Foreword

by Daniel L. Goroff
Vice President and Program Director
Alfred P. Sloan Foundation

This is an important Handbook, compiled by an important institution, on an important topic. The Alfred P. Sloan Foundation is therefore a particularly proud sponsor of the Innovations in Data and Experiments for Action Initiative (IDEA) of the Abdul Latif Jameel Poverty Action Lab (J-PAL), which has taken on this endeavor, and of work on administrative data generally.

Many think of J-PAL as an advocate for randomized controlled trials (RCTs). This is true, of course, and the world is better for it. Others realize that J-PAL stands for more than econometric improvements. J-PAL is also about collective responsibility, for example. By bringing the laboratory model to the social sciences, J-PAL promotes new ways of designing, staffing, documenting, crediting, and replicating experiments that produce reliable results. Indeed, researchers leading this movement seem to have priorities that go beyond producing yet another paper for their own CVs. The shared goal they pursue instead—relentlessly and with great integrity—is to discover meaningful answers to important questions.

How is J-PAL bringing about this reorientation of empirical social science as a profession? Taking a page from the behavioral economists, nudges tend to succeed by making change seem easy, attractive, social, and timely. As a replacement for how lone professors have traditionally worked with their graduate and postdoctoral students, the laboratory model goes a long way on each of these four dimensions, thus providing a new technology for producing reliable research results. Among

those interested in empirical evidence, there is ample demand for such results, too, as the world struggles with everything from poverty to pandemics and from prejudice to polarization. Large-scale surveys, a traditional source of insights about matters like these, are no longer seen as fully adequate to the task due to rising costs, slow turnaround, sampling frame challenges, and declining response rates.

So, when it comes to generating empirical evidence, we have a novel production technology together with weakening competition and robust demand for the outputs. What about the inputs? Besides the laboratory labor, there is also a need for data. Wait—don't we usually think of research data as a product of this process? Suitably refined and polished, after all, we store those data sets away in repositories in case someone else ever wants to admire them. This Handbook is not about that, but rather about the new and promising role that administrative data is beginning to play as an enabler of exciting research.

What counts as administrative data? There are many definitions. I, for one, take it to mean any information not originally collected for research purposes. That includes transaction descriptions and other records compiled while conducting public or private sector business of all sorts. Unlike when dealing with well-designed and well-curated research data sets, no metadata, comparison groups, representative samples, or quality checks can be assumed.

Some therefore refer to administrative data as *digital exhaust*. That characterization certainly evokes origins as an unintended byproduct but fails to convey the potential value. Others speak of *found data*. That brings to mind an oasis stumbled upon in the desert. Unlike exhaust but more like an oasis, many like to classify administrative data as a public good.

I argue that this Handbook suggests a better metaphor—at least implicitly. The contributors' more explicit goal is, of course, to help facilitate and promote the use of administrative data in the production of high-quality empirical evidence. In terms of nudging researchers in that direction, this is already an attractive and timely proposition. In fact, commercial applications of administrative data are all the rage

throughout the rest of society. Without more active roles for independent researchers and academic standards in this data revolution, there is a danger that only a few large and rather secretive institutions will either know—or think they know—what is going on in the world.

The challenge is that, as a goal to nudge toward, repurposing administrative data for use by researchers has been neither easy nor social. The Handbook chapters that follow present many examples of how the process can be made less burdensome for individuals and more beneficial for society. One way of appreciating the value of such advice is to consider the potential costs incurred without it:

Fixed Costs

Some holders of administrative data charge researchers for access. Even data that are supposed to be public by law, like the federal tax returns of charitable organizations, may only be available in bulk for a fee. Voter rolls and company registers must be purchased in certain states but are free to download in others.

Even after paying any such initial fees, administrative data sets usually need extensive preparation and attention prior to computing any statistics. The cleaning, documenting, linking, and hosting of files can be quite demanding. If the information is private or proprietary, then setting up an enclave or other protections also incurs expenses.

The case studies in this Handbook detail how much time and effort it can take to manage administrative data even before any research can begin. Currently, every investigator tends to start anew by negotiating their own access, doing their own cleaning, and making their own linkages with little incentive to share anything other than the final findings. We can do better. The lessons this Handbook proffers, and the coordination it suggests, show how.

Marginal Costs

Beyond routine maintenance, the budget implications of calculating one more statistic from a well-prepared, well-proportioned, and well-

hosted data set should be pennies at most. But there are other costs as well. When dealing with confidential information, for example, it follows from theorems described in this Handbook that every new query answered about a given data set leaks some privacy and depletes the *privacy loss budget* that should be fixed in advance. Even if the data set has nothing to do with people, every new query leaks some validity, too, and depletes the *statistical significance loss budget* that should also be set in advance. The chapters on disclosure avoidance methods and differential privacy explain how query mechanisms that satisfy ϵ-differential privacy control the rate at which simply trying to answer the questions that researchers submit about a given data set eventually and inevitably uses up the privacy loss and statistical significance budgets. Once spent, responsible curators are supposed to stop accepting queries altogether.

Remember this next time you hear that open data sets are a "public good" just like lighthouses or unpatented discoveries. Open data may serve the public good to be sure. Technically speaking, however, a research data set is not only excludable but also rival in the sense that with use it gradually loses its ability to generate safe and reliable evidence. This has consequences regarding the provision of administrative data for research purposes that the Handbook explores and that I will revisit below.

