1 THE PITFALLS, PROMISES, AND CHALLENGES OF DATA

Scientific inquiry has always depended on data and various manifestations of data science. The nature of that reliance, however, has metamorphosed dramatically in the twenty-first century. An unprecedented quantity and breadth of information, the ability to share data efficiently among disciplines, ever-expanding computational power, and the democratization of algorithms across domains continue to revolutionize the scientific landscape.

Still largely absent, though, are systematic approaches to using big data to solve the most urgent societal challenges across multiple domains. This book chronicles initiatives underway at the Massachusetts Institute of Technology (MIT) and elsewhere to address that deficit. Our goal in this book is to share key lessons we've learned through the launch of a new transdiscipline of Data, Systems, and Society that applies pioneering technologies to complex challenges. In doing so, we hope to encourage academicians, practitioners, students, and funders to join a growing worldwide effort to use data science for societal good.

Our story will touch on key topics in the history of computing, data science, systems thinking, and the social sciences that contribute to the new methodologies and habits of mind needed to solve previously insoluble problems. Along the way, we'll describe the structure and evolution of our new entity as well as some breakthroughs stemming from our new transdiscipline that demonstrate the promise of novel thinking and interventions.

A Seminal Challenge for Data, Systems, and Society

On March 2021, the Kaiser Family Foundation (KFF) published an analysis of the demographic characteristics of individuals who were vaccinated against coronavirus disease 2019 (COVID-19) throughout the US between mid-December 2020 and March 1, 2021. Citing Centers for Disease Control and Prevention (CDC) data, the report featured two alarming findings related to equity. Of those who had received at least one dose of the vaccine, 65% were White, 9% Hispanic, 7% Black, 5% Asian, 2% American Indian or Alaska Native, fewer than 1% were Native Hawaiian or Other Pacific Islander, and 13% reported mixed or other race.

This was troubling because, as had been noted in a December 2020 KFF report by Samantha Artiga and Jennifer Kates "preventing racial disparities in the uptake of COVID-19 vaccines will be important to help mitigate the disproportionate impacts of the virus for people of color and prevent widening racial health disparities going forward." And yet, KFF analysis of 41 states showed a consistent pattern – Black and Hispanic individuals were receiving smaller shares of vaccinations compared to their shares of infections, deaths, and their percentages of the total population.

Even more worrisome was the revelation that race/ethnicity was known for only slightly more than half (54%) of those who had received at least one dose. The aggregate concerns raised by KFF include not only disparities in vaccination rates but also the extent to which gaps, limitations, and in data collection were limiting the ability of policymakers to assemble a complete picture of who was and was not getting vaccinated. The bright side of KFF's reporting was that those crucial gaps in the data – and the attendant inequities – were brought to light during the early days of vaccine distribution in the US when mitigation and correction could be pursued productively. Efforts to obtain more targeted data are definitely a must if we are to address such disparities.

Bigger Doesn't Always Mean Better

Data always have been the key to scientific discovery. But the collection and analysis of data, in and of itself, does not guarantee results. Misperceptions, misunderstandings, and mistakes often can be

traced back to flawed data sets – poor sampling, inconsistent collection or reporting, overly narrow investigation of phenomena, just to name a few root causes of "dirty" data. Reliance on such data can unwittingly prejudice observations and conclusions and provide the rationale for inadequate or counterproductive policy initiatives. Still worse are scenarios in which relevant data are suppressed, intentionally misrepresented or manipulated, or selectively curated to serve the predetermined objectives of the data collector or analyzer.

Unfortunately, so-called big data hasn't yielded the solution to these potential pitfalls. Simply collecting massive amounts of data doesn't guarantee that you will have the specific types of information you need to solve the particular problem you are working on. The CDC vaccination data cited at the beginning of this chapter is just one example of how a very large data set can have limited utility for decision-makers in the midst of a crisis. Policies based on data that underrepresents key segments of the population run the risk of being ineffective at best and counterproductive at worst.

