5

Data as Labor

VALUING INDIVIDUAL CONTRIBUTIONS TO THE DIGITAL ECONOMY

Facebook: Jayla, why is Imani always trolling Deon's posts?

Jayla: I'm a little busy today, Facebook.

Facebook: I know, but rates are double today. If you can give me ten minutes to figure this out, you'll make $15.

"Ok what's up?"

"I'm trying to figure out what's going on between Imani and Deon. They used to only interact rarely, then not at all and now Imani is always mocking what Deon posts."

"Yeah that's what happens when romance goes astray."

"Ah, so they were romantically involved? They didn't post about it."

"Right, not everyone likes to announce their intimate engagements to the world."

"I guess it makes sense now, given some other things I saw them doing . . ."

"You shouldn't be telling me about that!"

"So, who broke up with whom?"

"Can't you tell: Deon dumped Imani! That's why she's always trying to make him out to look like a sissy in her comments. She's getting back at him and wants to make him feel like he's nothing without her."

"I get it. Could you tell that just from the posts or did you know the whole story?"

"Well, they kept it pretty quiet, but I guessed from what was going down online and then I got Imani to go to yoga with me to give me the low-down."

"Do you two often talk about private things at yoga?"

"It's kinda a private girl place, and the physicality of the workout opens you up to talking things out."

"Well, thanks for your help, Jayla. Next time I'll hopefully be able to pick up these dynamics on my own and maybe even help you notice them. In the meantime, anything you need help with today?"

"Given you used up the time I was going to spend finding gifts for my cousins, maybe you can help me fix that."

"You mean for Diwali?"

"How did you know that?"

"Well, it is coming up next week and Malik's wife is Indian, so I thought their kids might be celebrating it."

"Good call. So, the problem is, honestly, I don't know what Diwali is, or what you get for it, or what you would get kids these days."

"I think I have the perfect thing: a virtual reality game for the kids and some artisanal sweets for the whole

family. $25 total, plus $2 for the work I did finding it. Or you can subscribe to my personal assistant services for $100 a year. You've already spent $75 this year and we're not even halfway through."

"You're right, I should subscribe. You can charge my credit card. But I need the presents by tomorrow morning."

"Obviously, you do; I know when you're going to see them. The price includes a delivery by then. You're usually up by 9 and don't have anything scheduled then. Alright if the drone comes by to bring the sweets around then? The game will appear on their Oculus; how long after your arrival do you want that to happen?"

"Yeah that sounds good, and maybe 20 minutes."

"All set, I'll let you get back to your day."

"Thanks for the work and help, sorry I was a little cranky."

"No need to apologize to me. Just get some sleep, you were going hard last night."

"Good idea."

You probably find the idea of Facebook prying into the details of your friends' relationships, and paying you to help it, creepy. Yet this business practice, at one remove, is already ubiquitous. Why does Google enable us to plan our trips on Google maps? It learns traffic patterns, which it can then package into services it sells to ride-sharing and public transit platforms. Why does Facebook provide us a "free" space to build our social lives? Because we reveal personal information, which enables Facebook to match us with products we might be willing to buy. Why do Instagram and YouTube offer such useful ways to share media? The images and video they host are the inputs to

"machine learning" (ML) systems that power "artificial intelligence" (AI) services that they sell to customers—from face recognition to automated video editing. If you aren't aware of how much platforms know about you and profit from this knowledge, check out the account settings pages they increasingly are required to have, which display this full set of information; you may be surprised.

The primary difference between the scenario we describe above and present practice, other than some advances in chat capacities, is that in the world we imagine, Facebook is open and honest about how it uses data and pays for the value it receives with money. The user's role as a vital cog in the information economy—as *data producer and seller*—is highlighted.

Why is this important? Most people do not realize the extent to which their labor—as data producers—powers the digital economy. Consider how people think of AI. In some portraits, AIs are autonomous agents built by brilliant and possibly mad programmers like the reclusive genius in the 2014 film *Ex Machina*, who set into motion a system that runs itself. Reality is different, however, as "the inventor of virtual reality" Jaron Lanier highlights in his brilliant 2013 book *Who Owns the Future?*,[1] which inspired many of our ideas in this chapter.[2]

AIs run on ML systems that analyze piles of human-produced data. "Programmers" do not write ingeniously self-determining algorithms. Instead, they design the interaction between workers (meaning us, the users who produce data) and machines (computational power) to produce specific information or production services. Most of the difficult work is not deriving profound algorithmic designs. Instead, it involves tweaking existing models to fit the relevant data and deliver the desired service. Programmers of ML systems are like mod-

ern factory floor managers, directing data workers to their most productive outlets.

The powerhouses of the digital economy, firms like Facebook, Google, and Microsoft, exploit the lack of public understanding of AI and ML to collect for free the data we all leave behind in our online interactions. This is the source of the record profits that make them the most valuable companies in the world. Facebook, for example, pays out only about 1% of its value each year to workers (programmers) because it gets the rest of its work for free from us! In contrast, Walmart pays out 40% of its value in wages.[3] People's role as data producers is not fairly used or properly compensated. This means the digital economy is far behind where it should be, that the income from it is distributed to a small number of wealthy savants rather than to the masses, and that many of us have a false fear of AI creating mass unemployment when humans are more necessary than ever to our digital economy.

The Rise of "Data Work"

Data work, like "women's work" and the cultural contributions of African Americans at one time, has been taken for granted. In the case of women, the extensive labor required to raise children and manage the home was treated as "private" behavior, motivated by altruism, that was outside the economy and hence not entitled to financial compensation or legal protections.[4]

In the case of African Americans, many of the defining concepts of modern American music and dance originated in the private entertainment practices of African American communities. As depicted in films like *Show Boat*, this creativity was often exploited by white entrepreneurs for profit. At the

same time, African Americans were often not paid at all, as their contributions were dismissed as idle amusements.[5] Even when they managed to receive some compensation for performances, their intellectual property rights were usually disregarded, partly because they were excluded from the American Federation of Musicians, which was central to securing artists' rights, until the 1970s. The story of data work is less familiar than these iconic historical cases, but increasingly important.

Early in the life of what is now the Internet, its designers had to choose what information to record and what to discard. Many early designs supported technologies that would have made it easier for receivers of information to automatically pay the providers. These designs used two-way links where every piece of information would effectively carry its full provenance with it.[6] At various points in the development of the web, governments and companies made attempts to direct revenue to the diffused set of individuals who contributed value to the system. In France, the pre-Internet Minitel system had a system of micropayments[7], for example, and the America OnLine (AOL) service popular in the 1990s in the United States charged its customers a fee and used the revenue to pay for content it made available within its simplified "walled garden" interface. For a period, some Internet designers were trying to force email to carry postage stamps as a way to deter spammers from flooding inboxes with junk.

Yet, what eventually became the mainstream Internet did not start as a commercial or economic project. Instead, it was a collaborative platform within government, military, and academic circles where participants were assumed to be interested in collaboration for reasons external to commercial motivations. The World Wide Web interface of hyperlinks developed by Tim Berners-Lee and others therefore placed

emphasis on lowering barriers to participation rather than on providing incentives and rewards for labor. "Information wants to be free" became a slogan for entrepreneurs and a rallying cry for activists. It especially appealed to a Silicon Valley mentality that grew from the counterculture of the 1960s.[8]

During the 1990s, venture capital poured in to commercialize the booming Internet before online services had established how they would monetize their offerings. Internet companies relentlessly pursued users under the banner "usage, revenues later" (a "backronym" for "url"). While partly driven by the dot-com stock market bubble, this strategy was also influenced by the dominant position Microsoft had established by offering its operating system at relatively low cost and in a form compatible with many hardware platforms. The "network effects" created by this strategy were widely viewed as placing Microsoft in a position to reap enormous rewards.[9] This encouraged many venture capitalists to fund services that rapidly enlarged their user base even if their business model was unclear.

