Conclusion: Migration Research in Times of Ubiquitous Digitization

Rinken, Sebastian; Pötzschke, Steffen

Chapter 11
Conclusion: Migration Research in Times of Ubiquitous Digitization

Sebastian Rinken and Steffen Pötzschke

Microprocessors and the Internet are outstanding cases in point for the increasingly frenetic pace of technological innovation. Starting in the 1950s and 1980s, respectively, computing speeds and interconnected data volumes have grown exponentially to become two related processes that have triggered profound changes in all kinds of productive, commercial, administrative, cultural, and social activities. Yet, when surveying the history of technology from a bird's-eye perspective, it is striking how many inventions—especially since the advent of industrialization about two centuries ago—have made a decisive mark. Mankind's collective path from subsistence communities to ubiquitous digitization is littered with milestones, and the number of important innovations is so large that it is difficult to arrive at a convincing shortlist. Was the lightbulb a more disruptive novelty than the automobile, or vice versa? How does the smartphone compare to the steam engine?

A focus on information and communication technologies (ICTs) helps us appreciate the truly epochal status of the digital revolution. Throughout the entire history of mankind, only one similarly momentous achievement stands out in this realm—the emergence of written language in ancient Mesopotamia (oral language is pre-technological, since it lacks non-biological hardware) (Majó, 2012, pp. 67–69). About 4000 years ago, the combination of novel coding (signs, alphabets) and storage technology (papyrus) started to smash the barriers of time and space associated with the physical range of humans' hearing and sight. Nowadays, the combination of novel coding (bits) and processing technology (microchips, Internet) is smashing the barriers of time and space erected by a range of mutually

S. Rinken (✉)
Spanish Research Council (CSIC), Institute for Advanced Social Studies (IESA), Córdoba, Spain
e-mail: srinken@iesa.csic.es

S. Pötzschke
GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany
e-mail: steffen.poetzschke@gesis.org

S. Pötzschke, S. Rinken (eds.), *Migration Research in a Digitized World*, IMISCOE Research Series, https://doi.org/10.1007/978-3-031-01319-5_11

incompatible (physical, electric, chemical, and electronic) means of storing and transmitting text, sound, and images. This unprecedented conversion to one universal format makes an ever-increasing volume of information available, at least potentially, to everybody, anywhere, anytime. By comparison, Gutenberg's celebrated invention was relatively minor, since while facilitating economies of scale—a feat that changed the world, to be sure—it left the extant information coding and transmission system largely unaltered.

By definition, since the business of scientists is distilling information into knowledge, the digital revolution profoundly affects the scientific endeavor. This insight is especially true for empirically-minded scientific disciplines that devote considerable effort to obtaining data in the first place. In survey research, as in medical trials, the process of information-gathering must follow strict rules for results to be valid, the most basic of which concern the selection and handling of study participants. However, in recent years, survey quality standards have become increasingly difficult to achieve due to growing coverage and non-response biases (Groves, 2011) and—especially when tackling sensitive items—persistent (if hard-to-measure) response bias (Krumpal, 2013). When surveying migrant populations, these challenges are exacerbated by added difficulties: international migrants, including refugees, tend to elude established sampling procedures, are often difficult to locate, may resist or resent the interviewee role, and require multi-lingual questionnaires and/or linguistic assistance—in short, they are notoriously hard-to-survey (Tourangeau et al., 2014).

It seems obvious that innovative ICTs are a game-changer in this context, since traditional time-space restrictions are particularly bothersome when targeting highly diverse and mobile populations. Yet, are migration scholars seizing the opportunities afforded them by new technologies?

The rationale of this book is based on the hypothesis that migration studies have even more to gain from the digital revolution than most other fields of social research. Each contribution in its own way encourages migration scholars to explore the added advantages granted them by innovative technologies and approaches. None of the authors, much less the editors, advocate an uncritical adoption of new technology: we all agree that its benefits have to be weighed carefully against its limitations, advantages put into perspective, and risks adequately managed. Yet, all the contributors agree that inertia is not an acceptable option, and so all the book chapters prod migration researchers to explore new data types and technological tools actively, rather than continuing to depend exclusively on accustomed data collection procedures. Because such experimentation inevitably entails a learning curve, we believe that migration scholars stand to benefit, both individually and collectively, even from mixed experiences. Before resuming a general discussion on how migration research is affected by the relentless process of digitization, we summarize the objectives, procedures, and results of each chapter.

