New data sources and computational approaches on migration and human mobility

Albert Ali Salah, Tuba Bircan, Emre Eren Korkmaz

Abstract

Human migration and mobility is one of the most important societal issues of our time, with wide-ranging implications, and for which it is difficult to obtain timely insights. Policy makers require multi-faceted and accurate insights for implementing actions to ensure the well-being of populations. Recent advances in sensor technologies and data scientific tools allow observation of humans and their environment at great detail, and at great granularity. New data sources, such as mobile phones, digital applications, social media, and satellites are used for creating timely and detailed insights into human mobility, which is especially important for humanitarian crises that accompany some of these movements. However, there are also risks of surveillance at an unprecedented scale, loss of privacy, and unexpected changes to power structures. In this work, we provide an overview of the potential benefits and risks, as well as challenges of using big data analysis for migration and mobility.

Keywords: Big data analysis, migration, human behaviour analysis, quantitative research, data science, mobility analysis, computational social science

1 Introduction

Scholars from different disciplines have been seeking to understand the societal dynamics in migration[1] and mobility, which concerns itself with long-term human movements within and between countries, as well as ways to estimate and predict such movements.[2]

While qualitative research methods have been instrumental in gaining insights into the factors influencing human movements, quantitative approaches were essential for tasks like prediction and estimation. The digital era ushered in by extremely widespread use of digital tools, such as mobile phones and social media, brought new possibilities for quantitative analysis. Especially with the onset of the Covid-19 pandemic, the potential value of new data sources -such

[1] We provide definitions of some key terms at the end of the chapter.

[2] This is the uncorrected author proof. Copyright with British Academy. Please cite as: Salah, A.A., T. Bircan, E.E. Korkmaz, "New data sources and computational approaches on migration and human mobility," in Salah, A.A., E.E. Korkmaz, T. Bircan (eds.) Data Science for Migration and Mobility, British Academy / Oxford University Press, 2022.

as mobile phones- was recognised in capturing the statics and dynamics of large scale human movements (Oliver et al. 2020).

In this work, we provide an overview of the potential benefits and risks, as well as challenges of using big data analysis for migration and mobility. Our aim is to illustrate, with examples, how approaches based on large scale data analysis can be used to gain insights into the multi-layered and complex issues surrounding this area. Our take-away lesson is that such methods are not wholesale solutions providing definitive answers, but have different strengths and weaknesses compared to traditional methods, and as such, can be complementary to them.

This work is organised as follows. Section 2 is devoted to the conceptual groundwork. In Section 3, we discuss the potential benefits of new data sources, and give examples from the literature. Section 4 is about the stakeholders, and how data can be shared among these stakeholders. We do not discuss ethical, legal, and privacy related risks associated with the collection and usage of big data solutions in this chapter, but we provide a case study to illustrate one of risks more in depth in Section 5. We conclude in Section 6.

2 Conceptual groundwork

2.1 Migration and mobility

Although the migration process refers to relocating across an (inter)national boundary for a period of time, the precise operationalisation of this concept in practice varies amongst countries, institutes and disciplines. We are all 'mobile individuals' living in a world that has been structured by mobility and people move for various purposes and lengths of time. The forms of mobility and migration are diversifying globally (e.g. Favell (2011), Skeldon (2017), Vertovec (2007)). Hence, when focusing on different aspects of migration, it is important to distinguish different forms of movement and migration. Migratory movement can be: within or across borders; voluntary (for work, study or family reasons) or forced (as a result of conflict or natural disasters); regular (with documentation) or irregular (without documentation); and temporary, seasonal or longer term/permanent. Definitions can change during the process. For instance, if a non-native student (a voluntary, regular migrant) overstays her/his allowance for the academic year s/he was registered for, s/he may become an undocumented migrant. This will also lead to a gap in keeping the records and following up the migratory movements and status changes.

As a first step in studying migratory movements with new methodologies, the terminology of migration and mobility needs to be set clear especially for the scholars working on these topics and not having a social sciences background. In order to avoid any confusion between mobility and migration, we can start with distinguishing two terms. Every migratory movement is 'mobility' but the other direction is not true, every mobility movement cannot be counted as 'migration'. Therefore, for traditional migration statistics, the term 'mobility'

is not commonly used.

2.2 The data perspective

Computational social science is defined as "the development and application of computational methods to complex, typically large-scale, human (sometimes simulated) behavioural data" (Lazer et al. 2009). It is a research area that fosters (or rather, demands) collaborations between data and computer scientists, statistical physicists, and with many disciplines whose inquiries extend into issues touching on human behaviour, such as epidemiology, economics, policy, governance, public health, and naturally, migration. The most important aspect of these collaborations is the sharing (and subsequent analysis) of large-scale human behavioural data, which are gathered by diverse stakeholders for diverse purposes, but can be re-purposed for addressing specific research questions (Lazer et al. 2020).

In the context of migration and mobility, what we call 'new data sources' are the ones that are not widely adopted for studying questions in this area. To give a specific example, satellite remote sensing systems have been used frequently to study bird migrations, which are more regular and 'visible' than migrating humans, but such data are extremely useful in sensing environmental conditions that can be used to gain insights into economic factors, population estimates, environmental conditions, and proxy indicators of wealth (Bircan 2021). Similarly, mobile phone and social media usage leaves digital traces that can be very informative on different aspects of migration (Coimbra et al. 2021, Kim et al. 2021, Luca et al. 2021, Sterly & Wirkus 2021).

The interest from computational social scientists, computer scientists and data scientists for the utilisation of alternative data sources for varied societal challenges led to an increase in studies addressing human migration and mobility. This is also related with the newly burgeoning field of 'Data Science and Artificial Intelligence for Social Good,' which typically adapts United Nations' Sustainable Development Goals (SDGs) as a framework for addressing problems with socially beneficial outcomes (Tomašev et al. 2020, Cowls et al. 2021). From the data science and methodological perspectives, these studies consider 'migration' more of an application case, rather than contributing to the existing theoretical frameworks. Acknowledging the benefits of these innovative techniques, there are still numerous challenges to be tackled for paving analytical and sustainable paths to nurture interdisciplinary studies and to strengthen the new computational approaches to long lasting questions about migration.