As the bursting of the tech bubble cooled this euphoria, emerging tech giants like Google had to find a way to make money from their user base. Google's Sergey Brin and Larry Page initially considered user fees and paid subscriptions, while insisting they would never turn to advertising. But several factors forced them to change their minds.[10]

First, the extended period of free access to services in the late 1990s caused users to become accustomed to an Internet where payment for pure information services was infrequent. People developed a strong attachment to the idea of completely free services, an attachment that likely made this tradition hard to break later.[11] In fact, a social and business movement developed around the concept that online services

should be free, as embodied in entrepreneur and writer Chris Anderson's 2009 best-selling book, *Free: The Future of a Radical Price*.[12]

Second, many of the services provided online were, at least initially, occasional and small, with the result that investment in the development of infrastructure that would have been needed to keep track of payments was not cost-justified. In the late 1990s and early 2000s, many start-ups tried to create systems of micropayments. For example, usability guru Jakob Nielsen led a campaign for micropayments.[13] One of these efforts eventually became the payment platform PayPal. However, in practice (at least in its early years) the overhead costs of PayPal meant it was used only for large transactions. The emergence of social networking and blogging services of "Web 2.0," where many interactions are quick and superficial, made this problem worse. Required payments would have been too small to justify the costs on platforms like PayPal.

Third, in early days the Internet was an unfamiliar Wild West populated by many sophisticated young hackers who were willing to put up with inconvenience in exchange for "freedom." In this environment, dubiously legal services, such as Napster, thrived and could muscle out more secure legal services because mainstream alternatives struggled to keep up with technology. This made charging for anything, even established forms of intellectual property such as music, challenging.

Together, these forces established an environment where users were reluctant to pay for anything and the providers of services therefore searched for alternative means of staying afloat. Desperate for some way to monetize their massive user base, Google turned to advertising to stabilize its balance sheet. Facebook, YouTube, and others followed Google's lead.

Google's insight was that advertising online could be targeted more finely to user needs than is possible in traditional advertising media, like print newspapers or television. Because Google can glean the values and preferences of users from their search history, it can minimize advertising waste and noise. The personal ecosystem offered by Facebook, far more complex than a Google search, serves a similar function. Facebook learns details about users, which allows it to match them to advertisers who seek a narrowly targeted audience, and to place advertisements in social contexts by encouraging users to share advertising campaigns with their friends. Most important, it allows Facebook to identify the most opportune moments to hit users with a "reminder" to purchase something they had previously been considering, a feature that sometimes gives users the eerie sense that the service can read their minds.

Factories for "Thinking" Machines

The insight that *data* about users was the central asset for technology giants became increasingly salient with the explosion of interest in "big data," ML, and AI. Machine learning is a "second-generation" approach to building AI systems. The first generation, which largely died out during the 1980s, focused on building formal logical rules that represented intellectual human tasks like language or game playing. This approach had some notable successes, including the Deep Blue computer, which defeated World Chess Champion Gary Kasparov. But it failed in most commercial applications. During the 1990s and early 2000s, a new approach based on statistics and probabilistic prediction came to the forefront.

The core idea of ML is that the world and the human minds that intelligently navigate it are more complicated and uncertain than any programmer can precisely formulate in a set of rules. Instead of attempting to characterize intelligence through a set of instructions that the computer will directly execute, ML devises algorithms that train often complicated and opaque statistical models to "learn" to classify or predict outcome of interest, such as how creditworthy a borrower is or whether a photo contains a cat.

The most famous example of an ML algorithm is a "neural network," or neural net for short. Neural nets imitate the structure of the human brain rather than a standard statistical analysis. In the usual methods of statistics, different input variables are assumed to have relatively simple and independent effects on the "output" variables we want to explain. Being tall, being a man, and eating a sugar-rich diet are all assumed to be predictors of a high body weight in a relatively independent manner.

Neural networks work differently. Rather than inputs directly and independently determining outputs, the inputs are assumed to combine in complex ways to create "features" of the phenomenon being studied, which in turn determine other features, which eventually determine the outcome. Such complex relationships are familiar from everyday life. If we see a number of red pixels in our eyes, we may realize the image is predominantly red. If we see a trunk and floppy ears, we may recognize an elephant. Only once we have perceived both of these shapes, however, do we realize we are looking at a representation of the Republican party, commonly denoted by the color red and the shape of an elephant. A number of red pixels on floppy ears would not directly suggest "Republican"; it would be as likely to convey a wound, for example.

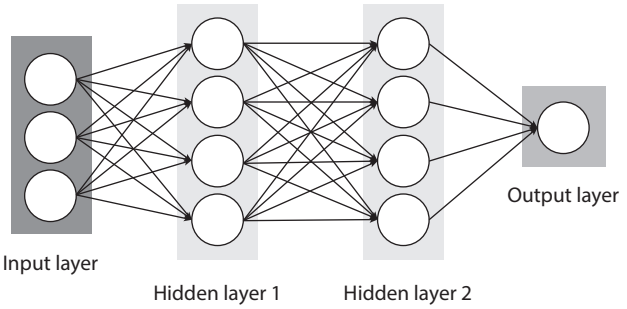

FIGURE 5.1: A stylized representation of a neural net.

A neural network is able to handle such sophisticated abstractions by learning the presence of more abstract features of data in its "hidden layers." Immediately apparent facts about an image, such as the shade of color of each pixel in an image, are represented by the activation of "neurons" or nodes in an "input layer." This input layer of neurons is then connected to a "hidden layer" meant to represent somewhat more abstract features. Neurons in this hidden layer will in turn activate when some weighted average of the inputs to that neuron surpass some "activation threshold." These activations tend to represent slightly more abstract and complex features of the image.

To achieve greater abstraction, this hidden layer is then connected to a second hidden layer, with the same properties, and so on. Eventually the last of these hidden layers yields to a final "output layer" that determines the eventual outcome of interest, such as a prediction of whether the photo is Republican campaign material. Figure 5.1 shows an example of a simple neural net with only two hidden layers.

Neural nets can, in principle, encode a very wide range of relationships, especially when the number of layers is large. Typically, each layer will encode a higher level of abstraction

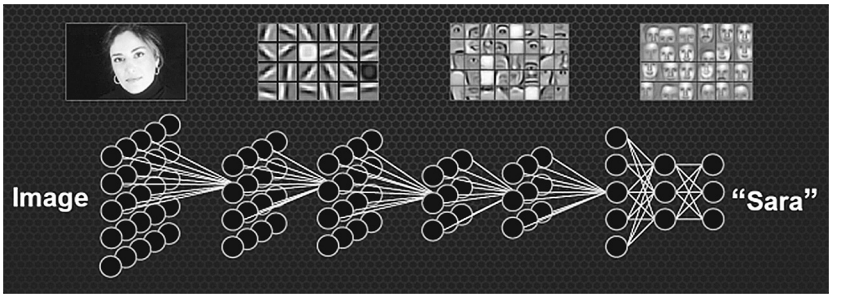

FIGURE 5.2: A facial recognition neural net. Deeper layers represent higher degrees of abstraction.

than the layer below it. Figure 5.2 represents an example. "Shallow" layers, near the input image on the left, represent relatively simple features of the image. On the far left we see a typical image input. Next, to its right, we see a shallow hidden layer. A typical set of patterns that leads to activation of this neuron is shown. This layer tends to detect lines and colors oriented in various directions, a relatively simple and concrete idea. A deeper layer, shown to its right, encodes elements of a typical face, such as eyes, ears, noses, etc. On the far right we see one of the deepest layers, closest to output. These show abstract versions of entire faces. Once a neural network reaches this level of abstraction, it is clear how it can detect faces: the firing of one or more of these deepest "facial recognition" neurons indicates that a face is present in the picture. Neural networks thus achieve astonishing intelligence through repeatedly reprocessing increasingly complex inputs into more complex ones through a series of layers, until they finally reach their desired prediction.