11.1 Added Traction: New Tools for Sampling and Data Collection

The five contributions to the first part of the book address a variety of ways that new technology can improve the collection of "designed" data from purposefully sampled respondents. Three chapters address the twin problems of sampling and locating highly mobile, as well as oftentimes dispersed, populations, and two chapters focus on the challenges of target populations' markedly varied sets of linguistic competences.

In *Innovative Sample Designs for Studies of Refugees and Internally Displaced Persons*, Stephanie Eckman and Kristen Himelein focus on the procedures for drawing probability samples of forced migrants. More specifically, they discuss approaches to sampling that are apt for implementation in face-to-face surveys conducted in developing countries. They distinguish between three contexts of sample recruitment that depend on the characteristics of the study population: forced migrants living in camps, urban settings, and those who lack even a moderately stable place of residence ("on the move"). Regarding each of these scenarios, the authors explore a range of sampling options and highlight specific challenges and avenues to address them. The implementation of these options relies on new technologies to varying degrees and in various ways. One fascinating example is the use of satellite pictures or images collected by aerial drones (which can be deployed on-site by research teams) for generating real-time maps of migrant dwellings, which enable interviewers on the ground to employ aleatory (route-based) sampling plans. Other inspiring cases of new technological options include the use of geographic information system software and GPS-equipped interviewers in the recruitment of highly mobile migrants into a sample, and the incorporation of digital trace data in sampling strategies. While the tools of choice depend on each particular study's objectives, budget, and time-frame, the general take-home-message of this chapter is that regardless of the data collection mode, it is worthwhile for researchers to think creatively about how new technologies can improve the research process. Also, as this chapter shows, the technological enhancement of rather traditional data collection modes can be just as advantageous as the incorporation of new data types.

In *Targeting on Social Networking Sites as Sampling Strategy for Online Migrant Surveys: The Challenge of Biases and Search for Possible Solutions*, Anna Rocheva, Evgeni Varshaver, and Nataliya Ivanova shift the focus from sampling for face-to-face interviews to sampling for online surveys. Specifically, Rocheva and her colleagues analyze the use of advertisements in two of Russia's leading social networking sites (SNS)—Vkontakte and Odnoklassniki—to capture participants for various Internet surveys of migrant populations in Russia. By doing so, the authors contribute to the growing body of literature that is investigating the use of alternative recruitment procedures when reliable sampling frames are unavailable or not feasible. Previous work on the SNS-based sampling of hard-to-reach populations has dealt mostly with Western European countries, the USA, and Australia, thus the

extant literature refers mostly to Facebook, given its predominance in these markets. However, as the authors stress, a disproportionate focus on one particular SNS has serious limitations, since Facebook is not available in all countries, and is not necessarily the most-used social network in the countries in which it can be accessed. Furthermore, since each SNS employs its own algorithms, targeting options, and general procedures, the knowledge obtained about one such platform cannot be simply extrapolated to others. Therefore, although these authors' conclusions largely confirm those of previous research, their detailed examination of Vkontakte and Odnoklassniki constitutes a significant addition to the extant literature. On the upside, SNS-based sampling has been found to enable researchers to investigate highly dispersed populations within a short time frame. On the downside, however, such procedures generate non-probability samples, with the added limitation of uncertainty about the algorithm parameters that underpin the target selection of SNS-based advertisements. In Russia as elsewhere, these algorithm parameters are anxiously guarded as proprietary information by platform owners, and may change without researchers' knowledge. The ensuing combination of selection and self-selection biases of unknown proportions suggests that, as long as those conditions persist, the results of SNS-based sampling have to be considered with caution.