Data science is a very broad term and encompasses all methods used to turn data into knowledge and insights. Machine learning, on the other hand, is about creating models that optimise certain functions over such data. Some of these models are concerned with prediction, in which case the data that are consumed by these models are treated as input, and the predictions as the output. For example, the number of immigrants and emigrants for a specific country can be predicted based on past data (Robinson & Dilkina 2018). However, this problem is very challenging, as there are many factors that play a role in migration,

and their interaction is complex. Other machine learning approaches are concerned with finding patterns in data (such as clusters and groups that exhibit differences), or with modelling spatio-temporal dynamics. A type of machine learning algorithms, called ' deep learning,' have received enormous interest in recent years, but these are typically mathematical models with millions of free parameters, and their success depends on very large datasets used for training (i.e. setting the parameters) of these models (Alpaydin 2016). What made machine learning and data science very interesting for migration and mobility is the availability of the new data sources we have mentioned, as they allow the development of complex mathematical models.

Big data are extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions. As Hilbert (2013) states, big data differs from data based on traditional household surveys as they do not refer to a random sample of individuals but to the totality of the population using, for instance, mobile phones or internet- based platforms, and these data are accessible in near real time. Big data are not only 'big' because of their volume; the speed ('velocity') at which they are generated and the complexity ('variety') of the information are also considered as distinguishing features of this kind of data (Hilbert 2013). In relation to the migration research, the term 'big data' refers to anonymised data that are generated by users of mobile phones, internet-based platforms (i.e., social media) and applications, or by digital sensors and meters such as satellite imagery. Big data analysis requires specific technical and analytical methods required to extract meaningful insights from the big data and transform these data into 'value' (De Mauro et al. 2016).

2.3 Approaches to empirical modelling

At this point, it is useful to stress some important distinctions in the way computational and social sciences approach their empirical problems. Since computer modelling is more limited in handling semantic information, compared to structured and numeric information, it is not a surprise that it supports quantitative approaches better than qualitative approaches. But what is more important is that in computational sciences, there is a heavier emphasis on predictive models, as opposed to explanation based models in social sciences (Hofman et al. 2021). One of the promises of computational social science is the possibility of creating models with both predictive and explanatory powers.

If we apply the conceptual model proposed by Hofman et al. (2021) to empirical modelling in migration and mobility, we get two main distinctions along which to look at research. The focus of the study can be on specific features and effects, to explain the factors that caused a certain phenomenon under study, or conversely, it can be about predicting outcomes, to try to see what would happen if some of these factors are changed, for example through policy decisions. The second dimension proposed by Hofman et al. (2021) is whether there are any interventions or distributional changes for the described situation. Table 1 summarises the four possibilities under this conceptual model. To exemplify the

	No intervention or distributional changes	**Under interventions or distributional changes**
Focus on specific features or effects	Descriptive modelling	Explanatory modelling
Focus on predicting outcomes	Predictive modelling	Integrative modelling

Table 1: Organizing empirical modelling along two dimensions proposed by Hofman et al. (2021), simplified from the original.

possibilities of new data sources and computational approaches in migration and mobility, we provide different examples, falling under these four quadrants.

Descriptive modelling is about describing the features or effects for a certain situation. Consider the question of integration and assimilation of migrants, which is typically studied via surveys. This limits the number of people from which data are obtained, but allows detailed insights for some of the factors. For example, the EUCROSS survey[3] that investigated the Europeanisation of immigrants included telephone interviews with 1000 subjects from different nationalities (Pötzschke 2015). Contrast this with the study of Marquez et al. (2019), where mobile phone data traces of over 1 million users (including over 200K Syrians in Turkey) were used together with 65K Tweets collected from Twitter to measure integration and sentiments against refugees in Turkey. The use of such large data sources allowed the researchers to analyse integration and segregation in great spatio-temporal detail, even though factors such as demographics were not visible due to the anonymisation of the data.

Explanatory modelling is for estimating the effects of changing some of the factors that affect a situation. This goes beyond descriptive modelling, in that there is an effort to explain the causes of changes in the observed variables. Many mathematical models in empirical sociology and economics fall under this quadrant, including the gravity and radiation models in migration analysis (Zipf 1946, Simini et al. 2012). As an example of using new data sources, we can consider the application of computational linguistic and text analysis to media content about migration (Allen 2021, Koch et al. 2020). Koch et al. (2020) study the empirical relationship between public debates in the media and asylum acceptance rates in Europe from 2002–2016 by using a large news repository called GDELT[4], which contains about 2.5 terabytes of media material per year. The advantages of using GDELT over traditional data sources are the large coverage and the possibility of having a rich representation for public debate as a factor.

Predictive modelling, the third quadrant in Table 1, is about predicting the outcomes, without paying much attention to the specific factors that play a role in the outcomes. This is a very typical setting for machine learning scenarios, where a complex model is trained for achieving supervised classification (i.e.

[3] `https://cordis.europa.eu/project/id/266767`
[4] `https://www.gdeltproject.org/`

the model learns to predict output labels from a set of input features, based on a training set of pre-labeled data points). The model can achieve excellent accuracy, which is the main criterion of success, but it may not give any usable insights about the input factors, especially if there are many such factors, and each contribute a little to the outcome, as well as having complex relationships with other factors. Bircan (2021) discusses the use of remote sensing data for migration and mobility research. In (Dietler et al. 2020), researchers develop a machine learning based prediction model for quantifying settlement growth in rural communities in Burkina Faso, using such remote sensing data. Since other data sources were lacking, a multi-annual training dataset was created using historic Google Earth imagery and good prediction results were obtained.

The main problem with such predictive modelling approaches is that the machine learning model only learns the variation in its training set, and has limited generalisation capabilities. When there is an unexpected change of conditions and dynamics, the models cannot handle these automatically. Instead, the experts need to intervene, and re-train the models by taking relevant factors into account. To illustrate the difficulty of fully automatic incorporation of the relevant factors in machine learning models, consider the example given by Earney & Jimenez (2019) about UNHCR's Jetson project in Somalia, where the predictive models for drought-related migration failed, until the researchers, based on interviews they have conducted, factored in the local goat prices. Because goats did not survive the journey to the border, they had to be sold before moving out.

The fourth quadrant, integrative modelling, seeks to predict outcomes in terms of interpretable factors and causal relationships between them. An example study is conducted by (Bosetti et al. 2019), where mobile phone data was used to create realistic models of refugee mobility and integration, to assess scenarios of a measles epidemic starting in different cities of Turkey. The vaccinated local population and largely un-vaccinated refugee population had different mixing opportunities in different cities, derived from the mobile data, and the influence of this integration indicator in the spreading of the epidemic was assessed via agent based modelling.