The sheer volume of data we are collecting about almost everything also has the potential to make reliable data-driven decisions impossible. Often, the breadth and depth of information at our disposal far outweigh our brain's ability to account for every data point and potential trend line in a coherent and empowering way. Machines may do this better as long as they are fed with the appropriate data. As is the case with big data circa 2024, we have a lot of data about many things but not enough data about any one particular thing we need to understand deeply. And the new mathematical methods we are perpetually inventing – artificial intelligence (AI) and machine learning (ML) being prime examples from the early twenty-first century – continue to have their own limitations and blind spots. Large language models (LLMs), such as Bard and ChatGPT, are trained on millions of parameters of deep neural networks (DNNs). Much of that data is unlabeled or partially labeled, which can produce good results across many tasks but cannot deliver precise results in the majority of tasks.

As the number of individual users on social platforms such as Facebook grows into multi-billions, so does the cache of data being gleaned from social networks. Such data sets, however, are not collected methodically, are generally unlabeled, and lack consequential information about individual nodes. Those characteristics complicate the task of making causal inferences that would assist researchers and

decision-makers with policy development and problem-solving. The field of high-dimensional statistics has contributed many useful low-dimensional models that filter out noisy and irrelevant data, but the speed at which large and messy data sets grow will continue to challenge us for the foreseeable future.

Statistics: A Definition

Statistics is the practice or science of collecting and analyzing numerical data in large quantities and then using that data to make inferences about a whole population based on representative samples. This involves transforming data into models to aid in decision-making processes such as prediction and regulation. Statistics plays a fundamental role in the process of scientific discovery and serves as a foundation for many quantitative fields. It also comprises the field of statistical learning theory, which addresses essential questions related to learning models from data. In linear regression, for example, a simple linear model is fitted to one or more data sets to predict or classify data points. The theory provides a probabilistic framework to assess how well the model represents reality, considering aspects such as sample complexity and model evaluation. ML is founded on statistical learning theory and often refers to unstructured learning problems.

Timescales and Shortfalls

Often, the data we need and obtain occur on multiple, even divergent, time scales and lack the dynamism of the real-world phenomena from which they are drawn. Such challenges help explain why the task of optimizing the US electrical grid has confounded public- and private-sector organizations for decades. Operating costs and pricing, human behavior (demand and usage patterns), the performance capabilities and limitations of technologies, and the inherently deliberate nature of decision-making institutions function on vastly different time scales across broad geographic areas. And the types of data we gather from those phenomena often fail to conform to a single analytical tool or methodology.

If too much data has its drawbacks, too little data is even more unfortunate. As we saw from the example at the beginning of this chapter, well-intentioned policymakers can be stymied from drawing

actionable conclusions when the data on hand fail to fully account for key components of the problem. Phenomena and systems that are characterized based on small and/or narrow data sets also are much more easily misrepresented or manipulated for nefarious purposes.

Another data scarcity pitfall arises when we try to measure and analyze rare, but critical, failures in networked systems such as power grids, financial markets, and transportation systems. Small, idiosyncratic disruptions occasionally cascade into major breakdowns. Because those events are infrequent, we have limited information with which to build the data-driven predictive models we need. The 2021 paper "Cascading Risks: Understanding the 2021 Winter Blackout in Texas" by Joshua W. Busby and nine coauthors about the collapse of Texas' power generation capabilities in February 2021 highlighted our need to understand the mechanisms behind such catastrophes. If we correctly analyze those mechanisms, we may be able to make predictions about potential future failures with much less data. We must understand the underlying phenomena generating the data in order to make reliable conclusions or decisions.

Causality

Back in the days before the data science revolution, I would introduce new students to my research group by walking them around our lab space and showing them where they would be sitting. I often used the opportunity to joke that if the new student agreed to sit at a particular desk, they were guaranteed to graduate. My evidence? Every previous student who sat at that desk had graduated. Most new students would laugh at my quip and recognize the absurdity of my argument. On rare occasions, though, a student would accept my rationale at face value – which left me worrying that they didn't understand the difference between causality and correlation.

Nowadays, the causality/correlation distinction is fairly well understood in everyday contexts (e.g., a CNN news broadcast), and it is ubiquitous in many scientific and technological endeavors. This spans a variety of fields, including drug design, recommendation systems, and economic policy.