How does a neural network learn, from the endless possible combinations of weights at each layer, which ones are right to predict the outcome of interest (the presence of a face in this

case)? There are three critical components that go into making a working neural net. First, "data," usually an extremely large collection of labeled examples; in this case, this would be a large number of photos tagged as containing or not containing a face. Second, "computation." Neural nets are usually run on large farms of servers. Last (and, as we will argue, least), "supervisors," the programmers who set up the structure of the net, help prevent it from getting stuck, and use various tricks of the trade to ensure it learns quickly and effectively.

Neural nets are nothing new. Researchers have been interested in them on and off since at least since the late 1950s. However, until about a decade ago, neural nets were widely viewed as useless: in 1995 one of the founders of ML, Vladimir Vapnik, bet an extravagant dinner that by 2005 "no one in his right mind will use neural nets."[14] The problem was that "shallow" neural nets, those with few layers, could not accomplish much. Most interesting properties of objects are much more abstract than these simple, shallow nets could detect. On the other hand, attempts to train deeper nets failed for years because of the lack of data and computational power.

Without sufficient numbers of labeled examples, the space of possible representations was simply too large for the neural net to search through. It would thus end up "overfitting" to irrelevant details of particular images, such as the fact that all images containing a face might have exactly three red pixels in the picture. The problem of overfitting—that is, of trying to fit a complex model to insufficient data—is nicely illustrated by the xkcd cartoon partially reproduced in figure 5.3. If we allow a complex set of rules to predict presidential elections, there are too few examples to fit these complex rules and thus our rules can easily "overfit" to inessential features of the elections, resulting in bad predictions. The more complex the rules we

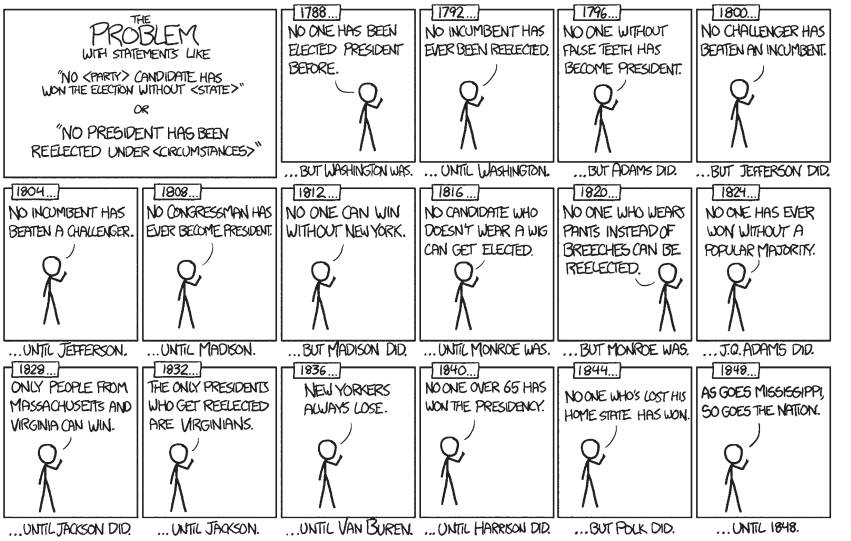

FIGURE 5.3: The problem of overfitting, illustrated by predicting presidential elections. *Source:* Excerpted from "Electoral Precedent" at https://xkcd.com/1122/.

want to fit (the deeper and more fully connected the neural net), the more data we need to avoid overfitting. Computer scientists and statisticians call the number of labeled data points needed to avoid overfitting for a problem (such as recognizing faces, or artistic styles) the "sample complexity" of the problem.[15]

Data alone, however, are insufficient to train a neural net. These data have to be stored and processed. More important, the process of actually training the net requires huge numbers of computations. Without ample computers capable of performing all these calculations, neural nets never find the right explanation of the observed data, no matter how much of it there is. The dramatic advances in computational and storage capacity on the cloud in the late 2000s was critical to allowing neural nets to be trained. The deeper and more complicated a

net, the greater computation and storage required to train it. The computation and storage requirements of a net are called its "computational complexity."

The last component of making a neural net function is programming. Programmers currently play important roles in tweaking the structure of the net and the procedure by which it is trained. However, these processes are being automated through a movement called "democratizing AI" led by Microsoft.[16] The number of programmers required, unlike the amount of data and computation, does not inevitably grow with the complexity of nets. More basic research, proposing new algorithms, or training techniques, can have a greater impact, but in practice the advantages granted by such algorithmic advances are usually short-lived and quickly replicated. The crucial components of success for nets are data and computational power.

While simple, shallow nets, which can solve basic problems such as detecting whether a picture is oriented horizontally or vertically, have low complexity (both sample and computational), more complicated, deep nets, which can solve more sophisticated problems like personalized facial recognition or providing blurbs describing the action in a photo, are much more complex in terms of both the data and computation they require.

This is why neural nets were hardly used prior to the late 2000s and then, beginning around 2010, exploded to become perhaps the hottest technology of the day. It was around that time that both the volume of data collected and the speed and depth of computation became sufficient to allow applications that made a difference in users' lives. Around that time the first ML-powered personal digital assistants and dictation services emerged; Siri, Google Assistant, and Cortana became familiar

features of everyday life. Even more ambitious applications are being developed, including virtual and augmented reality, self-driving cars, and drones that deliver goods to consumers at the click of a button.

Because these services have high "sample complexity," they require vast stores of data on which to train the ML systems. Thus, the vast data sets collected by Google, Facebook, and others as a by-product of their core business functions became a crucial source of revenue and competitive advantage. Companies that started as reluctantly free service providers in search of a revenue model and morphed into advertising platforms are now in the process of becoming data collectors, delivering services that lure users into providing information on which they train AIs using ML.

Sirens and Titans

Jaron Lanier describes such platforms as "siren servers." Their allure, he explains, derives from the combination of the free services they offer because of their scale and exceptional data access. Yet Lanier worries about the social and economic consequences of their business model. Because they do not pay their users for data, they do not give their users proper incentives to supply data that are most needed.

For example, right now Facebook receives a constant flow of hundreds of millions of new photos posted each day by users. These photos are good training grounds for ML systems that Facebook is developing to automatically label and even explain photos. Yet at present, there is a mismatch between Facebook's needs and the reasons that users post photos. Users often provide little information accompanying a photo because they expect their friends to understand the context of it.

The result is that the data that Facebook receives are low-quality. Facebook tries to nudge users to provide useful labels by inducing them to write comments explaining photos or by associating emotions with them. But what Facebook really needs is the capacity to ask users simple questions about the photos and receive answers from them.