3 Why New Data Sources for Migration?

3.1 Limitations of traditional data sources

Migration, influenced by multiple and interlinked factors, is growing globally in scope, complexity and diversity. Despite a large amount of official statistics and administrative data on migration, only part of the complexity of migration phenomena can be captured through traditional migration data. Notable amounts of data on migration have been collected by several countries, however, the mechanisms to centralise, disaggregate and cross reference all data collected from various branches of the government are lacking. Additional important data sources are provided by international organisations, such as the

UN International Organization for Migration (IOM), Organisation for Economic Co-operation and Development (OECD), and United Nations, and these also need to be integrated for wider insights.

The main purpose of much of these data is to answer local/national situations. Data collection methodology is therefore often tailored to local circumstances, which restricts data comparability across countries. Major gaps across the migration data landscape are summarised in four categories (Santamaria & Vespe 2018):

- Problems with existing data
- Issues with the way the data are presented or disseminated
- Data that are not widely collected or are not easily accessible
- Potentially useful data that are currently inaccessible

The global-level migration indicators are relatively under-developed and the problems with existing data include drawbacks and inconsistencies in definitions and typologies; drivers of migration, representation of gender and hidden populations, geographical coverage and timeliness (Ahmad-Yar & Bircan 2021, Bircan et al. 2020). To elaborate further, traditional data collection is mostly limited to certain periods. Country-wide survey studies are expensive and performed once every few years. Field work involves in-depth, concentrated data collection, but it does not have the spatio-temporal coverage of new data sources. Continuous collection of real-time data by various corporations such as mobile phone operators, financial institutions and technology platform companies serving people on the move can provide much more detail on some aspects. Such data seem to be largely untapped for global-level migration indicators at the moment.

These challenges can be addressed by improving the existing datasets by promoting different data sources, specifically big data for the estimations of migration, and by proposing validated best practices of data collection and sharing.

3.2 Challenges for new data sources

There are commercial and/or privacy issues to solve for having access to new data sources, and substantial research and analysis are required. The challenges with the use of big data for migration are (1) institutional interests of political actors shape the decisions about the data collection; (2) methodological heterogeneity due to the different definitions, methods and data sources; (3) nation-states' policies on quantifying migration (Scheel & Ustek-Spilda 2018). Moreover, researchers in the field have increasingly more often faced the paradoxical problem of data abundance. The questions about this new data upsurge are threefold: First, sifting through is a much bigger task than it was in the

past. This warrants larger resources. Secondly, there are limited and emerging regulations concerning data use and release. The third is the challenge of working with partial, incomplete and imperfect data.

These challenges with new data sources for migration studies give rise to a thorough exploration of innovation and cooperation regarding simulations, triangulation and big data analytics. In the meantime, the scholarly field of migration studies remains vulnerable and prone to the damage of pseudoscience or excitement of experiments with data.

3.3 Big Data and migration research in practice

Global Migration Group and others (2017) group main big data sources that have so far been used in migration-related studies under three broad categories:

1. Mobile-phone-based - e.g. call records
2. Internet-based - e.g. social media
3. Sensor-based - e.g. Earth Observation Data (satellite imagery).

In addition to those, there are other alternative data sources in use, such as financial datasets (Gürkan et al. 2021), and media archives (Koch et al. 2020). Increasing number of studies and initiatives are working on the insights on migration phenomena that can be provided through the analyses of big data. In this section, we provide a few illustrative examples of the use of data in practice.

Our first two examples are about the use of mobile call detail records (CDR). The great spatio-temporal granularity of mobile CDR allows detailed investigation of mobility within a country, but also serves to derive proxy indicators of socio-economic factors. In a landmark study, Blumenstock et al. (2015) used mobile CDR metadata from Rwanda to infer proxy indicators for wealth. In order to do that, they have complemented the CDR with a comparatively small (i.e. less than a thousand subject) sample of follow up phone surveys, which provided them with wealth ground truth. This in turn served as labels for a supervised model that was able to predict wealth across the country, at a fraction of the cost of a national household survey, and in a much faster way. This study is also a very good example of integrative modelling discussed earlier, as the analysis revealed mobile data factors about wealth prediction. The authors noted that "features related to an individual's patterns of mobility are generally predictive of motorcycle ownership, whereas factors related to an individual's position within his or her social network are more useful in predicting poverty and wealth" (Blumenstock et al. 2015).

A second interesting study is by Bakker et al. (2019), where mobile CDR is used for measuring social integration of refugees. Using the Data for Refugees Challenge data (Salah et al. 2018), the researchers developed a number of proxy indicators for measuring social integration for each city, such as the amount of space where Syrian refugees and Turkish locals come together, and the ratio of calls made from refugees to locals, as opposed to other refugees. Using 2017

data, they illustrated that some of the cities in Turkey had very low integration scores, and indeed, violent events in 2018 confirmed these insights. An obviously interesting aspect of such analysis is that it can be performed continuously, and used to monitor changes over time, as well as to assess the effect of policy decisions, or changing social factors.

Two important problems in migration research, for which big data approaches based on social media may be useful, are estimating and predicting migrant stocks in a country, and the migration flow from one country to another (Sîrbu et al. 2020).

In nowcasting migrant stocks, typical data sources are official statistics and administrative data, but different countries have differences in the amount and quality of such data, and it is difficult to combine multiple sources. Some of the data will be outdated or not properly updated. Zagheni et al. (2014) investigated the use of Twitter to nowcast migrant stocks. Using either geo-located Tweets, or using migrant language (when it is different from the host country's language), it is possible to find migrants in a country. However, Twitter users have a particular demographic distribution, and only a fraction of the population under study will be using it. Subsequently, such biases must be carefully taken into account. Combining social media and survey studies may result in even better estimates (Alexander et al. 2020).

For predicting migration flows, one problem of traditional models like the gravity model is to predict variations over time for a single migration corridor, as these models depend on variables that do not change over time, such as distance between the countries, or colonial past history. Böhme et al. (2020) illustrate that Google trends data, which collects the frequency of words used in search queries in a country, can be used to predict migration intentions, provide faster estimates compared to official data releases, and improve conventional model estimates. However, as the 'parable of Google Flu' (see (Lazer et al. 2014)) illustrated before, algorithmic changes in Google's search recommendation system, possibilities of external manipulations, and drift in the system may eventually invalidate such a system for prediction. Since Google trends data are not designed for the express purpose of migration flow estimation, such changes may damage its validity and reliability over time. Controlling for dependencies among data, paying attention to measurement and construct validity (Lazer et al. 2014), creating explanations for more transparent processing of the data will enable researchers to use this data source for a longer time. For more specific migration flows, other social media channels or data repositories can be used. For instance, LinkedIn data can be used to investigate migration of professionals (State et al. 2014), or Scopus can be used for migration of researchers (Miranda-González et al. 2020).