A fun illustration of causality and the presence of a confounding factor involves ice cream and sunburns. If we examine data on ice cream consumption and sunburns, we may observe a seemingly inexplicable

correlation that suggests that eating a frozen dairy dessert and getting over-tanned are related. Once you incorporate a third variable such as going to the beach, however, the correlation makes more sense. We can deduce the probability that eating ice cream and experiencing sunburns, when conditioned on going to the beach, are simply independent of one another.

Causality plays a crucial role in recommendation systems, with these systems serving as a primary example of extracting causal information from observational data. When examining sparse customer ratings data from Netflix, the challenge arises in assessing a customer's interest in a particular movie for recommendation. For instance, can we determine if customer A is interested in movie X based on their observed behavior?

Nearest neighbor methods attempt to address this by identifying another customer B, who has rated past movies similarly to customer A and has also given a high rating to movie X. The proximity argument is then used to infer that customer A may be interested in movie X.

However, there are several caveats to this approach. First, it may be challenging to find a single person who is highly similar to customer A. This challenge has led to the development of synthetic control groups, where a collection of individuals collectively approximates the behavior of customer A. More importantly, the absence of a rating for movie X from customer A may not be random; it could be due to a lack of interest. This introduces the possibility that the absence of a rating is confounded by some other unmeasurable variable. Consequently, it becomes challenging to confidently ascertain customer A's interest in movie X.

Drug designers cannot bring a product to market without establishing causation through randomized clinical trials (RCTs) and control groups that demonstrate the causes and effects of a drug. To mitigate the influence of unknown confounders, two populations of randomly selected subjects are created. The randomness ensures that these groups are not biased in specific ways, such as being composed entirely of women or individuals of a particular age.

In the trial, one group is administered the drug, while the other group receives a placebo. The effects are measured by calculating the average outcome within the first group and comparing it to the average outcomes of the second group. This comparison yields what is known as

the average treatment effect, providing a reliable assessment of the drug's causal impact while controlling for potential confounding variables.

At the random population level, RCTs are instrumental in determining whether there is a measurable difference between applying the drug and not applying it. However, despite these insights, we still face uncertainty at the individual level. In other words, we cannot definitively predict what will happen if a specific person who initially received a placebo is later given the drug.

The personalized treatment effect represents a deeper layer of causality that is of significant interest. While RCTs may contain information that allows for the assessment of this effect, it necessitates a more in-depth analysis of the data. Understanding how the drug interacts with an individual's unique characteristics, health conditions, and other factors requires a more nuanced examination beyond the broad conclusions drawn at the population level. This personalized treatment effect is crucial for tailoring medical interventions to individuals and optimizing healthcare outcomes.

Causality from observational data is the holy grail of statistics, and one that has a profound impact on how we explain data. But running experiments can be expensive and impractical, so we often find ourselves short of data that demonstrate causality. Human health initiatives – think gene therapy or finding a cure for Lyme disease – are notoriously difficult to study in live subjects, for example. Because the search for cures must persist in the face of data scarcity, we must increase and enhance our ability to examine mechanisms using models in research and development.

When working with purely observational data – including initiatives that don't lend themselves to RCTs – we must devise alternative approaches for determining cause and effect. This is often the case in econometrics when researchers construct instrumental variables and synthetic control groups that limit potential dependence on confounders. My colleague Alberto Abadie's article "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects" is a good reference for those who want to explore the topic in detail.

We see synthetic controls at work in contemporary American life when the board of governors of the US Federal Reserve System (the Fed) grapples with where to set interest rates. Historically, policymakers have observed that raising interest rates can produce a measurable effect

on reducing inflation. But observational data are not free of confounders. If the Fed is buying more federally issued bonds at the same time it is raising interest rates, for example, the resulting effect on inflation cannot be attributed solely to higher rates. Full-scale live experiments on the US economy are out of the question, so the governors must do the best they can with observational data and statistical modeling to establish causal relationships.