For now, note that we can only slow the rate at which privacy and validity evaporate with data use. No technological advances or other cleverness can prevent such leakage altogether, according to the theorems. What to do? Moving to new data sets, say either resampled ones or "set-asides" reserved from the original, can not only refresh budgets but also provide new perspectives. Another strategy is rationing direct access to data that would otherwise be overused. Exploratory research can be performed on high-quality synthetic data without impacting privacy or validity budgets at all. Tentative statistical or modeling conclusions obtained that way can then be sent to validation, or verification, servers for confirmation. These servers do access the original data but are designed to use only small portions of the privacy or validity budgets. The only researchers able to query the original data

would be those whose explicit, important, and pre-registered hypotheses cannot be tested otherwise due to linkage or other requirements. Such a regime has been shown not only to generate publishable results but also more reliable results than research based on p-hacking, data dredging, selective reporting, and other common practices.

Transaction Costs

Negotiating a Data Use Agreement (DUA) often requires considerable time, tact, and trust. As described in the chapter on data use agreements, legal technicalities and bills can be formidable but surmountable. All may seem to go well until some new player or policy sends everything back to square one. Case studies in this Handbook highlight just how to engineer mutually beneficial relationships between data holders and data users by avoiding or overcoming such frictions.

Economists who study transaction costs suggest that, when frictions are onerous, the solutions are often institutional. There is a role here for intermediaries who can deal with entire sectors of similar data holders on the one hand and with entire classes of data users on the other. This has to be more efficient than everyone negotiating pairwise agreements one at a time.

Examples range from the Institute for Research on Innovation and Science (IRIS) at the University of Michigan, which processes, protects, and provides administrative data gathered from universities about grant expenditures, to the Private Capital Research Institute (PCRI), which does the same with data from private equity firms as described in the PCRI's chapter in this Handbook. Some refer to such intermediaries as *Administrative Data Research Facilities*. The staff of each includes experts on data governance who also know the data-holding sector and the data-using sector well enough to deliver valuable benefits to both.

Opportunity Costs

Professors lament that, absent such intermediaries, the time and effort they spend trying to secure administrative data keeps them from pur-

suing more valuable tasks few others can address. This has particularly been the case, for example, in their quest for social media data held by tech platforms. Arguably, researchers have paid insufficient attention to challenges such as protecting privacy, identifying specific hypotheses suitable for testing with the data if obtained, compensating for the fact that such data do not constitute a representative sample of a well-defined population other than the users of a particular platform, devising ways to combine administrative data with survey or experimental data, etc.

Indeed, obsession with "getting the data" may blind researchers to other approaches or considerations. Most administrative data, after all, are only observational. Unless it describes suitable treatment and comparison groups, such data can rarely, if ever, yield robust causal conclusions. Running a well-designed RCT can, of course. RCTs usually require not just access to administrative data, but also the active cooperation of administrators in carrying out an experiment. Chapters in this Handbook provide examples from around the world where concentrating on how to answer an important question, instead of just how to obtain an attractive data set, has paid off handsomely.

***

Faced with all these costs, researchers naturally look for funding to cover expenses. That includes making proposals to grant-making organizations like the Alfred P. Sloan Foundation. When describing my work there, I often say that I am in the public goods business. That framing, when invoked in discussions of open data as a pure public good, suggests that the provision of data depends on solving a collective action problem, that is, a game where the natural Nash equilibrium fails to be Pareto efficient.

Under such circumstances, social science lore recommends nudging players to take their social obligations seriously and to internalize more of the benefits that might accrue to others. J-PAL and similar groups have made progress this way, as described above, motivated by compelling goals like the alleviation of poverty and supported by substantial grants from private and public sponsors. But while philanthropy

can proudly provide start-up funds, the sustainable provision of public goods ultimately depends on fundamental shifts in cultural, institutional, or legal support.

In other words, calling a commodity a "public good" may sound like praising it as worthy for funding. But to a grant-maker, the technical term "public good" just signals that, short of tax dollars or philanthropic support, financing will be difficult and sustainability will be very difficult. Cases where grants do help a community solve a collective action problem and provide a public good can be very productive, compelling, and gratifying, of course. The Handbook describes excellent examples, including the tools, systems, knowledge, and access mechanisms that facilitate research on administrative data.

Not everything of social value has to be a public good like this in the technical sense. As chapters in the Handbook indicate, conducting research on a data set—administrative or not—uses up its evidentiary value, especially if the data describes sensitive information about individuals. Talk of budgets, in this case for privacy and validity, evokes the way economists usually analyze the provision of commodities other than public goods.

From this point of view, we have a familiar scarce resource problem—but with high initial costs, low marginal costs, and the potential to enable a wide range of valuable activity over time. Solutions to such problems are often called infrastructure projects, particularly ones that result in reduced transaction costs, too. Monopolies or duopolies tend to play a role, justified by the positive externalities associated with sound infrastructure. Financing is not necessarily that much easier than for a public good but can also generate significant social benefit if designed well. Like railway or communications nodes, institutional intermediaries in this case could be connected to form an efficient network that traffics in administrative data by following trusted standards and practices.

Building these nodes, whether they are called Administrative Data Research Facilities or not, thus represents capital investment in research infrastructure. The Alfred P. Sloan Foundation's enthusiasm about providing data for economics research is, like the chapters that follow,

based on realism both about the economics of research data and about the promise of administrative data in particular. Others wishing to join this adventure may similarly find inspiration in this Handbook's account of how capital and labor can be organized to help answer important questions by transforming administrative data into high-quality evidence.