Satellite imaging, the third major data source on our list, can be used to gain insight into dynamic settlements of refugee encampments. Quinn et al. (2018) proposed the use of deep learning algorithms on satellite images to estimate refugee occupancy rates in 13 such settlements in Sudan, Nigeria, and Iraq, by detecting structures automatically. While it is possible to use the satellite images for a manual estimation, this will take significant amount of time, as

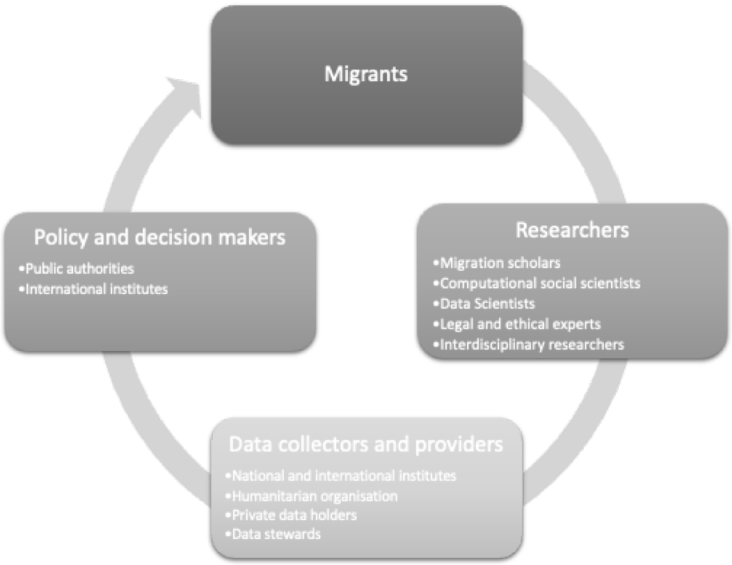

Figure 1: Stakeholders and collaborators in data science for migration and mobility.

there are thousands of structures in each of the investigated settlements.

4 Practices of data sharing

Migration research adopting new technological developments and methodologies can only improve and attain its goals when strong cooperation between the collaborators who provide and/or require the new data sources is established. There are commercial and/or privacy issues to solve for having access to such data and substantial research and analysis are required before it can meet ethical requirements from both research and societal perspectives.

4.1 Stakeholders and collaborators

Figure 1 illustrates the stakeholders and collaborators network in data science for migration and mobility, which includes migrants (who are often the true data owners), researchers, data collectors, data providers, policy and decision makers. We briefly discuss the main concerns and issues from a stakeholder perspective.

Researchers: New data sources offer new tools of analysis to researchers, but also create new methodological challenges, particularly in assessing the representativeness of such datasets. Analytical techniques from computer science are increasingly being used to solve social science problems, including migration

and mobility related challenges, but these are mainly implemented by information and communication technology (ICT) experts and data scientists, and require interdisciplinary collaborations.

The collaborations between data scientists and social scientists require negotiating the interdisciplinary language gap, paying attention to both data-driven and theory-driven aspects of the problem, and applying rigorous methodologies that satisfy both disciplines. A focus on predictive modelling, such that is common in computational sciences, requires careful controls on model complexity and generalisation. A focus on explanation based modelling, on the other hand, demands caution about underlying assumptions, valid research design, and spurious relationships between factors.

Social scientists should be able to understand the assumptions that drive algorithms when assessing their societal implications. Algorithmic approaches affect society in different ways. They may be used to generate analysis results, which are taken up by policy makers and thus influence choices in the public sphere. They can also be incorporated into decision making systems, and directly influence the decisions on a day-to-day basis. For researchers, it may be alluring to use new modelling tools, such as those based on machine learning, or agent-based simulations, to build more powerful and insightful models, and as we argued in the previous section, to integrate explanation based models with predictive models. These approaches can enhance social scientific inquiry in novel ways, address data gaps, save time and other resources, allow more granular and continuous analysis, and with greater coverage.

Data collectors and providers: Both public and private data collectors have a stake in this area. For private companies, the purpose of sharing data can be the realisation of humanitarian goals, and social good (Letouzé & Oliver 2019). In order to do that, the legal and ethical problems of data sharing must be solved (Canca et al. 2021), and the financial aspects must be considered. In particular, for sustainable and long-term social good projects, financing the data collection and sharing is essential.

Policy and decision makers: These groups, including the international institutes, non-governmental and inter-governmental institutions, have a very clear stake. They need knowledge to operate efficiently and effectively, and data may by the key to obtaining this knowledge. However, policy makers are centres of authority and having them access large amounts of data is not without its risks.

Migrants: The final and arguably most important stakeholders are the migrants and refugees themselves, as data owners, as well as the group most influenced by the policy decisions and decision making systems that consume the data. Alencar & Godin (2021) argue that initiatives developed for migrants and refugees under social good initiatives may fail to improve the realities of refugees, or even create technological dependencies that are harmful in the long run. Grassroots movements by migrants and refugees, as well as re-appropriation of data and technologies, are essential in expanding the discussions in this area. It is also important to adopt an international human rights perspective to look at technology use in migration and 'migration management,' with all the debates

that come with the term (Geiger & Pécoud 2010, Molnar 2019).