Lacking direct input of this kind, Facebook sometimes employs "crowd workers" to label the images after the fact. But these workers will rarely understand a photo as well as the person who posted it does. If, instead of hiding their ML algorithm's use of data from users, Facebook were to make users aware of the role they played, and to reward them for inconvenient but valuable contributions, ML systems would have better data to work with. This alternative world, sketched in our opening vignette, would allow them to supply better AI services to their customers and clients.

Another example is YouTube, to which 300 hours of video are uploaded every minute, according to the website. Yet the producers of this content receive minimal compensation. While the analytics are a bit complicated, a typical YouTube content creator receives roughly $2 for 1,000 views of a video. Given that an average YouTube video lasts about 4 minutes, this means that creators can expect about five hundredths of a cent per minute their videos are viewed. In contrast, Netflix pockets about half a cent per minute a typical user watches its videos, or roughly ten times as much.[17] It is not a great surprise, therefore, that Netflix has produced critically acclaimed television series like *Orange Is the New Black* and *House of Cards*, while YouTube videos are less celebrated for their cultural value. Similar calculations apply to the contrast between traditional news outlets and Twitter. These prices are all likely a small fraction of the value users derive from watching. People's time

is worth more than a few percent of a cent. This phenomenon is broader than video, however; the siren servers have thrived on devaluing creative content from news to music and appropriating the value it generates for themselves rather than creators.[18]

Lanier also worries about the distributional and social consequences of the failure to pay for data and online creative production. There is widespread concern that AI systems will displace many human workers. A widely discussed engineering study found that nearly half of all jobs in the United States are likely to be automated in coming decades.[19] While skepticism is warranted, even the possibility of massive long-term job loss justifies thought about how to limit the negative distributive and social consequences. Experience with automation suggests that communities where "robots take the jobs" are usually hard hit, not just in terms of income, but also with regard to the sense of purpose and will to live of community members.[20]

Job turnover and displacement have always been unfortunate consequences of technological progress. New types of jobs regularly replace old ones: artisans were replaced by factory hands, human computers by electronic ones, buggy whips by taxi drivers. In each generation, new techniques for producing existing goods offered new kinds of jobs and new goods appeared, which required workers. What strikes many as uniquely worrisome about AI, from this perspective, is that it seems not just to make humans more productive. It holds out the possibility of entirely replacing humans in a wide range of tasks while offering no alternative role for human work.

Nor do these fears seem unwarranted by the economic data. According to one of our ongoing projects with collabora-

tors including Lanier, the share of income going to labor in the largest tech companies is roughly 5–15%, lower than any industry other than extractive ones such as oil, and dramatically lower than service-sector companies like Walmart, where labor's share is roughly 80%.[21] Labor economists have argued that the rise of powerful companies with large monopsony power has been driving down labor's share of income.[22] Their data are too aggregated by confidentiality restrictions to determine the exact sectoral nature of these changes, but it seems plausible that the high-technology industry plays a major role in it. *If* these AI-driven companies represent the future of broader parts of the economy without something basic changing in their business model, we may be headed for a world where labor's share falls dramatically from its current roughly 70% to something closer to 20–30%.

That is a big "if." Forecasting the course of technology is notoriously difficult. Lanier's insight, however, is that even if this does come to pass, AIs are not actually the free-standing replacement for human labor they appear to be. They are trained with and learn from human data. Thus AI, just as much as fields or factories, offers a critical role for ordinary human labor—as suppliers of data, or what we will call *data as labor*. Failing to recognize data as labor could thus create what Lanier calls "fake unemployment," where jobs dry up not because humans are not useful but because the valuable inputs they supply are treated as byproducts of entertainment rather than as socially valued work. Even if AI never lives up to its hype, data as labor may offer important supplemental earning opportunities and sense of social contribution to citizens affected by rising inequality. Yet none of this will happen unless people change their attitudes toward data.

Diamonds in the Rough

Lanier's view might strike some readers as pessimistic. In the existing system, people disclose huge amounts of data about themselves in return for the services the Internet provides—searching, mapping, digital assistance, and so on. Why is it important for people to be paid for data in money rather than in-kind in the form of valuable services?

The leading advocate of this view is Hal Varian, chief economist at Google, who has argued that data are omnipresent these days and that what is scarce are the talent and computational power needed to make sense of these data. Varian thinks that all that is needed for AI services to succeed is for nothing to stand in the way of "natural" collection of data by siren servers, and ample rewards to talented engineers and perceptive investors for their contribution to the mechanics and infrastructure. In this view, data are much more like capital than labor: they are a naturally available resource, harvested from the public domain (where they are freely available), and transformed into something useful only by the hard work of programmers, entrepreneurs, and venture capitalists who then deserve to own the data.[23]

Another way to think about this view is in relation to Adam Smith's classic "diamond-water" paradox. Smith found it paradoxical that water was so valuable in use and yet had little value in exchange, while diamonds have such limited uses and yet have great value in exchange. This diamond-water paradox was finally resolved by the "marginal revolution" of the late nineteenth century in which William Stanley Jevons, Léon Walras, and Carl Menger (the first two of whom you may recall from chapter 1) argued that the exchange value of a good is determined by the *marginal* value of the last unit of a good available,

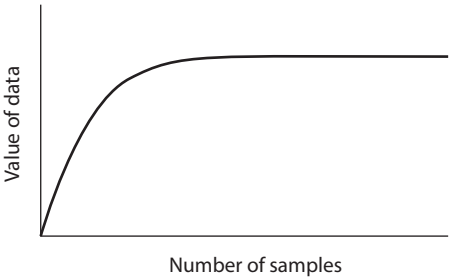

FIGURE 5.4: The value of data as a function of the number of observations in a standard statistical estimation problem. The marginal value declines rapidly. Thanks to Nicole Immorlica for providing this graph.

rather than the *average* value gained by its consumption. While the average value of water is high, its marginal value is low as it is so plentiful. Varian's argument is that while data may have enormous value in total or on average, *on the margin* no individual's data are worth much.

Varian's argument is persuasive if we focus on traditional uses of data in classical, pre-ML statistics. In standard statistics, the goal is to measure some parameter of interest; the simplest example would be the average of something (say, income) in a population. Under common assumptions, the marginal value of an additional individual's income in allowing you to measure the average income in the population diminishes rapidly, because the more you see, the less uncertainty you have about this average. The marginal reduction in uncertainty dies off as the 1.5 power of the number of individuals; this mathematical relationship is depicted in figure 5.4.

For example, if the marginal reduction in uncertainty from one more individual's data when there are only one hundred individuals observed is one unit, by the time we observe a mil-

lion individuals, the value is a mere one one-millionth. Moreover, it is rarely useful to know a quantity extremely precisely. Most of the time knowing it roughly serves our purposes. An entrepreneur who wants to open a wealth management firm in a neighborhood wants to know whether the average income is $100,000 or $200,000, but doesn't need to know that it is $201,000 rather than $200,000. Initially collected data not only reduce uncertainty by more: those initial reductions (from huge uncertainty to reasonably bounded guesses) are more valuable than are later refinements. Thus, in a standard statistical world, data rapidly lose their value. For standard statistics, "big data" are mostly useless. Small data suffice.