A similar note of caution is raised in the chapter *Web-Based Respondent-Driven Sampling in Research on Multiple Migrants: Challenges and Opportunities* by Ágata Górny and Justyna Salamońska, which explores how web-based respondent-driven sampling (web-based RDS) could be used to recruit multiple migrants into a web survey. Just like the authors of the previous chapter, they address non-probability sampling for online surveys, yet their context shifts from a heterogeneous and broadly defined target population residing in one specific country (as explored in Chap. 3) to a narrowly defined target group—Polish migrants who have resided in several foreign countries—scattered across a potentially large number of places, a situation that accentuates the "hard-to-identify" component of the manifold complications that typically make migrants a hard-to-survey population. Conceptually, the RDS approach resolves this difficulty by asking respondents to recruit as additional study participants all those among their friends and acquaintances who meet the target group definition. In combination with a diversified pool of first-round interviewees ("seeds"), this rule of recruiting *all* eligible contacts is expected to give RDS an edge, in terms of representativeness, over other non-probability sampling strategies. However, in this particular case, it turned out that the target definition (multiple migration experience) was not salient enough to generate extensive recruitment chains; therefore, first-round seeds accounted for the majority of participants. Accordingly, one of the conclusions of the chapter is the need for researchers to verify, prior to a study's launch, that the target population is defined in terms of a salient self-definition as a social group. Second, whereas traditional face-to-face RDS interviewers can explain to their respondents the importance of recruiting additional participants, this crucial step can be exceedingly challenging regarding web-based RDS. Third, again with a view to increasing the likelihood of referral to additional interviewees, the authors highlight the importance of keeping the questionnaire short and engaging. Fourth, the authors point out that

the management of incentives, a vital ingredient of the RDS methodology, also needs especially careful thought and preparation with respect to an online-only research setting in which staff and participants are literally scattered across the world. To recapitulate, this study will be extremely helpful to migration scholars interested in web-based RDS.

In *Computer-Assisted Migration Research: What We Can Learn about Source Questionnaire Design and Translation from the Software Localization Field*, Dorothée Behr provides a fascinating example of knowledge transfer. Diverging from the well-trodden path of scientists' insights being applied to other (political, economic, commercial, etc.) realms, on this occasion academics are at that transfer's receiving end. Behr draws on the know-how of multinational technology companies to outline the manifold steps, sophisticated workflows, and multi-professional expertise required for the seamless implementation of pluri-linguistic questionnaires in computerized surveys. Specifically, she details the procedures used by the software localization industry to ensure that the vast and continuously changing range of technology products are equipped with customized versions of instructions and user interfaces that consumers anywhere on the globe may require. Behr's examination of the complexity and resource requirements of this blueprint sends a sobering message to the oftentimes atomized and underfunded community of migration scholars: state-of-the-art multilingual questionnaires for digital surveys require painstaking forward planning and a seamless cooperation of many distinct professionals, and thus, they require top-drawer organizational capabilities and a substantial budget. Small-scale surveys entailing only two or three languages may still be manageable with more artisanal means, but Behr's chapter illustrates how, in migration studies and other cross-national survey operations, the research landscape is evolving towards increasingly large and complex management structures. Although only unusually well-resourced projects can hope to emulate the procedures that Behr outlines, her study should appeal to a much broader audience, since it exemplifies how cross-cultural adaptation, rather than the niche concern of specialized scholars, is quite literally mainstream business.

The second of our chapters on linguistic matters—*Surveying Illiterate Individuals: Are Audio Files in Computer-Assisted Self-Interviews a Useful Supportive Tool?* by Florian Heinritz, Gisela Will, and Raffaela Gentile—provides methodological reflections on a research design that had been optimized on substantive grounds. Thus, rather than employing distinct methodological options alternatively in an experimental setting, several such options were combined in the fieldwork. Despite the ensuing limitations of this approach, the chapter contains interesting observations on the tools that can be used with an especially hard-to-survey population—international migrants who lack reading skills in their native tongue. Heinritz and colleagues addressed this challenge by preparing audio recordings for all the questionnaire items in a range of languages; in addition, they translated the questionnaire into these languages and deployed fieldwork staff competent in these languages. Initially, the audio recordings were meant to be used by respondents who preferred to administer part of the (computer-assisted) questionnaire themselves, with a view to preserving the anonymity of chosen

responses and thus preventing response bias due to social desirability concerns. However, since the interviewers were present at all times throughout all the interviews, the distinction between computer-assisted self-interview (CASI) and computer-assisted personal interview (CAPI) was blurred in practice, and the intended safeguard against social desirability bias became largely elusive. In addition to illustrating the need for researchers to carefully envision and pretest the whole fieldwork process to detect unanticipated glitches, this study cautions that the incorporation of technologically advanced features does not automatically guarantee enhanced data quality.

11.2 A New Dimension: Leveraging "Found" Data for Migration Research

The four contributions to the second part of the collection address options and tools for accessing and using *found data*, i.e., data that were either collected actively by third parties or generated passively—without a prior research design or specified scientific purpose—by digital sensors or devices.