The Challenge of Interconnected Systems

Interconnections among subsystems create very different dependencies and phenomena from systems in isolation – though we often take these interconnections for granted. When we are sailing along smoothly at 30,000 feet on a commercial flight between Atlanta and Boston, we take it for granted that the plane is engineered to travel safely between the two cities in the hands of a skilled pilot (or more likely autopilot). Even if we are aware of the complex guidance and decision-making systems that assist pilots, we seldom think about the interconnection between the airplane and the global control systems that coordinate flight paths, takeoffs, and landings. Disconnecting an airborne commercial airliner from this complex system could be catastrophic, and attempting to understand the workings of the global aviation industry apart from this system would be futile.

Those interconnections can also result in cascaded failures, such as what happened in 2016 when Atlanta experienced a three-inch snowfall (a storm that Boston residents would consider negligible). As a result of that single weather event, air traffic experts projected three-hour delays for certain flights leaving Los Angeles. We all understand that delays in airports are interlinked and that delays at Atlanta International Airport could result in cascaded delays at LAX. But why were the flights in Atlanta delayed in the first place?

Airport operators in Atlanta certainly have the necessary equipment to remove three inches of snow from the runways without interrupting takeoffs and landings. The city of Atlanta, by contrast, does not invest in infrastructure for snow removal at the same scale because three-inch snowstorms are rare events. When the 2016 storm congested Atlanta-area roads, pilots and crew members were unable to reach the airport in time for their scheduled flights (despite the availability of transit to the airport). The interconnections and dependencies between

ground and air transportation were the true cause of the systemic failure in airline travel that day. We'll dive deeper into interconnections, feedback, and complexity later in this book.

I'll have more to say in upcoming chapters about the strategies we employ at the Institute for Data, Systems, and Society (IDSS) – abstraction, dissecting mechanisms of failure, statistical analyses, causality, and mapping human and institutional behavior through the lens of social sciences – to zero in on viable solutions to complex, data-rich challenges. Before I launch into an overview of our transdisciplinary methodologies, however, we must consider the issues of privacy, bias, and fairness in data collection and use.

Exploring the Limits of Privacy

In March 2021, journalist Kashmir Hill reported in *New York Times Magazine* that a little-known company called Clearview AI had created a database with three billion images of people. The photos came from social media, employment sites, YouTube, and Venmo – all part of the public web – and included links to the sites where each of the images originated. When the activities of Clearview were first exposed by journalists, several companies (Facebook, Google, and LinkedIn among them) pursued cease-and-desist actions, all of which failed.

The existence of such web-scraping technologies was not, by itself, newsworthy. The startling aspects of the revelations were the scale of the human image base (many times larger than similar products used by law enforcement at that time), the fact that individuals depicted had no idea their images had been collected, and the list of people and organizations accessing the technology. According to Hill's reporting, *BuzzFeed* leaked an inventory of users that included Bank of America, the NBA, and a billionaire investor in Clearview who used the image base to ID his daughter's dinner date who was otherwise unknown to him.

Although this example represents only a tiny portion of personal information being collected without our knowledge, it highlights well the extent to which individual privacy has been compromised by massive data-collection activities. As we race ahead to solve pressing societal challenges with rich new information sources, must we accept the losses of personal information we've already sustained and seek to limit the pitfalls? Or should we be fighting to claw back some of the

privacy we've unwittingly sacrificed? The push and pull implicit in these questions reflect a societal debate that may never be fully resolved.

Privacy Gains and Losses

One reason we are unlikely to return to pre-digital-age notions of privacy is that people simply aren't inclined to do so. A January 2019 survey by the Center for Data Innovation found that 58% of Americans would give sensitive personal data – such as location, medical, and biometric – in exchange for immediate or long-term benefits. Expected ROIs ranged from increased convenience (easier logins, free navigation assistance, etc.) to cures that might improve the health and well-being of ourselves and others.

On the flip side, the Pew Research Center reported in November 2019 that 66% of Americans considered the potential risks of allowing governmental entities to collect their personal data to outweigh the benefits. Wariness was even greater (81% expressed an aversion) when applied to private companies. Yet businesses forge ahead – Amazon, for example, began requiring its delivery drivers to submit to AI surveillance of their locations, movements, and biometric data in 2021 – while US regulatory bodies seem to be frozen in place like the proverbial deer in the headlights.