Box 1 - Data sharing models

Limited data sharing model is when a data holder signs a non-disclosure agreement (NDA) with a selected group and shares data with them. Some of the data challenges in the social good domain have used this model, including the Data for Refugees Challenge (Salah et al. 2018, 2019). The NDA has some legal guarantees about what is done with the shared data, but not really the practical means to enforce them. This model is also less usable in a continuous analysis scenario, where data insights are regularly provided to policy makers.
Remote access model also requires an NDA, but allows access to some form of data remotely. Typically, the data are internally sanitised, and privacy-sensitive information is removed by anonymisation and data aggregation. This sharing model can be somewhat more flexible than the limited data sharing model, but essentially, the difference is in logistics of data.
Application Programming Interfaces (APIs) and Open Algorithms are more systematic ways of approaching the remote access scenario, where an interface is designed to allow data access by queries. The principle is that the entity sending the queries will not access the actual data, but only the outcome of the queries, which are designed in a way to satisfy privacy concerns (Letouzé 2019). This approach provides transparency and mutual trust, but the development of the necessary APIs is costly. This model is called the Question and Answer approach in (De Montjoye et al. 2018).
Pre-computed indicators and synthetic data model, as the name implies, shares only transformed data. Pre-computed indicators are variables that are computed from the raw data, such as wealth proxies based on mobility patterns (Blumenstock et al. 2015). Synthetic data are typically produced by simulations, based on feature distributions of the original data. While producing such transformed data can be easily done in a way to protect individual privacy, group privacy still needs to be considered.
Data collaboratives are "public-private partnerships that allows for collaboration and information sharing across sectors and actors" (Verhulst & Young 2019). This is a general approach in which 'data stewards' within private companies curate and facilitate data sharing for public good. Issues of consent and legal permissions are regulated by the private partner.
Differential privacy seeks to develop data access mechanisms that permit gaining insights about the population without learning anything specific about individuals (Dwork et al. 2014). The core idea here, particularly relevant for the migration domain, is that the anonymisation and aggregation of data, while facilitating data sharing, means a loss of precious information, such as demographics. Differential privacy allows creation of models without developers accessing the raw data. For example, in **federated learning**, a model is sent to many data centres holding some sensitive data, and each time a slightly updated model is passed onwards (Li et al. 2020). This way, the data are not aggregated in a single centre, and the developers of the model do not have actual access to the sensitive details.

4.2 Found data and private data

Found data describes data collected for a purpose different than its present use.[5] For example, mobile call detail records are originally collected by telecommunications companies for marketing insights and customer relationship management, but as they contain very detailed information about human movements in a country, they can be re-purposed for analysing mobility (Salah et al. 2019). Some of the found data repositories are publicly available, but the telecommunications data are typically not. These are both sensitive, and valuable assets for these companies. Accessing such privately owned data sources is difficult. Researchers' access to privately-owned data is a significant step for enabling further computational applications to address diverse questions about migration. However, critical privacy questions arise about the legal requirements, confidentiality, and rules of engagement. Any approach to securing personal data and protecting residents from being predicted, manipulated, or outright controlled via their personal data requires strong encryption and cybersecurity – without back doors.

Industry can generate data and data products that could be potentially accessible to scientists, practitioners, and the public. However, the integration at global level of public data curated and owned by industry is not always feasible or practical. The reasons can be summarised as the lack of shared open access policies and the potential conflict between the regional / national interests of industry and the international dimension and perspective of data sharing. Nevertheless, new industrial collaborations of this kind do occur. While broad strategies and regional mechanisms are being explored, a sharper focus on accountability and monitoring mechanisms are needed. The growing role of the private sector in the governance of AI and data science applications, also for migration studies, highlights the movement away from state responsibility. We discuss some of the risks this may entail in Section 5.

Before we go into the potential benefits and risks of new data sources, we briefly discuss in Box 1 various models of accessing large scale private data. Based on an earlier taxonomy proposed by De Montjoye et al. (2018), Letouzé & Oliver (2019) discuss five major data sharing models, which we summarise here. These models have different advantages and disadvantages.

5 A Case Study on Risks: Digital Identity, Big Data and the Power of Corporations

Until now, we discussed what data science offers for studies in migration and mobility, and discussed the significance of accessible data. In this section, we consider the practice of implementing technological solutions, and analyse a concrete case study on irregular migration for discussing risks. Biased and unfair algorithms, data gaps, non-transparent algorithmic decisions, function

[5] In social sciences, such data are generally called 'secondary data'.

creep for collected data, and loss of privacy and control are all issues raised as significant risks in this domain (Canca et al. 2021). But there are also issues related to conflicting interests of the stakeholders, which we illustrate with an example here.

Consider the role of corporations and their relations with the UN agencies and humanitarian organisations to collect, store and process data in the name of financial integration of refugees at digital identification projects. This is a hot topic in the humanitarian context, and here, we aim to demonstrate how the collection, processing and storage of refugee data could create risks by ignoring the fundamental rights of refugees and serve the corporate interests. This debate also illustrates that large-scale data analysis is not only a methodological and technical issue, but it is a part of a highly political power struggle.

5.1 The identification problem

As the number of refugees and forcibly displaced people worldwide has reached 79.5 million[6], the issue of what to do for their financial and social integration in the countries they take refuge comes to the fore.[7] The push to find solutions is all the more urgent, considering that conflicts, civil wars and other factors that cause immigration will continue, and those who leave their homes and countries cannot return for years, maybe even for generations. Likewise, more than 1 billion people around the world currently cannot benefit from basic rights and services, nor can they participate in financial life due to their lack of identity. Insufficient state capacity or exclusionary political decisions may explain why these people lack formal identification. The United Nations' Sustainable Development Goals scheme aims to overcome the identification question and granting digital identity is one of these solutions (UN 2015), which is mainly promoted by financial, technological corporations and mobile phone operators (Coppi & Fast 2019).

A calculation of the lives potentially transformed by solving the identification question goes from 26 million refugees to 80 million when displaced people are included, and even upwards of 1 billion people when those currently lacking a legal identity are considered. Subsequently, this is first and foremost a data collection and management problem. Regardless of their status, each of these individuals faces obstacles to enjoying their fundamental rights and services, and are relegated to remain outside of financial services and markets.

There are two main reasons for solving the identification question. It is a critical problem that people without a legal identity cannot access essential public services such as education and health, because they are not officially registered. This is particularly true for refugees, asylum seekers and undocumented immigrants. The other issue is to fulfil the 'Know Your Customer' (KYC) requirement that financial institutions and mobile operators must comply with when they provide services to their customers, and acting from this, paving the

[6] https://www.unhcr.org/figures-at-a-glance.html

[7] https://www.gsma.com/mobilefordevelopment/blog-2/using-mobile-technology-provide-functional-identities/

way for the refugees to receive cash transfers as a part of the humanitarian assistance. In this way, it will also be possible for refugees and anyone who does not have an identity card to open a bank account or buy SIM cards for their mobile phones, and become a customer of these corporations.

5.2 Stakeholders and suppliers

The global companies entering the field of refugees and humanitarian aid with innovative solutions became prominent upon the call by the United Nations on the private sector to reach the Sustainable Development Goals by 2030 (UN 2015). The 'UN Guiding Principles on Business and Human Rights' explains how companies can play a progressive role in solving social problems. To integrate refugees into social life, global companies are requested to support the development of refugee-hosting countries through investments and creating employment. In this way, they are encouraged to position refugees as employees or suppliers in the global supply chain, and refugees as employees or entrepreneurs are expected to contribute economically to the host country.[8]

Here, companies are given a progressive, transformative role. In the 1960s and 70s, global companies were accused of being the reason for underdevelopment and the exploitation of the labour force with underground and aboveground resources. However, with these calls, they are attributed a positive role of bringing the corrupt administrators of undeveloped countries to their knees, introducing international standards in these countries, and developing the domestic workforce in terms of technical and qualification.