The world of ML is different from the world of standard statistics for two reasons that mirror the reasons why data have so little value in the classical statistics perspective. First, the difference in approach between ML and standard statistics is how they relate to complexity. Recall that different problems of different complexity require different amounts of data. In statistics, the goal is to solve a single, simple problem. In ML, as data grow we try to teach the AI system new and more complicated things, to solve problems with increasing sample complexity.

For any one, well-defined learning task, data only tend to have marginal value for a limited range of data sizes, those close to the sample complexity of the problem. When the available data are much below the sample complexity, there are not enough data to even get started on learning. Above this size, most learning has already taken place, so additional data quickly run into the diminishing returns we highlighted above.

This pattern of data values is pictured in figure 5.5. Each vertical line represents the sample complexity of some problem in machine vision; more complex problems lie to the right.

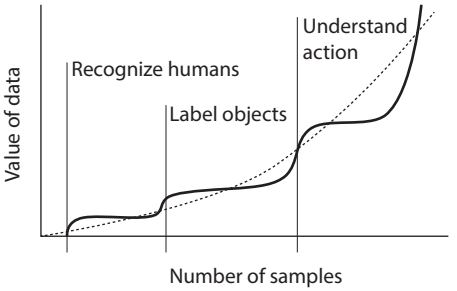

FIGURE 5.5: The value of data as a function of the number of observations in a typical ML domain, here machine vision. Each vertical line represents the sample complexity of a particular problem. Thanks to Nicole Immorlica for providing the graph.

Notice that, after the sharp rise around the sample complexity point, the shape of the curve, at least for a while, closely matches what we saw in classical statistics. Once we have reached enough data to make progress on a particular ML task (such as recognizing whether there is a human in a photo), this problem becomes like a classical statistics problem and additional data lose their value at a similar rate. Until we reach this point, data go through a long period of being useless for the opposite reason and then being incredibly useful over a very short range where the data teach the system what it needs to know.

However, while this pattern holds for any *given* task the ML system wants to learn, the overall learning of the system is quite different, as the figure illustrates. While at any given time the system is only in the data range of learning one or a few things, at any given time it is most likely learning *something*. In the figure, a vision system that is a third of the way along the chart has already mastered recognizing the presence of a

human, and additional labeled photographs are of little value. It also is not yet close to having enough data to understand the nature of the action in the photograph; this is much too complex a problem. However, between these two complexities it is learning to label all discrete objects in the photograph. Thus, additional data are now useless for both the recognition and analysis problems, but very useful for the labeling problem. From this perspective, the primary determinant of the marginal value is not the statistics of a given ML problem, but rather the *distribution of complexity across different problems*.

Just as with classical statistics, there is a second critical question that determines the marginal value of data: how important it is to solve each of the problems data allow ML to tackle. If simple, early problems have much greater value than later, more complex ones, data will have diminishing value. However, if later, harder problems are more valuable than earlier, easier ones, then data's marginal value may increase as more data become available. A classic example of this is speech recognition. Early ML systems for speech recognition achieved gains in accuracy more quickly than did later systems. However, a speech recognition system with all but very high accuracy is mostly useless, as it takes so much time for the user to correct the errors it makes. This means that the last few percentage points of accuracy may make a bigger difference for the value of a system than the first 90% does. The marginal value grows to the extent that it allows this critical last gap to be filled.

To understand these dynamics, consider the oft-abused analogy to human learning. The analogy we are drawing is between the learning processes; we *do not* mean to say that AIs really are like humans. For learning any given skill, studying is mostly useless, then very useful, and then mostly useless again.

For example, until you are advanced enough to grasp calculus, studying math will do little or nothing to advance your understanding of calculus; it will seem impossibly complex. And once you know calculus passably well, additional study will quickly become wasted and redundant. Yet for a critical period, the study is extremely valuable in learning calculus.

At most points in a mathematics education, you will be mastering some more or less useful skill (multiplication, trigonometry, calculus, probability, etc.) and study will be valuable at acquiring that skill, but of little immediate use for others. Whether the marginal returns to studying math overall increase or diminish as you learn more depends on whether the more complex skills have greater or lesser value than the simpler ones. This depends on many factors, and the relationship may not always have a clear direction: multiplication may be more useful than geometry, but less useful than calculus, which you learn even later. But overall evidence on the labor market returns to education suggests that the value of additional years of schooling does not trail off very quickly: advanced degrees often boost earning power by more over what someone with a basic education earns than a basic education does over none.[24]

We suspect something similar is true of ML. While additional data may not improve some services that have matured (like selecting movies you like), the same data may improve other services that are at an early stage (virtual reality, speech translation). In many cases the more complex and sophisticated services are more valuable. This is shown in figure 5.5, where the value gained by later services is greater than the value gained from earlier services. If this is true, then data may actually have *increasing* rather than diminishing returns, as more data allow for the solution of more complicated and

more valuable problems. Furthermore, since human culture is always developing in rich new ways, AI/ML will always need more data to keep up. Even if AIs do eventually "learn everything" and data run into diminishing returns, that day will arrive only in the distant future, once we have AI systems capable of mimicking not only an individual but all collective human intelligence.

Technofeudalism

Why, then, do siren servers not voluntarily pay their users to supply the high-quality data that would allow them to develop the best services? If data production is labor, why doesn't a market for data work emerge as a part of the broader labor market?

In fact, we have seen tentative first signs of markets for high-quality, labeled data. Many researchers and some companies use Amazon's Mechanical Turk (mTurk) marketplace to pay online workers to label and clean data sets, and to participate in social-science experiments. This is not entirely new. Television ratings are still determined by Nielsen, which pays households a small fee to record their viewing.

Notice, however, that the buyers of data in these settings are for the most part *not* the siren servers we have been discussing. Instead, they are smaller companies, academic researchers, and financial firms with no direct access to data. Many of these businesses have exciting prospects. Work Fusion, for example, offers a sophisticated incentive scheme to workers to help train AIs to automate business processes. Might Ai hires workers to label maps and road images and sells the labeled data to companies producing self-driving cars.

However, the total size of these markets is tiny compared to the number of users who produce data used by the siren servers. The number of workers on mTurk is in the tens of thousands, compared to billions of users of services offered by Google and Facebook.[25] The data titans (Google, Facebook, Microsoft, etc.) do not pay for most of their data. The most important players, those who have the scale of data necessary to tackle the most complex problems, are mostly absent from these markets, instead relying on "free" data passively collected from their user base. Of course, these data are not really free; the siren servers provide services to users in exchange for receiving their data.

This arrangement, in which users have the right to prescribed services and the company gains all the upside of the data they generate, may sound novel, but it is actually very old. Prior to the rise of capitalism, feudal labor arrangements worked similarly. Lords insulated their serfs from fluctuations in markets and guaranteed them safety and traditional rights to use the land and to keep enough of their crop to survive. In exchange, lords took all the upside of the market return on serfs' agricultural output. Similarly, today, siren servers provide useful and enjoyable information services, while taking the market value of the data we produce in exchange. We thus refer to this contemporary system as "technofeudalism."

This arrangement is far from optimal. Users who have exceptional skills or knowledge, but who are not particularly enthusiastic about using social media, stay away and deny the value of their contributions to online social life and ML systems. So, too, do people who are poor or otherwise marginalized. Conversely, the lack of payments in the digital world makes it impossible for anyone to specialize in adding value

through their data: one cannot live on the free services that Facebook and Google offer. Technofeudalism also stunts personal development, just as feudalism stunted the acquisition of education or investment in improving land. The inability to earn real money in these environments undercuts the possibility of developing skills or careers around digital contributions, as technoserfs know any investment they make will be expropriated by the platforms. At best, by becoming an exceptionally active member of a digital community one can earn some kudos, badges, and recognition that one can hope to parlay into some vaguely related work offline. At worst, regardless of how much you contribute, you still receive the same digital services as anyone else.