People are conflicted. They want to believe that big data can benefit both the individual and society without subjecting either to high levels of risk or intrusion. Those of us who work with and conduct research into data and systems believe we can achieve this balance, but we have a lot of convincing to do before we can gain the necessary buy-in from the general public.

Data Collection, Biases, and Algorithmic Fairness

While many of us view data as key to solving our most complex challenges, it also has created seemingly intractable, society-wide problems. One of the earliest and most notorious misapplications of data was the introduction of racial categorizations into the US Census in the mid-nineteenth century. A substantial majority of researchers today agree that the genetic argument for those categorizations was tenuous at best – the differences in genetic structures are continuous and far from belonging in distinct, enumerated clusters. Phenotypes in genetic

structures do not differentiate based on these divisions, and sociologists assert that racial structures emerged from social relations that the Census categories reinforced. My colleague, MIT Chancellor Melissa Nobles, covered this topic well in her book *Shades of Citizenship: Race and the Census in Modern Politics*.

Many far-reaching public policies – voting rights, housing, healthcare, biometrics, policing, and crime prevention, to name a few – have relied on these categories and reinforced patterns of structural bias and individual racism. Over time, such policies reinforced divisions and created socially defined races that favored some groups and disadvantaged others. Decades of redlining by home mortgage lenders, for example, prevented Black borrowers from buying property in higher-value zip codes. As Chat Travieso observed in "A Nation of Walls: The Overlooked History of Race Barriers in the United States," the practice created a self-fulfilling cycle that prevented the accumulation of generational wealth and widened the wealth gap between Black and White Americans.

In recent decades, patterns of surveillance data collection in predominantly Black neighborhoods have created another self-fulfilling cycle of bias. Biased assumptions about higher crime rates in Black neighborhoods were used to justify greater surveillance. Closer observation resulted in more arrests, and the data were used to justify even more surveillance. Now you have a data set showing that more arrests occur in predominantly Black than in predominantly White neighborhoods, lending further credence to biased policing and crime prevention policies. Readers will find a rigorous analysis of those counterfactual effects in the 2017 paper "Counterfactual Fairness" by Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. When fed into ML algorithms, data showing higher arrest records generate predictions of more crime and justify even more police presence. It should come as no surprise that the policies and practices borne of this biased data collection and analysis actually have diminished public safety in affected areas.

Members of the data science community are diligently seeking to bring fairness to algorithmic processes by applying methods that correct for biased data sets. "Racially unaware" algorithms attempt to remove all references to race from data, but externalities in data sets complicate this process and leave many unconvinced of the efficacy of this approach. A different method known as statistical parity (SP)

compares algorithmic decisions across two or more categories of people to identify comparability – or lack thereof – in assessments of risk, false positives, false negatives, and other factors. Unfortunately, SP has proven inconsistent in detecting bias when applied to complex, real-world systems. For a detailed tutorial on how fair-minded ML practitioners can better serve marginalized and oppressed populations, I recommend "The Fairness Field Guide: Perspectives from Social and Formal Sciences" by Alycia N. Carey and Xintao Wu.

Making Data Count

On any given day, business management newsletters and periodicals publish more advice on monetizing data than the average person can read in 24 hours. Many offer some version of "You're collecting all this data, now here's how you can use it to boost profits." That plethora of guidance speaks to the fact that accumulating massive amounts of data became a minor obsession for many organizations long before their leaders had any idea what to do with that information.

Facebook and Google, of course, led the way in converting big data into cash flow, with Amazon hitting its data-monetization stride in the late 2010s. Robert J. Shapiro, former undersecretary of commerce in the Clinton administration, and Siddhartha Aneja, a policy analyst at the Georgetown Center on Poverty & Inequality, reported in 2019 that Amazon appears to have more than doubled its earnings on user data between 2016 and 2018. The pair also calculated that Facebook profited from users' personal information to the tune of $35.2 billion in 2018 – 63% of Facebook's total earnings that year.