Understanding this approach is important in terms of seeing how global companies mobilise and use other institutions as intermediaries. With the desire to expand the market and access new data, both as a supplier of these institutions and as a strong 'stakeholder' in this field, together with states, UN organisations and humanitarian aid organisations. One of the first steps in this context is to give identity to refugees, forcibly displaced and those who do not have official registration. It is preferable that this identity is not a paper or plastic one, because it is not possible to fill it with all the information, it can also be lost and forged, and more importantly, it remains with the holder, hence preventing a technology or financial company from accessing these data. However, if the identity is digital, it may be possible to store a wide variety of data items, including name, birthplace and date, as well as vaccine shots, diplomas, and aids received. In addition, security issues such as a person's relationship with criminal organisations can be checked. This way, states, UN agencies, and corporations can access very important information about the background of a person besides the name and place and date of birth.

The question is, who will collect this information and transfer it to digital media? The United Nations and humanitarian organisations are taking on this hard work. These institutions, which set up camps where people took refuge,

[8] https://www.business-humanrights.org/en/big-issues/un-guiding-principles-on-business-human-rights/

distribute aid, determine the refugee status, and perform other difficult and risky tasks, also query the identity of the applicants for aid and transfer the data they obtain to the digital environment. These technological solutions are expensive, and require serious investment. They require not only a technical infrastructure, but also expert teams. Since UN agencies or humanitarian organisations that need to produce quick solutions to emergencies in short-term projects do not have the money and capacity to make this investment, these organisations have to work with mobile phone operators, banks and technology companies, to which they have to give a significant part of the funds they receive. Of course, donor countries can also fund some organisations to develop their own technological infrastructure, but this is not preferred.

This is where the risk of data exploitation enters the picture. As global supplier companies provide technical infrastructure, they are able to access the data of millions of people and control the shaping of the market, thanks to the work of these institutions. In this way, banks organise cash transfers in the field of humanitarian aid, mobile operators distribute SIM cards and offer communication and mobile banking services, while technology companies offer and develop a range of technological products from blockchain to big data analysis and machine learning. The technological literacy required to understand how systems like blockchain operate is very high, and it would be naive to expect the refugees to fully understand the implications. The controversy surrounding UN's collection of biometric information from ethnic Rohingya refugees (discussed by Canca et al. (2021)) illustrated that the refugees often did not understand the forms used in data collection fully, or were afraid of negative consequences in case they did not share data.

As a summary, although digital solutions enable the financial integration of refugees, forcibly displaced, undocumented immigrants, and those without an official identity, as well as their participation in the capitalist economy as customers and employees, this process remains at the initiative of companies, and within the limits set by them. The issue of the rights and freedoms of these communities remains incomplete. Having many stakeholders creates additional risks for the vulnerable, because while data are shared, the conditions under which this is achieved and the influence of different parties in determining what is stored and processed, is not fully transparent.

6 Conclusions

Migration is a complex and a dynamic phenomenon that requires multifaceted expansion of the quantitative approaches through a multi-dimensional lens. We have illustrated in this work that data science and the analysis of secondary data sources can provide valuable insights with high spatio-temporal granularity. One can argue that so far, scattered efforts have been made with regards to the attempts for enhancing the knowledge base on migration- and mobility-related challenges in quantitative terms via new data sources, but coordinated and significant systems have not been established yet.

The use of big data and machine learning allows creation of complex models, with which simple measurements and behaviours, accumulated and integrated over time, can turn into reliable estimators. However, issues of generalisation and model validity, as well as societal and ethical issues stand as challenges, before these approaches turn into mainstream solutions. Traditional approaches using well-controlled and in-depth measurements cannot be replaced by analysis of secondary data, but only be complemented.

When these new tools are added to the toolbox of the researcher, they offer enhanced flexibility in dealing with practical issues. Each country is different in its conditions, resources, and needs. Official statistics and national surveys provide excellent data in some countries, which can afford to dedicate plenty of resources for high-quality data. In other countries, secondary data sources may address much needed data gaps. The data scientific tools developed naturally serve primary data sources as well, and in time, cause better theoretical models to be developed.

The use of secondary data sources to observe human movements is not without its societal risks. Unchecked, it may bestow undesired surveillance capabilities on governments and corporations. The analysis outcomes, or systems built on such data may end up acquiring biases, or providing outcomes that are not sufficiently transparent or accountable. Clearly, these risks need to be openly discussed and solved, before these technologies become usable. We believe the efforts will be worthwhile.

7 A Selected Glossary on Migration

A full blown glossary of technical terms in migration and mobility is beyond the scope of this chapter, but we provide definitions for a few selected terms here. IOM maintains an up to date glossary on migration.[9]

Migration: Most widely adopted definition of IOM (IOM 2019) is also based on the UN definition from 2012 and defines migration as the movement of persons away from their place of usual residence, either across an international border or within a State.

Mobility: Generally, human mobility studies make reference to movements rather than the groups that made them and the places where they occurred. We define human mobility as the temporary movement of persons without the aim of changing their residence, unlike migration.

Immigration: From the perspective of the country of arrival, the act of moving into a country other than one's country of nationality or usual residence, so that the country of destination effectively becomes his or her new country of usual residence (IOM 2019). EU's definition for immigration is the action by which a person establishes his or her usual residence in the territory of a Member State for a period that is, or is expected to be, of at least 12 months, having previously been usually resident in another Member State or a third

[9] `https://publications.iom.int/system/files/pdf/iml_34_glossary.pdf`

country.[10].

Emigration: From the perspective of the country of departure, the act of moving from one's country of nationality or usual residence to another country, so that the country of destination effectively becomes his or her new country of usual residence (IOM 2019). Similar to immigration, EU defines emigration by a projected residence change of at least 12 months.

Internal migration: Internal migration is the movement of people between different residences within the same country (Bartram et al. 2014, Rees 2011).

International migration: Following the UN recommended definitions (2012), we define the (long-term) international migration as the movement to a country other than that of the migrant's usual residence for a period of at least a year (12 months), so that the country of destination effectively becomes the new country of usual residence.

Internal mobility: Internal mobility is the temporary (either short- or long-term) mobility of the residents within the country borders.