This lack of effective incentives forces siren servers to set up their services so it is simple and convenient for the users to supply this data. Any inconvenient data labeling, or the supply of data from people not inclined to use the services provided by the siren servers, is impossible in a pure feudal system. While interaction environments can be designed to prompt users for useful information (e.g., by making available emoticons that allow users to label their interactions with their corresponding emotions), there are limits to the detail and usefulness of the tags that users will supply purely for fun in the course of entertainment and consumption.

This fact does not escape the siren servers. Most have their own crowd-sourcing platforms, which label huge sets of data they collect through other means to improve the value, reliability, and usefulness of these data. Siren servers go to extraordinary lengths to hide the role of human data work in producing their "magical" services, to the point where efforts to expose this work have become something of a social movement among Internet workers' activists,[26] as described by an-

thropologist Mary Gray and computer scientist Sid Suri in their upcoming book *Demanding Work*.[27] For example, Google quietly subcontracts more than 10,000 human raters to give feedback on the quality of its search results in cases where organic user feedback is insufficient, yet it took investigative reporting to uncover this practice.[28] Thus, while siren servers clearly need help from ordinary users, they wastefully contort themselves to go around the most natural channel (asking those organically interacting with their services for feedback) and make minimal payments to workers outside of this chain, hiding this practice and its importance from the public eye. At the same time, these companies have come to occupy a position of commanding influence over media and policy discussions because of their role in curating information and funding policy research.[29]

Digital Whitewash

In ongoing work with Lanier and other collaborators, one of us is trying to explain why siren servers have tolerated this wasteful state of affairs. A useful analogy is a story from Mark Twain's *Tom Sawyer* in which Tom tries to unload his responsibility for whitewashing a fence onto his friends. His first approach is paying them, but it fails. He soon realizes that if he pretends to be enjoying the task, they will not only agree to perform the work for him but pay him for the privilege. As extensive literature in psychology has shown that, in the right social context, labor becomes leisure; work becomes entertainment.[30]

Siren servers followed in Tom's footsteps. They began collecting user data in the normal course of business, only to find that users were happily laying golden eggs for them to entertain themselves. Users of social networks provide precious

labeled photographs for free to connect with their friends. Google powers its ML analysis of videos from funny YouTube posts. Very few users are paid much for their contributions, allowing the siren servers, who can sell advertising and, increasingly, AI services, to pull in large profits.

Siren servers, especially the leaders in data collection (Facebook and Google), are unlikely to begin paying for data to improve its quality or volume of their own accord. The basic problem is that there are only a few siren servers to compete for user data. Each one knows that if it starts paying for some data, competition among the services will quickly force them to pay for all the data they are currently receiving for free. Paying users, even in a relatively limited set of valuable contexts, is likely to undermine the siren server business model of exploiting free data for several reasons.

First and most basically, the market power (what economists call *monopsony* or *oligopsony* power) of siren servers means that any change to the market which causes users to be paid for their data will increase the siren servers' costs.

The importance of monopsony power in markets for data labor was first highlighted in a paper by Gray and Suri along with economist Sara Kingsley.[31] Since that time, empirical analysis by Suri and his collaborators has confirmed that task posters in mTurk have a remarkable degree of monopsony power, even if they are not very large players in the market, given the time and task-type specificity of "turkers'" interest in completing jobs.[32]

The monopsony power of the siren servers is dramatically larger. They constitute a far larger fraction of all potential available work of this form. While it is difficult to quantify, it seems very likely that a majority of all valuable online and perhaps all

digital data are collected by Facebook and Google; in 2015, Google's share of Internet searches (with which most browsing begins) was 64% and on average Facebook's 1.5 billion users spent 50 minutes every day on the site or app.[33] The sheer fraction of the market controlled by these giants means that they would bear most of the brunt of increased prices for what are currently free data.

Given that most of the productive work to be done is not separate "crowdsourcing" that workers explicitly seek out, but rather work in the course of entertaining online interactions, competing companies would have to build up services of comparable quality and user devotion before they would be able to make productive use of querying users for valuable data. Several start-ups have adopted this model in an effort to attract users to an alternative social network (e.g., empowr) or data management service (e.g., Datacoup). However, they have attracted only a few users with an ideological attachment to the idea. Most users prefer a network that is used by most of their friends and that offers higher quality services.

One start-up that has succeeded in eliciting more useful data from users is reCAPTCHA, familiar to most Internet users as the puzzles one is often asked to solve to prove one is not a bot in order to access an online service. While the CAPTCHAs that reCAPTCHA asks users to solve serve a security purpose, they were designed as a data source for digitizing text and increasingly as a data source for training automated text recognition and other ML-based systems. Note, however, that reCAPTCHA was successful precisely because it partnered with existing siren servers, was incorporated into their product offerings, and never offered monetary payment. After Google acquired reCAPTCHA for a reported $30 million in

2009, a Massachusetts user unsuccessfully sued Google for violating labor laws based on the theory that reCAPTCHA is unpaid labor.[34]

Most potential data labor market competitors for the siren servers would find it hard to make use of data in anywhere near as productive a way as the siren servers can. As we highlighted above, the highest-end AI services only become possible with truly massive computational and data capacities. These capacities are only within the grasp of a few very large digital titans. Of course, a start-up could gather data in the hopes of selling it to the siren servers, but they would have just as strong an interest in avoiding paying for data through the back door as they would through any other route. In short, the siren servers have occupied the central piece of real estate in a "digital commons" that has room for only a few players, and their interests are now opposed to paying technoserfs who are at present voluntarily tilling this land.

Beyond the market structure and the nature of AI technology, the nature of social media makes these sites particularly resistant to competition. Most users want to be part of a social network that includes all of their friends. These *network effects* can make it difficult for competitors to enter the market unless they have enough financial backing to subsidize users for years—and the social norms around money not changing hands makes even that strategy challenging to pull off. Many social scientists have also recently argued that siren servers use techniques similar to those employed by casinos to make their content addictive.[35] Together these properties raise the power of siren servers to lock users into patterns that may not serve their long-term interests.

Second, as highlighted by economist Roland Bénabou and Nobel Laureate Jean Tirole in their incisive 2003 and 2006

analyses of situations like the Tom Sawyer problem, paying for an activity often undermines *intrinsic* motivations (such as entertainment and social pressure).[36] Paying for online data provision may signal to users that the activities they currently view as entertainment are actually labor benefiting the siren servers and for which they should demand payment, undermining the entertainment value. Paying may also undermine the perceived motives of social collaboration and participation that may yield social rewards to users for "being part of an online community." On a darker side, paying may also undermine the stickiness of content as it "breaks the spell" of online entertainment by making clearer the nature of the economic relationship.