Clearly, the pitfalls of big data discussed earlier in this chapter aren't preventing companies from using our personal information to haul in a great deal of money. I'm not suggesting that we should – or could – dramatically restrict data monetization, provided those activities meet widely accepted societal standards for consent, transparency, and equity. I will argue, however, that boosting the bottom lines of businesses is far from the greatest societal benefit we can reap from big data.

We know, for example, that heterogeneous, dynamic data on climate, soil conditions, and disease outbreaks can be just as life-changing for small- to medium-sized potato farmers in Peru as deep public health data were to the entire world during the COVID-19

pandemic. Hundreds of projects with similar societal objectives are underway throughout the world on any given day, and many thousands more will follow in the coming decades if we can gain peoples' trust that our data-gathering activities are governed by rigorous ethical, inclusive, and respectful guidelines.

At IDSS, we contend that data collection – and the ethics, rationales, and methodologies behind it – are at least as important as the algorithms we devise to analyze the information we gather. Human behaviors and incentives must be carefully accounted for across the full spectrum of applicable populations if we are to create or improve the systems that dominate contemporary life. Climate, food, energy, transportation, healthcare, education, finance, commerce, media, governance, and even our understanding of what it means to be human all stand to be enhanced with broader and more effective uses of data.

At the same time, we must empower individuals to secure their personal information from unnecessary intrusions, unauthorized disclosures, and intentional or unintentional misuse by third parties. If we clearly articulate the problems we are trying to solve and are transparent about how we will use the data we are collecting, I believe we will increase trust among the general public. In doing so, we also will boost individuals' willingness to participate in the types of data-gathering efforts that will yield long-term societal benefits.

The Income-Inequality Debate: Case in Point

Thomas Piketty's influential research, including his seminal work *Capital in the 21st Century*, has heightened global awareness of escalating income inequality. His comprehensive analysis of recent decades in the US exposes a widening gap between the top 1% and the bottom 50% of earners. Between 1980 and 2020, for example, the share of pre-tax income among the lower 50% of earners declined by 9.6%, while the share among the top 1% of earners rose by 8.2%. Piketty attributes this trend to a persistent phenomenon – the growth rate of inherited wealth surpasses earned income during periods of sluggish economic performance. The higher return on capital during those periods perpetuates the concentration of capital and wealth among a privileged few. Factors related to managerial power also contribute to the exacerbation of inequality over time. Piketty argues that mitigating effects of that overall trend in the first half of the twentieth century – wars,

industrialization, education, and economic growth – were no more than temporary setbacks to the broader dynamic of rising income inequality that continues to this day.

Piketty proposes a range of governmental interventions to mitigate the widening gap, emphasizing increased taxation for the wealthy (e.g., capital gains and inheritance taxes) as well as intensified efforts to combat tax evasion. He also advocates for substantial social investments in areas such as education, healthcare, and debt cancellation. The effectiveness of these pivotal changes, he asserts, rests on the conclusiveness of the data analysis conducted, underscoring the paramount importance of rigorous and responsible data science.

Piketty's prescriptions, though widely embraced, also have inspired pushback. Gerald Auten from the Office of Taxation at the US Treasury and David Splinter of the Joint Committee on Taxation of the US Congress, for example, contest Piketty's methodology in their article "Income Inequality in the United States: Using Tax Data to Measure Long-Term Trends." They argue that Piketty and collaborators Emmanuel Saez and Gabriel Zucman at the World Inequality Lab may have systematically attributed more wealth to the top 1%, thereby potentially overstating the issue of inequality.

Auten and Splinter contend that the increase in income inequality during recent decades is approximately half the amount estimated by Piketty, Saez, and Zucman. The ongoing debate has evolved through multiple iterations and rebuttals between the two camps, culminating in the 2024 publication of the most recent Auten and Splinter paper in the *Journal of Political Economy*. This scholarly duel highlights the intricacies of data science and emphasizes the critical role methodology plays in deriving accurate and reliable conclusions from economic data. It also underscores the inherent challenges in achieving a data-driven understanding of income inequality, in particular, and of complex socio-economic issues, in general.