Sebastian Rinken's and José Luis Ortega's *Leveraging the Web for Migration Studies: Data Sources and Data Extraction* provides an introduction to the second part of the collection. They explore the implications of the "data revolution" for migration research, i.e., the availability of ever-increasing amounts of mostly unstructured data through the Internet. Rinken and Ortega argue that such new data sources are particularly useful for migration studies, given the limitations of traditional research techniques and data sources. In addition to highlighting the wealth of third-party surveys and administrative datasets accessible on the Internet via a range of generalist data portals, specialized sites, data repositories, and search engines, the main contribution of this chapter is its discussion of some of the techniques that enable researchers to extract non-structured data from the Internet. Rather than a hands-on crash course, this introduction to *web-scraping* as a data collection method aims to alert migration researchers of the need to broaden their skill set, both as individuals and as cross-disciplinary teams. In Rinken's and Ortega's view, unstructured data could offer extraordinary opportunities for gaining insights into migration flows, integration patterns, and migration-related attitudes, to name three areas of outstanding relevance. As they strive to advertise this potential, the authors may at times strike an overenthusiastic note in that important issues, such as data protection, privacy considerations, and data quality, are hinted at rather passingly and certainly warrant more sustained consideration. Thus, in the context of the overall structure of the collection, this chapter serves as an appetizer inviting migration scholars to actively explore new data types and their ensuing research options. Although Rinken and Ortega do not suggest that traditional types of data will disappear, they anticipate that their added value, when compared to "found" data, will become less and less obvious as digitization affects an ever-growing share

of more and more people's daily lives. Thus, their advice to migration scholars is not to sit on the fence, but rather enter the fray!

The remaining three chapters in the book's second part exemplify, each in its own way, how migration scholars can leverage new data sources for their research objectives. In *How Canada's Data Ecosystem Offers Insights on the Options for Studying Migration in an Unprecedented Era of Information*, Howard Ramos and Michael Haan focus on the use of administrative records as resources for scientific inquiry, paying special attention to the interconnections between different datasets. Drawing mainly on their intimate knowledge of the Canadian data environment, Ramos and Haan highlight innovative approaches that facilitate the utilization of different administrative data sources and their linkage for research purposes. While attesting to the value that such data hold for migration studies, the authors identify a number of challenges that need to be addressed to fully harness their potential. A first issue concerns the diversity of definitions and measurement options employed by distinct data providers. With respect to this concern, the authors insist on the need for scientists and practitioners to develop standardized definitions and instruments. This solution seems quite ambitious even at a national level, not to mention a cross-national perspective, considering the breadth of statistical operations involved and the diversity of the specific goals pursued by distinct data providers and data collection operations. A second major challenge concerns data access and curation. Even in Canada, where administrative records are increasingly available to researchers, access usually requires physical presence at a specific institution. This rather anachronistic prerequisite poses added difficulties especially for scholars based in other countries. As for curation, making administrative records available to the scientific community requires a considerable additional workload, and thus investment in qualified personnel by the data providers. Finally, Ramos and Haan also draw attention to the fact that the secondary use of administrative sources raises important data security and data protection issues. Notwithstanding these challenges, they urge researchers and data producers to cooperate nationally and internationally with a view to creating the data infrastructures and data handling protocols necessary for making administrative data more readily available for scientific analysis.

In *Assessing Transnational Human Mobility on a Global Scale*, Emanuel Deutschmann, Ettore Recchi, and Michele Vespe switch the focus to an even more ample notion of *found data*, the truly huge volumes of data that are gathered for purely operational reasons, without any inherent relation to researchers' conceptual definitions and needs. Deutschmann and his colleagues combine datasets on world-wide tourism and air passenger traffic to generate a plausible estimate of a target variable that is not provided by any of those two sources, namely, *cross-border mobility*. Since they manage to uncover something that was "visibly absent" in those sources, their dataset merging procedures seem to be touched by magic. Put more prosaically, the authors depart from a painstakingly crisp description of the content covered by their two baseline sources. Next, they analyze the relation of these found data with their information needs. By detecting the overlapping kinds of information and defining rules of transformation for specific data categories, they are able to outline a pathway toward a merged dataset. While the exact procedures and steps are

obviously specific to each study, such a formalized approach (that translates neatly to mathematical formulae) is likely to be appropriate whenever scholars face analogous challenges. Thus, the contribution by Deutschmann and his colleagues is an excellent example of how extant data collections can be repurposed for research needs. By the same token, this chapter showcases the virtues of datasets that cover entire populations—in this case, all of the world's cross-border overnight stays and air travel.