International mobility: Unlike international migration, which emphasises the aim of a change in residence between countries, international mobility is defined as cross-border, but temporary movements of the residents of a given country.

Stock migration: For statistical purposes, the total number (at a particular point in time) of international migrants present in a given country, who have ever changed their country of usual residence (UN 2012, IOM 2019).

Migration flows: The number of international migrants arriving in a country (immigrants) or the number of international migrants departing from a country (emigrants) over the course of a specific period (UN 2012, IOM 2019).

Migrant: As defined by the IOM (2019), migrant is an umbrella term, not defined under international law, reflecting the common lay understanding of a person who moves away from his or her place of usual residence, whether within a country or across an international border, temporarily or permanently, and for a variety of reasons. The term includes a number of well-defined legal categories of people, such as migrant workers; persons whose particular types of movements are legally defined, such as irregular migrants; as well as those whose status or means of movement are not specifically defined under international law, such as international students.

Refugees and asylum seekers: Refugees and asylum seekers are migrants who have left their countries and request international protection on account of persecution, war or other factors that put their lives or security at risk (Bartram et al. 2014). In countries with individualised procedures, an asylum seeker is someone whose claim has not yet been finally decided on by the country in which he or she has submitted it. Not every asylum seeker will ultimately be recognised as a refugee, but every recognised refugee is initially an asylum seeker (IOM 2019).

[10]Regulation (EC) No 862/2007 on Migration and international protection

Acknowledgments

This study is supported by European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 870661.

References

Ahmad-Yar, A. W. & Bircan, T. (2021), 'Anatomy of a misfit: International migration statistics', *Sustainability* **13**(7), 4032.

Alencar, A. & Godin, M. (2021), Exploring digital connectivities in forced migration contexts: digital 'making do' practises, *in* 'Data Science for Migration and Mobility', Proceedings of the British Academy.

Alexander, M., Polimis, K. & Zagheni, E. (2020), 'Combining social media and survey data to nowcast migrant stocks in the united states', *Population Research and Policy Review* pp. 1–28.

Allen, W. (2021), Applying computational linguistic and text analysis to media content about migration: Opportunities and challenges for social scientific domains, *in* 'Data Science for Migration and Mobility', Proceedings of the British Academy.

Alpaydin, E. (2016), *Machine learning: the new AI*, MIT press.

Bakker, M. A., Piracha, D. A., Lu, P. J., Bejgo, K., Bahrami, M., Leng, Y., Balsa-Barreiro, J., Ricard, J., Morales, A. J., Singh, V. K., Bozkaya, B., Balcisoy, S. & Pentland, A. (2019), Measuring fine-grained multidimensional integration using, *in* 'Guide to Mobile Data Analytics in Refugee Scenarios: The 'Data for Refugees Challenge' Study', Springer Nature, pp. 123–140.

Bartram, D., Poros, M. & Monforte, P. (2014), *Key concepts in migration*, Sage.

Bircan, T. (2021), Remote sensing data for migration research, *in* 'Data Science for Migration and Mobility', Proceedings of the British Academy.

Bircan, T., Purkayastha, D., Ahmad-yar, A. W., Lotter, K., Iakono, C. D., Göler, D., Stanek, M., Yilmaz, S., Solano, G. & Ünver, Ö. (2020), 'Gaps in migration research: Review of migration theories and the quality and compatibility of migration data on the national and international level.', *HumMingBird* (July).

Blumenstock, J., Cadamuro, G. & On, R. (2015), 'Predicting poverty and wealth from mobile phone metadata', *Science* **350**(6264), 1073–1076.

Böhme, M. H., Gröger, A. & Stöhr, T. (2020), 'Searching for a better life: Predicting international migration with online search keywords', *Journal of Development Economics* **142**, 102347.

Bosetti, P., Poletti, P., Stella, M., Lepri, B., Merler, S. & Domenico, M. D. (2019), Reducing measles risk in Turkey through social integration of Syrian refugees, *in* 'Data for Refugees Challenge Workshop'.

Canca, C., Salah, A. A. & Erman, B. (2021), Ethical and legal concerns on data science for large scale human mobility, *in* 'Data Science for Migration and Mobility', Proceedings of the British Academy.

Coimbra, C., Fatehkia, M., Garimella, K., Weber, I. & Zagheni, E. (2021), Using facebook and linkedin data to study international mobility, *in* 'Data Science for Migration and Mobility', Proceedings of the British Academy.

Coppi, G. & Fast, L. (2019), Blockchain and distributed ledger technologies in the humanitarian sector, Technical report, HPG Commissioned Report.

Cowls, J., Tsamados, A., Taddeo, M. & Floridi, L. (2021), 'A definition, benchmark and database of ai for social good initiatives', *Nature Machine Intelligence* **3**(2), 111–115.

De Mauro, A., Greco, M. & Grimaldi, M. (2016), 'A formal definition of big data based on its essential features', *Library Review* .

De Montjoye, Y.-A., Gambs, S., Blondel, V., Canright, G., De Cordes, N., Deletaille, S., Engø-Monsen, K., Garcia-Herranz, M., Kendall, J., Kerry, C. et al. (2018), 'On the privacy-conscientious use of mobile phone data', *Scientific data* **5**(1), 1–6.

Dietler, D., Farnham, A., de Hoogh, K. & Winkler, M. S. (2020), 'Quantification of annual settlement growth in rural mining areas using machine learning', *Remote Sensing* **12**(2), 235.

Dwork, C., Roth, A. et al. (2014), 'The algorithmic foundations of differential privacy', *Found. Trends Theor. Comput. Sci.* **9**(3-4), 211–407.

Earney, C. & Jimenez, R. M. (2019), Pioneering predictive analytics for decision-making in forced displacement contexts, *in* 'Guide to Mobile Data Analytics in Refugee Scenarios', Springer, pp. 101–119.

Favell, A. (2011), *Eurostars and Eurocities: Free movement and mobility in an integrating Europe*, Vol. 56, John Wiley & Sons.

Geiger, M. & Pécoud, A. (2010), The politics of international migration management, *in* 'The politics of international migration management', Springer, pp. 1–20.

Global Migration Group and others (2017), *Handbook for Improving the Production and Use of Migration Data for Development*, Global Knowledge Partnership for Migration and Development (KNOMAD).

Gürkan, M., Bozkaya, B. & Balcisoy, S. (2021), Financial datasets: Leveraging transactional big data in mobility and migration studies, *in* 'Data Science for Migration and Mobility', Proceedings of the British Academy.

Hilbert, M. (2013), 'Big data for development: From information-to knowledge societies', *Available at SSRN 2205145* .