The Core of the Debate

One intriguing aspect of the debate between Piketty, Saez, and Zucman on one side and Auten and Splinter on the other is that both camps rely on common data sets for their analyses. Those data sets include information from the Internal Revenue Service, Federal Reserve Board, Bureau of Economic Analysis, and miscellaneous data sets and

research that estimate unreported incomes in various ways. The two sides also share a common understanding about the timing of various government interventions that might have impacted some misreported income (e.g., which tax laws affected outcomes when they took effect). The sides agree, as well, that income should be associated with individuals. Despite all that common ground, other factors cause the two camps to diverge.

The Piketty/Saez/Zucman method ranks incomes based on the total amounts reported by tax filers (with jointly reported income split evenly between filers) and then compares the total income of the top 1% to that of the bottom 50% of earners. The Auten/Splinter approach, by contrast, ranks incomes by integrating the number of children in a family and then normalizes income by the unit's size (i.e., dividing by the square root of the number of people). Once ranked, Auten/Splinter assigns the full income (not the normalized one) to its respective grouping.

When you apply the Auten/Splinter approach to a couple married filing jointly with four children and a $500,000 income, the effective income for ranking is $204,124. In contrast, if you apply the Piketty, Saez, and Zucman approach, the effective income for ranking is $250,000. Since after ranking both groups agree that the income for that family unit remains $500,000, the Auten and Splinter approach creates a higher average income for the lower 99% of earners as compared to the Piketty/Saez/Zucman ranking method. By placing higher-income units into lower groups, the Auten and Splinter method effectively reduces the overall calculation of the income gap. As one might expect, the Piketty/Saez/Zucman team rejects that approach and asserts that children and other dependents who are not income earners should be considered part of a consumption model rather than an income distribution model.

Undisclosed Taxable Income

In a related line of inquiry, numerous researchers have examined the complex issue of misreported income that should be – but is not – taxed. Although audit data can provide insights into this aspect of wealth accumulation, it also poses a sophisticated modeling problem. Audits lack randomization and reflect government policies that are grounded in prior beliefs about income-hiding behaviors within certain

groups. Those beliefs and behaviors undoubtedly fluctuate whenever tax laws change. As a consequence, researchers in this domain must create estimates of hidden taxable income for various income groups by inputting the data on undeclared income and correcting for biases. Such models must also estimate the percentage of evaders within each group (often referred to as the frequency of evasion).

Both the Piketty/Saez/Zucman and Auten/Splinter camps rely on the same frequency-of-evasion research to quantify and allocate misreported income that should be taxed (i.e., evaded income). Nonetheless, the allocation of evaded income turns out to be the largest driver of differences in income-share estimates between the two camps. Piketty and his collaborators allocate evaded income to each group of taxpayers in proportion to reported income. Auten and Splinter, however, use information about the frequency and magnitude of tax evasion to allocate evaded income to randomly selected filers by reported income group. Both camps then re-rank taxpayers by income to determine the top 1% and bottom 50% of earners. Just as with their divergent approach to income ranking, the Auten/Splinter methodology for tax evasion attributes more generated income for the bottom 99% of taxpayers than the Piketty/Saez/Zucman's approach and produces a narrower gap between the two income groups. While both approaches are sensitive to estimates of unreported income, I contend that the Auten/Splinter approach results in a more granular allocation of evaded income.

Transparency and Data Availability

The intense debate between the opposing income-inequality camps is, in my view, a fine example of the robust scientific discourse we seek to promote within the Data, Systems, and Society domain. Piketty and his collaborators have made their data available to the scientific community to analyze and potentially replicate. They also have acknowledged errors in their work, even as they have defended their overall assumptions and findings. Without such transparency, progress on complex socio-economic challenges would be greatly impeded.

The income-inequality debate highlights the multifaceted nature of analyzing and responding to societal-level problems. Allocating government expenditures in areas such as healthcare, housing, or

education, for example, illustrates the need for a consistent and defensible methodology. Accounting for auditing behaviors and trends highlights the need to understand factors that introduce bias into data. And despite many shared assumptions and information, a few key divergences in methodology produced significantly different – and comparably robust – findings. With respect to income inequality, the work has significantly advanced our understanding of the challenge. In a broader sense, the combined efforts of both camps have demonstrated the power and potential of data and systems thinking.