Third, despite media accounts about the data economy, most users are still unaware of the value companies harvest from their data.[37] In order to pay users for supplying the most valuable data to siren servers, servers would have to make explicit requests for labels, comments, and other user input. As users become aware of the "creepiness" of the current situation, their attitude toward online interactions is likely to change in a manner that will be both costly and disruptive to siren servers, as well as unpredictable. Publicity about Facebook's experimentation with the emotional valence of its users' newsfeeds created a public backlash, and research suggests that users who become aware of the "creepy" surveillance of technology tend to become distrustful of digital services or to use them in ways that reduce the value of their data.[38]

Finally, realizing Lanier's vision for data as labor would require building a variety of sophisticated technical systems. The architecture of many digital systems would have to be adjusted to keep track of the origin of and uses of data, so users could be rewarded at least for the average value their data create but ideally to some extent for the unique value their data may oc-

casionally end up yielding.[39] ML systems would have to be designed to determine particularly valuable data to them; then their requests for data would need to be channeled to consumer facing products; and finally, these products would need to be designed to query the users for extra data in a minimally intrusive way.

Another part of this problem is that users could find it burdensome to transact with the Internet on a regular basis. We imagined that Facebook would offer Jayla $15 for a few minutes of her time, but what if the actual value of the information supplied by Jayla is worth 15 cents or 15 thousandths of a cent? Personal advisor systems would have to be built to guide user choices and receive only occasional user feedback while handling all payments. Even with such systems, a basic shift in user perceptions of and social attitudes toward online interactions would be necessary.

Conversely, there would have to be a more effective way for siren servers to ensure the quality and value of the data they are receiving. Several years ago, when Microsoft experimented with paying users for their data, large numbers of bots sprung up to exploit the system and extract large amounts of money without providing real value to the company. Without some careful way to keep track of users, which would necessarily impose further burdens on the users themselves, paying for data could easily be exploited.

The last three factors we highlight are mostly reasons that treating data as labor might also be socially undesirable. We believe these factors would be outweighed by the benefits in the medium term. However, when these factors are combined with their monopsony power, network effects, and interests in manipulating user psychology, it is unsurprising that siren servers have not yet undertaken this ambitious transition.

On the other hand, it is possible that siren servers that are poorer in data and lag behind these leaders, such as Amazon, Apple, and Microsoft, could have both the scale to make competition possible and the incentive to break up this unproductive monopsony. By creating an alternative ideology to the prevailing focus on "free" stuff online, they could help break the dominant business model of their rivals and open up a chance to compete. However, it is also plausible that the structure of the industry makes it unlikely that any private entity will voluntarily and on its own shift to a more productive model; and that broader social and regulatory pressure is necessary to catalyze change.

Workers' Struggle

Many aspects of the story we have told are unique to present technology and the norms that have developed around the Internet. However, the idea that monopsonistic power created by technologies with strong economies of scale would lead to undercompensated labor and thus retard both economic development and equality is not a new one. It is one of the classic themes of economic history and the central idea of the most famous economic historian of them all, Karl Marx.

A central intellectual aim of Marx's 1867 first volume of *Das Kapital* was to explain why the wealth and well-being of proletarians (workers without property) had, as of the mid-nineteenth century, improved so little since the end of feudalism.[40] Marx claimed to identify a necessary tendency of capitalists to "exploit" workers by holding their wages below the value they generated. Marx argued that these labor practices created what his collaborator Friedrich Engels called a "reserve army of the labor" (that is, a class of unemployed)

whose even more squalid condition would persuade workers to do anything to maintain their jobs.[41]

As economist John Roemer showed, Marx's conclusions are extremely unlikely to prevail if employers compete for workers.[42] However, they are exactly what one would anticipate in a world where capitalists conspired with each other, or had sufficient unilateral power, to hold down wages. Beatrice and Sydney Webb, a dynamic pair of late nineteenth-century British Radicals, advocated collective bargaining by workers, arguing that it would make production more efficient by raising wages above the levels that drove workers out of the labor force.[43] John Kenneth Galbraith, the mid-twentieth-century American economist we met in chapter 2, hailed unions as a necessary form of "countervailing power" required to maintain balance the power of monopsonists.[44]

This view has been partly vindicated by the research of subsequent economists. Economic historian Robert C. Allen shows that prior to the emergence of unions, British wages during the early process of industrialization hardly advanced at all despite improvements in technology.[45] Once unions managed to counter the monopsony power of British industrialists, not only did wages quickly increase, but the pace of overall productivity radically accelerated. Economists David Autor, Daron Acemoğlu, and Suresh Naidu believe that the breaking of monopsony power through labor unions, government labor regulation, minimum wages, and other reforms was critical to the further acceleration of productivity.[46] Beyond their role in collective bargaining, unions served other functions that helped support the "Fordist" mode of assembly line–based production that prevailed in the twentieth century: they screened and guaranteed the quality of the work produced by their workers

and helped them learn the skills required by a rapidly changing work environment.

To be sure, many other things were happening at the same time, making it difficult to trace clear lines of historical causation. Unions also brought many inefficiencies and rigidities, caused strikes, may themselves have accumulated significant market power, and so forth. The hostility they thereby created and the extent to which they became inflexible and outmoded has led to their decline in the last several decades.

Yet even as unions have declined, some of the conditions we describe above have important resemblance to the conditions that helped stimulate their growth and benefits. The monopsony power of siren servers, we have argued, may be holding down wages for data laborers at 0 (or more precisely at the value of the services and entertainment these laborers derive from using digital services). This may suppress the productivity of the digital economy by reducing the quality and quantity of data and contribute to the maldistribution of gains from AI technologies. An individual data worker lacks bargaining power, so she cannot credibly threaten to withdraw her data from Facebook or Google unless she receives a fair reward.

Furthermore, to realize the gains from data as labor, data workers will need some organization to vet them, ensure they provide quality data, and help them navigate the complexities of digital systems without overburdening their time. These triple roles, of collective bargaining, quality certification, and career development, are exactly the roles unions played during the Industrial Age.

It may be time for "data workers of the world (to) unite" into a "data labor movement."[47] A striking feature of the data labor market is that it is an international market, one that is

almost completely unaffected by borders and government regulation. Once people awaken to their role as data laborers—obtain a "class consciousness," if you will—organizations (sort of like unions) may emerge to supply data laborers with the means to engage in collective action. Imagine, for example, a data labor union that solicited members—the data laborers—by promising them higher payments for their data. Once the union obtained a critical mass, it could approach Facebook or Google and threaten a "strike" (also, effectively, a boycott because data laborers are simultaneously consumers of Facebook's and Google's services). The technical details would be complex, but we can imagine a range of possible approaches.

The union could simply call on its members to stop using Facebook or Google for a day if the companies do not negotiate. A more complicated approach could involve routing data labor through platforms set up by the union, so that the union could disrupt the supply of data if and when the Internet companies on the other side refused to pay reasonable wages. A Facebook user would reach her Facebook account through the union's platform, so that the union could enforce collective action among users by shutting down the account or providing limited access to the account for the duration of the strike. At present, an Internet Service Provider could organize such an action, though it would need to structure itself as a labor union to avoid antitrust charges.

It seems to us that these unions could be effective. Unlike traditional unions, they combine labor stoppages and consumer boycotts—because, as noted, data laborers are simultaneously consumers. During a strike, Facebook would lose not only access to data (on the labor side) but access to ad revenues (on the consumer side). It's as if autoworkers could pres-

sure GM or Ford not only by stopping production but also by refusing to purchase cars. Also unlike traditional unions, which must struggle to maintain solidarity during strikes, the data unions could enforce the "picket line" electronically. Furthermore, the very network effects that entrench digital monopolies would work against them in this scenario: it would not be much fun, but would be extremely embarrassing, to break a Facebook strike if all your friends were striking on the same day.