Finally, in *Google Trends as a Tool for Public Opinion Research: An Illustration of the Perceived Threats of Immigration*, Reilly Lorenz, Jacob Beck, Sophie Horneber, Florian Keusch, and Christopher Antoun provide a stimulating example of how scientific inquiry can benefit from data that are readily available on the Internet. Lorenz and her colleagues demonstrate that the usefulness of search engine data (specifically, Google's Trends feature) depends not only on the careful selection of search terms, time periods, and territorial references, but also on recognizing inherent limitations. Google Trends provides information on relative frequencies regarding chosen search terms and reference periods, but not on user profiles or absolute frequencies, although estimates of these absolute frequencies are available to the paying customers of Google Ads. Rather than directly revealing specific behavioral or attitude patterns, such data speak to the relative salience of search terms. Side-stepping these limitations, the authors' strategy of external validation detects a lagged aggregate correlation of negatively worded search queries on perceived immigration threats with voting intentions for a virulent anti-immigrant party. Their study shows not only what kinds of threat perceptions are associated most closely with the electoral fortunes of right-wing populism, but also that immigration-related concerns translate into anti-immigrant voting preferences with several months of delay (at least in the particular case examined here—voting intentions for the "Alternative für Deutschland" party from 2013 through 2019). Since it is difficult to see how these findings could have been obtained with traditional survey instruments, they highlight the virtues of Internet-based research tools (granularity, customizability, timeliness, etc.). Also, while little knowledge exists so far on the potential inhibitions of choosing query terms, search engine use seems less prone to social desirability bias than surveys, even those self-administered online. However, this chapter also suggests that, with a view to detecting individual-level covariates or determinants, surveys continue to be an important part of researchers' toolkits.

11.3 The New Frontier: Distilling Knowledge from Accrued Data

To point out the obvious, the papers collected in this volume do not constitute a representative sample of migration scholars' use of innovative technology. Even if that claim had been plausible at the book's inception, its veracity would have

diminished inevitably by the time of its publication. Since the opportunities afforded to researchers by emerging technologies and the uses thus enabled by these options are constantly evolving, any conclusions related to the subject-matter of this book are necessarily tentative. Such caution regarding the adoption of new technologies applies to any research domain, yet it seems especially appropriate with regard to migration studies. For example, in comparison to the vibrant exchange between survey researchers and computational social scientists at the biannual Big Data Meets Survey Sciences (BigSurv) conferences (see Hill et al., 2020 for a collection of papers from the first such event), the migration research community appears to be relatively slow at taking up new technological options, especially with regard to mining the web for actual insights. Although we can only speculate about the reasons, it seems plausible to assume that data access issues, on the one hand, and insufficient data management skills, on the other, contribute decisively to this situation. Many migration researchers would probably argue that the complex nature of their field of study requires data of a more qualitative kind than those traced, in one way or another, on the Web. However, in our view, this line of argument underestimates the huge potential of next-generation multi-method approaches.

Admittedly, a comparison to the avant-garde of interdisciplinary cross-fertilization between data scientists and survey researchers is somewhat unfair to migration scholars. However, setting the bar high seems appropriate to raising the game, so this book has aimed at spurring some added diligence (for a complementary effort see Salah et al., forthcoming). Initiatives such as the Big Data for Migration Alliance (https://data4migration.org) and the HumMingBird project (https://hummingbird-h2020.eu) suggest that the migration-research landscape could change particularly fast in the coming years due to a combination of a relatively low "market share" of ICTs and other innovative technologies (as gauged by the papers presented at IMISCOE's annual conferences, for example), on one hand, and the added mileage that can be obtained potentially by their adoption, on the other. This book does not aim to provide a step-by-step guide to competence-building, but rather offers an incentive for migration researchers to constantly assess what kind of empirical evidence is most appropriate for achieving their goals, and how to obtain such data.