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S. et al. (2021), 'Integrating explanation and prediction in computational social science', *Nature* .

IOM (2019), *Glossary on migration*, number 34 *in* 'International Migration Law', International Organization for Migration.

Kim, J., Pollacci, L., Rossetti, G., Sirbu, A., Giannotti, F. & Pedreschi, D. (2021), Twitter data for migration studies, *in* 'Data Science for Migration and Mobility', Proceedings of the British Academy.

Koch, C. M., Moise, I., Helbing, D. & Donnay, K. (2020), 'Public debate in the media matters: evidence from the European refugee crisis', *EPJ Data Science* **9**(1), 1–27.

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014), 'The parable of Google Flu: traps in big data analysis', *Science* **343**(6176), 1203–1205.

Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H. et al. (2020), 'Computational social science: Obstacles and opportunities', *Science* **369**(6507), 1060–1062.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M. et al. (2009), 'Life in the network: the coming age of computational social science', *Science* **323**(5915), 721.

Letouzé, E. (2019), Leveraging open algorithms (OPAL) for the safe, ethical, and scalable use of private sector data in crisis contexts, *in* 'Guide to Mobile Data Analytics in Refugee Scenarios', Springer, pp. 453–464.

Letouzé, E. & Oliver, N. (2019), *Sharing is Caring Four Key Requirements for Sustainable Private Data Sharing and Use for Public Good*, Data-pop Alliance and Vodafone Institute for Society and Communications, London.

Li, T., Sahu, A. K., Talwalkar, A. & Smith, V. (2020), 'Federated learning: Challenges, methods, and future directions', *IEEE Signal Processing Magazine* **37**(3), 50–60.

Luca, M., Barlacchi, G., Oliver, N. & Lepri, B. (2021), Leveraging mobile phone data for migration flows, *in* 'Data Science for Migration and Mobility', Proceedings of the British Academy.

Marquez, N., Garimella, K., Toomet, O., Weber, I. G. & Zagheni, E. (2019), Segregation and sentiment: estimating refugee segregation and its effects using digital trace data, *in* 'Guide to Mobile Data Analytics in Refugee Scenarios', Springer, pp. 265–282.

Miranda-González, A., Aref, S., Theile, T. & Zagheni, E. (2020), 'Scholarly migration within mexico: analyzing internal migration among researchers using scopus longitudinal bibliometric data', *EPJ Data Science* **9**(1), 34.

Molnar, P. (2019), 'Technology on the margins: AI and global migration management from a human rights perspective', *Cambridge International Law Journal* **8**(2), 305–330.

Oliver, N., Lepri, B., Sterly, H., Lambiotte, R., Deletaille, S., De Nadai, M., Letouzé, E., Salah, A. A., Benjamins, R., Cattuto, C. et al. (2020), 'Mobile phone data for informing public health actions across the Covid-19 pandemic life cycle', *Science Advances* **6**(23).

Pötzschke, S. (2015), 'Migrant mobilities in Europe: comparing Turkish to Romanian migrants', *Migration Letters* **12**(3), 315–326.

Quinn, J. A., Nyhan, M. M., Navarro, C., Coluccia, D., Bromley, L. & Luengo-Oroz, M. (2018), 'Humanitarian applications of machine learning with remote-sensing data: review and case study in refugee settlement mapping', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **376**(2128), 20170363.

Rees, P. (2011), The dynamics of populations large and small: Processes, models and futures, *in* 'Population Dynamics and Projection Methods', Springer, pp. 1–28.

Robinson, C. & Dilkina, B. (2018), A machine learning approach to modeling human migration, *in* 'Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies', pp. 1–8.

Salah, A. A., Pentland, A., Lepri, B. & Letouzé, E. (2019), *Guide to Mobile Data Analytics in Refugee Scenarios*, Springer.

Salah, A. A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.-A., Dong, X. & Dağdelen, Ö. (2018), 'Data for refugees: The D4R challenge on mobility of Syrian refugees in Turkey', *arXiv preprint arXiv:1807.00523* .

Santamaria, C. & Vespe, M. (2018), 'Towards an EU policy on migration data: Improvements to the EU migration data landscape', *Luxembourg: Publications Office of the European Union* .
URL: *https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technicalresearch-reports/towards-eu-policy-migration-data*

Scheel, S. & Ustek-Spilda, F. (2018), 'Big data, big promises: Revisiting migration statistics in context of the datafication of everything', *Border Criminologies* .

Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. (2012), 'A universal model for mobility and migration patterns', *Nature* **484**(7392), 96–100.

Sîrbu, A., Andrienko, G., Andrienko, N., Boldrini, C., Conti, M., Giannotti, F., Guidotti, R., Bertoli, S., Kim, J., Muntean, C. I. et al. (2020), 'Human migration: the big data perspective', *International Journal of Data Science and Analytics* .

Skeldon, R. (2017), 'International migration, internal migration, mobility and urbanization: Towards more integrated approaches', *Population Division, Department of Economic and Social Affairs, United Nations* .

State, B., Rodriguez, M., Helbing, D. & Zagheni, E. (2014), Migration of professionals to the US, *in* 'International conference on social informatics', Springer, pp. 531–543.

Sterly, H. & Wirkus, L. (2021), Analysing refugees' secondary mobility using mobile phone call detail records (CDR), *in* 'Data Science for Migration and Mobility', Proceedings of the British Academy.

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D. C., Ezer, D., van der Haert, F. C., Mugisha, F. et al. (2020), 'AI for social good: unlocking the opportunity for positive impact', *Nature Communications* **11**(1), 1–6.

UN (2012), *Toolkit on international migration*, United Nations Department of Economic and Social Affairs.

UN (2015), *Transforming our world: The 2030 agenda for sustainable development*, A/RES/70/1.

Verhulst, S. G. & Young, A. (2019), The potential and practice of data collaboratives for migration, *in* 'Guide to Mobile Data Analytics in Refugee Scenarios', Springer, pp. 465–476.

Vertovec, S. (2007), 'Introduction: New directions in the anthropology of migration and multiculturalism', *Ethnic and racial studies* **30**(6), 961–978.

Zagheni, E., Garimella, V. R. K., Weber, I. & State, B. (2014), Inferring international and internal migration patterns from Twitter data, *in* 'Proceedings of the 23rd International Conference on World Wide Web', pp. 439–444.

Zipf, G. K. (1946), 'The P 1 P 2/D hypothesis: on the intercity movement of persons', *American sociological review* **11**(6), 677–686.