Finally, a data labor union might help foster digital competition by breaking the stranglehold on data of a few of the most powerful siren servers. The unions might find it optimal to share data between many different digital companies, rather than causing it to accumulate in one place. Of course, there are downsides as well—data unions, like traditional unions, might abuse their authority. However, we believe that at the present time, in light of the absence of any market—Radical or otherwise—in data labor, the gains exceed the losses.

A Penny for Your Thoughts

A first and necessary step before any of this is possible, however, is getting a quantitative grip on the value of data. Things that are not measured are not priced, and often once something is measured precisely, it begins to be priced organically. Systems for measuring the carbon footprint of individuals, companies, cars, and so forth have developed in the past decade. Even in the absence of legal carbon taxation, growing numbers of economic agents have begun to account for these carbon costs through voluntary offsets or by using them to guide company planning partly under social and consumer

pressure and partly because of concerns about potential future regulations. In this spirit, we believe the first step toward valuing individual contributions to the data economy is measuring these (marginal) contributions.[48] The field of "active learning" within computer science considers how to optimize the search for data (possibly at some cost) and offers a rich store of ideas to build on in answering these questions.

Second, appropriate technological systems would have to be built for tracing and tracking the value created by individual users. These systems would have to balance a number of competing concerns. On the one hand, they should try to measure with some precision which users are individually responsible for what data contributions, especially when these contributions are disproportionately large and/or those individuals would be unlikely to supply and invest in the unique data that make these exceptional contributions unless they receive these monetary incentives. Creators of valuable entertainments, experts in obscure languages who can aid computer translators, specialized masters of video games that can help teach computers to expertly play them as companions in multiplayer games, wine aficionados who can help train a computer nose: these are unique skills deserving of exceptional rewards. On the other hand, trying to track every detail of the ordinary use of a Facebook post is overkill and certain classes of data should be commoditized and paid an "average price" based on meeting overall quality standards, both to reduce the burden on the system and to insulate users from unnecessary risks based on whether their data end up being valuable.

Third, users will not want to have to make a cost-benefit analysis of the monetary value versus the hassle cost of every online interaction. While it is important that users are aware

of and acknowledged for the contributions they make and that the costs of services they use not be hidden from them, it would be impractical for most users to think through the financial value of every digital choice. Instead, most users will require guidance from an intelligent digital advisor that will filter and suggest opportunities that are lucrative relative to the hassle they impose—services that are worth it for users. This system will filter out "spam" that does not make sense for the user and will present the user with opportunities that do. Users can provide feedback, rating individual experiences or more likely giving comments or responses to system queries to help it learn user preferences.

Finally, a fair digital labor market would require a new regulatory infrastructure adapted to it. Minimum wage laws and related employee protections are poorly adapted to a world of flexible work where users make a variety of small contributions that supplement their main income streams. Governments would have to ensure that individual digital workers have clear ownership rights over their data, a step the European Union has moved toward with its General Data Protection Regulations, and that they have the right to freely associate to form data labor unions. Empowering users to not just be aware of their data but to be able claim the benefits of it will require allowing trusted agents to whom they delegate to access data in appropriate formats. This sort of technically literate and creative thinking about appropriate regulations for data as labor and related flexible work in the digital age (such as driving for ride-hailing services or hosting for home sharing) is at an early stage. But competition and countervailing union power will succeed only if regulations allow the flexibility for them to help shape a productive and fair digital labor market.[49]

A Radical Market in Data Labor

Suppose that the internet started paying you for your data. How would this change things? The first thing to understand is that it is not a quick path to riches for the masses. Even if the entire market capitalization of Google and Facebook were divided among American citizens, each would receive only a few thousand dollars. Divide the market capitalization among billions of users throughout the world, and the amount is even less. To be sure, the system we propose would increase the efficiency of the digital economy and therefore make more value available for everyone. But in the first few years, typical users would supplement their incomes with several hundred or perhaps a few thousand dollars.

How important a source of income data labor would become after a few years depends on how important AI turns out to be. Some commentators believe that AI will automate much of the economy. If true, data labor will represent a much greater source of income and wealth in coming years than it does at present, and in fact much of the market capitalization of digital companies is based on this possibility. If this is realized, data labor may grow to become a substantial fraction of many people's income. However, it is also possible that AI will to have limited applications, in which case data labor will never be more than a modest supplement to people's income.

To make a ballpark estimate of what gains we might expect, we suppose that over the next twenty years, AI that would (absent our proposal) not pay data providers comes to represent 10% of the economy. We further assume that the true share of labor if paid in this area of the economy is two-thirds, as in the rest of the economy; and that paying labor fairly expands the output of this sector by 30%, as seems quite reasonable given

the experience of productivity gains accompanying fairer labor practices in the early twentieth century. Then our proposal would increase the size of the economy by 3% and transfer about 9% of the economy from the owners of capital to those of labor. Applying the same logic as in chapter 4 about the effect of such transfers, this would lower the top 1% share of income by about 3 percentage points. While this may sound small relative to the whole economy, it would be a substantial contribution to median income for a household of four, raising it by more than $20,000, as much as during the thirty years following the world wars.

Yet even if data labor does become an important source of many people's income, there is no guarantee its fruits will be evenly distributed. Some people may have idiosyncratic cultural knowledge or abilities that will be particularly valuable to ML, while others will be too ordinary for their data to have much marginal value. Some data workers may contribute a little bit to a wide range of different ML processes, while others may contribute greatly in one area (such as language learning or cultural awareness) but little or nothing in other areas. We hope that the range of opportunities such a world would offer might allow individuals to specialize across a broader range of niches than at present, some opting for diversity and a more recreational work experience and others focusing on a concentrated passion. However, it is entirely possible that large inequalities would emerge and have to be disciplined by future reforms.

Beyond the direct income implications, paying people for data may also change the social understanding of the digital economy. Rather than feeling like passive consumers of internet services, users might see themselves as active producers and participants in the creation of value. We suspect that the

term AI would gradually give way to a more accurate understanding of the sources of value in digital systems such as "collective intelligence." Users would treat the useful insights of Siri and Alexa not as advice from robots, but as assemblages of human contributions, in the way they understand an encyclopedia or the insights on their Facebook wall.

As a psychological matter, this view does not seem impossible. People living in democracies seem to feel more empowered and active in politics than people living in dictatorships, even though the contribution of one's vote to policy outcomes is very small. When we "buy American" cars, we think of ourselves as purchasing the product of the labor of our fellow citizens, even though any individual American plays at most a tiny role in producing such products.

Yet, in many ways this change in the perception of consumers may be less important than the changes seeing data as labor may create for the data laborers themselves. Paying people for their data might make them feel like more useful members of society. In recent years, economists have begun to wonder whether large segments of the population will be unable to find work in an economy that places the most value on technical work that requires advanced education. Recent research suggests that the rise of video gaming is an important cause of the decline in labor force participation among young men.[50] Given current attitudes toward such activities, it seems plausible that such young men, some of them Internet trolls or bullies, may have a less than healthy relationship to the broader society.

Most people derive a sense of self-worth from making a contribution to society. In a world where individual digital contributions were appropriately valued by society, many video gaming young men could convert their enjoyment of

gaming into a productive skill. Given the trend toward the "gamification" of many productive tasks, it is not hard to imagine that the skills these young men have acquired in their life as gamers might help them earn a living if data were treated as labor. The untapped capacity of expert gamers deserves more respect, and more attempts at harnessing it for the social good, than it receives today. This would encourage gamers to develop their ability in a more socially valuable manner, yielding a sense of both personal dignity and political responsibility.