narayanan and Shmatikov 2008). Data re-identification is the process of discovering 'the identity of an individual who contributed data that subsequently had anonymization techniques applied' (Curzon et al. 2021, 102). In fact, as research by computer scientist Yves-Alexandre de Montjoye and colleagues shows, it is 'increasingly difficult, if not impossible, to anonymize a dataset' (Montjoye et al. 2012; cited in Kammourieh et al. 2017, 46). Data de-identification advocates, however, argue that sufficiently robust de-identification techniques, combined with proper data-protection practices, minimize the risk of de-identification (see, e.g., Cavoukian and El Emam 2014, 2).

Actors can re-identify data when de-identification practices are flawed or are insufficient to prevent re-identification or when actors combine datasets that were meant to be kept apart (see Lubarsky 2017; Ohm 2010). Combining a small number of attributes extracted from various datasets, such as gender, data of birth, postal code and marital status, is often sufficient to re-identify individuals with a high degree of confidence (see Rocher et al. 2019). What's more, these attributes need not relate to personal data, as re-identification can also be undertaken by combining personal data with non-sensitive,

non-personal data, such as movies watched, locations visited or web browsing histories (Narayanan and Shmatikov 2019). Every data point, even those revealing something seemingly innocuous 'abets further reidentification' (Ohm 2010, 1705). As de-identification attacks are improving over time, computer scientists Arvind Narayanan and Vitaly Shmatikov argue that de-identification techniques 'should rest on provable guarantees rather than the absence of known attacks' (Narayanan and Shmatikov 2019, 1).

A full account of the technical processes of de-identification and risks of re-identification lies outside the scope of this book (but see Lubarsky 2017; Ohm 2010). What's important to our argument, however, is that corporate claims about the effectiveness of de-identification practices reveal a fundamental truth about data: 'Data can be either useful or perfectly anonymous but never both' (Ohm 2010, 1704). While perhaps an overstatement, what this means is that data utility and privacy are 'intrinsically connected' because 'as the utility of data increases, the privacy decreases' (Ohm 2010, 1705–6). There is therefore an incentive for actors to re-identify data, either for their own use or to sell to others.

Here, again, we see a fundamental tension between privacy and the collect-it-all mentality characteristic of our data-driven society. Actors reliant upon pervasive data collection are understandably resistant to the argument that de-identification does not effectively protect privacy as there are strong financial incentives to safeguard the 'simplicity of the de-identification paradigm' (Narayanan and Shmatikov 2019, 2).

Aside from the data-maximalist attitudes of various states and companies, fundamental changes in the data economy have contributed to the risk of data re-identification. The number of datasets, both public and private, has grown, meaning that there is a risk that datasets may be combined (Kammourieh et al. 2017). Here, the risk is that disparate data sources may not individually reveal personally identifiable information but their combination may do so (Curzon et al. 2021, 7). The growing data broker industry, moreover, has as its primary purpose to amass, link and combine datasets from consumers, companies and even governments to uncover potentially valuable patterns in data of use to those interested in forecasting or influencing behaviour. New data sources in the last two decades also provide richer data, such as genetic information, fitness wearables, social media and mobile phone data (Taylor et al. 2017b, 3). Further, technological advances enable the collection and processing of mass amounts of data that, as noted by the British Academy and the Royal Society, 'generate unexpected patterns or insights which go far beyond the original intended purpose of data collection' (The British Academy and the Royal Society 2017, 34; cited in Rinik 2020, 347).

## CONCLUSION

The eight characteristics (and one inconvenient truth) outlined in this chapter describe how data 'works' in our own historically contingent knowledge-driven society. A knowledge-driven society is naturally predisposed to favour ubiquitous surveillance and control over knowledge (in this case, data) because this control is seen as a fundamental element of political, economic and social power. The particular form of this control is, in turn, linked to the interests of the actors involved. For the information-imperium state this means commodified data for market-based actors (i.e., companies), while for states, it entails data that serves state goals of protection/security (as with the system revealed by the Snowden leaks) and the delivery and management of public services. It is the logic to which state and non-state actors – digital economic nationalists and knowledge feudalists, whether authoritarians or democrats – must respond.

Chapters 6 through 9 explore how private actors and governments are increasingly amassing and using data in order to wield economic and political power. In particular, this book studies the accumulation and, importantly, the interpretation of data as a key power vector in the global economy and also considers those who benefit and those disproportionately affected by the rise of a data economy. In short, how can (and how *should*) data be governed, by whom and for what purposes?

None of this should be read to imply that there is either anything natural or inevitable about ubiquitous surveillance or data commodification. One of the lessons we can draw from Karl Polanyi's discussion of fictitious commodities is that the harm caused by treating human beings or nature as commodities can be reduced or eliminated. Policymakers can, in the name of human rights, limit the economic and social pressures of a data-driven society by restricting data commodification – think data-minimization efforts or exempting children from online data-collection efforts – and ubiquitous surveillance.

We would be remiss not to acknowledge that enormous pressures against such efforts are, to an extent, built into the system. These are evidenced most directly by the ongoing charade that terms-of-service agreements are anything but 'the biggest lie on the internet' (Obar and Oeldorf-Hirsch 2020), as well as by the faith that has been placed in de-identified data to square the surveillance-privacy circle. As we will see in the next chapter, the biggest obstacle towards more humane data and IP policies is not material power but ideology. The emergence of a knowledge-driven society has not only reshaped the economy and foundational institutions like private property but also how we think about the world itself.

## NOTES

1. We purposely avoid using the term 'big data' because, as boyd and Crawford (2012, 663) remark, 'big data' involves not just the ability to collect and analyse large datasets but also the 'belief that large datasets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy'. We explore the implications of this belief in chapter 5.

2. See chapter 5 for a more in-depth discussion of dataism and chapter 8 for a discussion of how governments are both using automated data practices to deliver public services and battling data companies to access the data necessary to regulate sectors like housing and transportation.