The collection's nine chapters (not including this closing chapter and the introduction) exemplify, or showcase, two main ways that innovative technology may contribute to enhancing the methodological arsenal of migration studies. The five contributions gathered in the book's first part explore how ICT and other emerging technologies help to improve the viability and quality of conceptually traditional studies, i.e., researcher-defined data collections pursuing extrapolation with respect to oftentimes very specific target populations. Despite their intrinsic difficulties, inescapable limitations, and considerable cost, such sample-based studies continue to be held in high esteem by academics and migration-managing institutions alike. This is true especially in times of intensifying migration flows. However, even in those countries with excellent systems of public statistics, information on newcomers' characteristics, needs, and skills cannot be delivered adequately by extant administrative sources or general-population surveys, nor can qualitative studies

alone provide the input needed for planning and implementing the services and procedures required in such circumstances. More generally, academics, practitioners, and evidence-focused policymakers cherish the possibility of converting survey items into predictors of key outcome variables. With due respect for other sources and approaches, surveys of migrant populations may therefore be seen to have been the linchpin or "frontier" of migration studies in recent decades. Although some of the techno-methodological options described in this book's first part (e.g., the use of social network sites for respondent recruitment) are of interest for qualitative data collections as well, their main field of application is the improvement of migrant surveys.

Shifting gears, the four contributions to the collection's second part address researchers' use of data that, instead of relying on samples, cover entire populations (at least in principle), and instead of deriving from specified research designs, were produced for administrative or purely operative reasons. In the book's second part, such data come into focus not with respect to sample design and implementation, but in terms of the actual clues they provide about people's behaviors and mindsets.

Even though social scientists have extracted information from administrative records and censuses for decades, the Internet now enables access to a previously unimaginable wealth of such sources. This situation opens up exciting new research options, especially when various datasets can be linked, and provided that privacy can be protected and informed consent achieved across an enlarged user base—two big "ifs" that require careful attention (Ramos and Haan, in this volume). Even in light of these constraints, administrative records represent the more accessible part of the new data universe, since they have been generated in purposeful ways and are conceptually rather similar to researcher-defined data collections (Connelly et al., 2016). This affinity is illustrated by the fact that a large share of the contributions to Hill et al. (2020) refer to the combination of survey data with administrative datasets.

In contrast, from a researcher's point of view, most types of Internet data are messy and oftentimes arcane. Different from censuses, which collect the same (relatively sparce) information from a given territory's residents at long intervals, and different from administrative records, which refer to well-defined (yet typically rather isolated) events or procedures, Internet data transcend traditional time-space restrictions and accustomed notions of what counts as data in the first place. In the new data universe, space and time cease to be constitutive features of a study's data collection plan, since they become customizable parameters for data extraction. By the same token, the traditional business of operationalizing variables of interest gives way to the perhaps even more challenging task of *repurposing* extant information, an endeavor that presupposes a capacity to identify relevant bits of information among overwhelming quantities of non-sense. Also, once a pattern is recognized, it requires interpretation, yet the dataset may (and most likely will) lack information concerning *explanatory* variables.

In short, *big data* is a categorically different kind of input into the research process – with far-reaching consequences. The specialized literature offers a range of descriptive adjectives, including *organic* (Groves, 2011), *found* (Connelly et al., 2016) and *non-sampled* (Hill, 2020). We would like to suggest the term *accrued*

data as an addition to this semantic cluster to highlight their status as mere derivatives of behaviors or even, increasingly, interconnected digital gadgets. Whatever the label, it seems vital to flag the categorical difference separating purposefully designed research data from data that lack intrinsic meaning, relevance, and even intelligibility.

The frontier has moved.

In our view, the leap from more or less circumscribed, conceptually-driven data collections to unbounded, operationally-driven data generation entails differences at least as momentous as those between the two broad categories of research designs (qualitative *versus* quantitative) that have co-existed, competed, and complemented one another in the social sciences throughout the past century or so. Of course, both survey-based and internet-based data require (and support) computer-assisted quantitative analyses. However, the accustomed statistical techniques are inappropriate for handling the enormous, and exponentially growing, volume of data that are by-products of an ever-expanding range of everyday activities (Spiegelhalter, 2014). As the role of automated computation increases, so does the incidence of spurious correlations. In the emerging world of data-driven social sciences, a crucial challenge is to derive *meaningful* knowledge from a deluge of information that undermines acquainted notions of pertinence and transparency (Kitchin, 2014).

The abundance of real-time data at zero or very low marginal cost risks obscuring, rather than highlighting, significant patterns. The sheer volume of information may blind researchers to its skewed nature. For various reasons, it would be misleading to understand these new data as mere reflections of the world as is (Vinck et al., 2019). Digital devices, software applications, and algorithms are not only conceived for specific purposes and subject to coverage, malfunctioning, and/or misreporting biases of oftentimes unknown magnitude, but also are designed and written by human beings and therefore, directly or indirectly, are influenced by their creators' socializations and cultural backgrounds. Furthermore, the predominance of private-sector data generates access hurdles and secrecy regarding fundamental definitions and procedures, as is the case with proprietary algorithms. In short, the opportunities afforded by big data's volume, velocity, and variety (cf. Laney, 2001) also entail pressing concerns regarding validity and veracity (McCoach et al., 2020).

The skills honed with a view to collecting and analyzing researcher-defined datasets are anything but obsolete in this context; instead the opposite is the case. As Hill et al. (2020) have stressed, initial excitement about big data's competitive advantages in terms of scope, timeliness, and cost has given way to the realization that many of its most notorious challenges can be reframed in terms such as coverage bias, imputation error, or total error (Biemer & Amaya, 2020; Sen et al., 2021). With respect to the interdisciplinary cooperation envisioned by Hill et al. (2020), data science requires survey researchers' input especially with regard to conceptualization, context, and data quality. Also, domain expertise of a more qualitative nature will continue to be indispensable for interpreting any patterns observed in organic data (Salah et al., 2019). These considerations suggest that survey research (and other "low-volume" data types and techniques) will continue to play a relevant role in the methodological portfolio of the social sciences in general and migration

research in particular. However, we anticipate that their contributions will be defined increasingly by next-generation "mixed methods" approaches. In a research landscape increasingly shaped by creative combinations of sampled and non-sampled data (Groves, 2011) and highly interdisciplinary teams (King, 2014), administrative data will be used to ease the burden for respondents, just as surveys will be employed to add depth and context to the behavioral and attitudinal patterns revealed by digital traces. Historical retrospect supports this prediction: throughout the second half of the twentieth century, rather than being supplanted by the increasing sophistication of surveys, qualitative studies continued to play their part. Just as in-depth interviews or focus-groups can offer more fine-grained and contextualized insights into motivations than surveys, so can surveys provide better insights into motivations when compared to organic data.

Research ethics is a second area in which accrued data represent a momentous departure from established procedures. With regard to surveys, administrative datasets, and censuses, ethical considerations have long been guided by the principles of privacy protection and informed consent. Since migrants are oftentimes vulnerable to abuse and discrimination, rigorous procedures are required to prevent any possibility of privacy breaches. However, in the brave new world of data mining, these principles are both partially unviable and patently insufficient, at least as traditionally conceived. As Deutschmann et al. and Lorenz et al. illustrate in their contributions to this collection, *a posteriori* repurposing of data may not be covered by explicit authorizations from their original sources (in these two specific cases, international travelers and search engine users), and such consent may not be necessarily pertinent with respect to huge aggregate datasets that do not contain any *personal* information. In other contexts, especially when user-created content is at the center of analysis, reflection may be needed on the exact nature of any given consent. More specifically, a question arises as to whether approaches that are technically covered by services' terms of use nevertheless need additional and specific consent by individuals when they become participants in, or objects of, research (Leurs & Prabhakar, 2018; Vinck et al., 2019). However, the need for ethical conduct transcends the personal sphere, for example, automated big-data analysis may reveal the real-time location of vulnerable groups. The new data universe requires an anticipatory incorporation of ethical considerations into computing procedures ("ethics by algorithm", cf. Dignum, 2018). The emerging consensus in the debate on big-data ethics demands complementing the range of principles that govern bioethics (beneficence, non-maleficence, autonomy, and justice) with an added imperative of explicability: the procedures of artificial intelligence must be made intelligible to lay citizens, and accountability for any misguided outcomes must be assured (Floridi et al., 2018). In more general terms, the use of accrued data, especially, yet not exclusively, in research on migrants and other minority groups, necessitates that scholars reflect on the potentially unintended consequences of their research. Upholding the "do no harm" principle must always take precedence over scientific curiosity and the allure of testing innovative procedures of knowledge production. Thus, on the one hand, scholars need to contemplate the consequences of publishing specific findings, especially on vulnerable

subpopulations such as irregular migrants. On the other hand, researchers need to consider that political actors may employ newly developed methods to carry out their own agendas (Franklinos et al., 2020; Vinck et al., 2019). Another challenge is to make the data accrued by private corporations accessible to scientific repurposing. The forthcoming regulatory battle at this regard presupposes the public's trust in appropriate ethical safeguards.

To resume, we envision a future of multi-disciplinary collaboration where migration specialists with various disciplinary backgrounds, survey methodologists, and data scientists upgrade and transfer skills with one another. If pushed to single out one take-home message from this book's contributions, we advise migration researchers to actively participate in the process of multi-disciplinary skill building in terms of project design, networking, and human resources development (training and staffing). Without wishing to scare people into action, the history of technology abounds in examples of doom spelled by the failure to adapt to disruptive technology. Without any hype, it seems safe to rate ubiquitous digitization among the most disruptive waves of technological innovation ever.

That disruption continues relentlessly as Internet-wary cohorts diminish and ever-increasing shares of daily behaviors become subject to digitized scrutiny. In time, sample-based data collections might come to be seen as anachronistic. Remember those chemistry-based photographs?

References

Biemer, P. P., & Amaya, A. (2020). Total error frameworks for found data. In C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, & L. E. Lyberg (Eds.), *Big data meets survey science* (pp. 131–161). Wiley. https://doi.org/10.1002/9781118976357.ch4

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research, 59*, 1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015

Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology, 20*(1), 1–3. https://doi.org/10.1007/s10676-018-9450-z

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Franklinos, L., Parrish, R., Burns, R., Caflisch, A., Mallick, B., Rahman, T., Routsis, V., Sebastián López, A., Tatem, A., & Trigwell, R. (2020). *Key opportunities and challenges for the use of big data in migration research and policy* [Preprint]. https://doi.org/10.14324/111.444/000042.v1

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly, 75*(5), 861–871. https://doi.org/10.1093/poq/nfr057

Hill, C. A. (2020). Moving social science into the fourth paradigm. In C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, & L. E. Lyberg (Eds.), *Big data meets survey science* (pp. 713–731). Wiley. https://doi.org/10.1002/9781118976357.ch24

Hill, C. A., Biemer, P. P., Buskirk, T. D., Japec, L., Kirchner, A., Kolenikov, S., & Lyberg, L. E. (Eds.). (2020). *Big data meets survey science. A collection of innovative methods* (1st ed.). Wiley. https://doi.org/10.1002/9781118976357

King, G. (2014). Restructuring the social sciences: Reflections from Harvard's Institute for Quantitative Social Science. *PS: Political Science and Politics, 47*(1), 165–172. https://doi.org/10.1017/S1049096513001534

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), 2053951714528481. https://doi.org/10.1177/2053951714528481

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity, 47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9

Laney, D. (2001). 3-D data management: Controlling data volume, velocity and variety. *META Group Research Note, 6*(70), 1.

Leurs, K., & Prabhakar, M. (2018). Doing digital migration studies: Methodological considerations for an emerging research focus. In R. Zapata-Barrero & E. Yalaz (Eds.), *Qualitative research in European migration studies* (pp. 247–266). Springer. https://doi.org/10.1007/978-3-319-76861-8_14

Majó, J. (2012). Evolución de las tecnologías de la comunicación. In M. de Moragas (Ed.), *La comunicación: De los orígenes a internet* (pp. 65–89). Gedisa.

McCoach, D. B., Dineen, J. N., Chafouleas, S. M., & Briesch, A. (2020). Reproducibility in the era of big data: Lessons for developing robust data management and data analysis procedures. In C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, & L. E. Lyberg (Eds.), *Big data meets survey science* (1st ed., pp. 625–655). Wiley. https://doi.org/10.1002/9781118976357.ch21

Salah, A. A., Korkmaz, E. E., & Bircan, T. (forthcoming). *Data science for migration and mobility*. Oxford University Press.

Salah, A. A., Pentland, A., Lepri, B., & Letouzé, E. (2019). *Guide to mobile data analytics in refugee scenarios. The "data for refugees challenge" study*. Springer. https://doi.org/10.1007/978-3-030-12554-7

Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A Total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly, 85*(S1), 399–422. https://doi.org/10.1093/poq/nfab018

Spiegelhalter, D. J. (2014). The future lies in uncertainty. *Science, 345*(6194), 264–265. https://doi.org/10.1126/science.1251122

Tourangeau, R., Edwards, B., Johnson, T. P., Wolter, K. M., & Bates, N. (2014). *Hard-to-survey populations*. Cambridge University Press.

Vinck, P., Pham, P. N., & Salah, A. A. (2019). "Do no harm" in the age of big data: Data, ethics, and the refugees. In A. A. Salah, A. Pentland, B. Lepri, & E. Letouzé (Eds.), *Guide to mobile data analytics in refugee scenarios. The "data for refugees challenge" study* (pp. 87–99). Springer. https://doi.org/10.1007/978-3-030-12554-7_5