

# ARE “DATA FOR GOOD” GOOD DATA?

Research and data quality conceptions  
on the use of big data to study migration



Author: Anna Pascale

Director: Dr. Kenneth Horvath

Expert: Dr. Caroline Roberts

*Thesis submitted to the Faculty of Social and Political Sciences,  
Of the University of Lausanne,  
In Partial fulfillment of the Requirements for the degree of  
Master of Arts in Public Opinion and Survey Methodology*

  
UNIL | Université de Lausanne

*Summer, 2021*



# Acknowledgements

The Master's in Public Opinion and Survey Methodology represented for me a precious occasion for enriching my background and providing my heterogeneous professional profile with a specialization I often felt to miss. I would therefore like to thank all the professors of this course who shared their experience and knowledge with dedication and the highest professionalism. Among them, I want to express a special thanks to Dr. Kenneth Horvath who patiently guided me through the elaboration of the current dissertation and invited me to consider facts and statements under different perspectives, pushing my analysis towards unexpected outcomes. I extend this gratitude to Dr. Caroline Roberts who, besides accepting to act as expert for my defence, has demonstrated how passionately she cared about this Master and its students, and with her kindness and disposition to help made the course easier and more pleasant. Other thanks go to my POSM colleagues whose congeniality made my long and frequent train trips from Geneva funny and less tiring.

I would also like to express hereby my sincere appreciation to the group of experts that have kindly accepted to take part in this study and share with me their time, experience, enthusiasm, and concerns about the developments of a big-data-oriented research in the field of migration.

Along the way that brought me to this dissertation, I bumped into many significant changes and challenges in my life that have made harder for me to complete the path. It is for this reason that I consider this thesis not only as an achievement, but as a little tangible piece of a new stage in my life, a stage where no personal effort or sacrifice is individual but shared with my loved ones. In particular, I share this effort with my husband, Francesco, who I thank very much as he shows me every day how hard he believes in me and that together we might struggle from time to time, but never give up on our ambitions! I would also like to acknowledge here my daughter Layla who, surely, did not make the writing of this thesis easier, but with her smiles and fondness made any free moment more desired and enjoyable. A big thanks to my friends in Geneva who have always offered encouragement and support. Another great appreciation is, finally, for my parents in law and, above all, my parents, and sisters who, despite the distance, made any possible effort to offer their help and made this achievement possible.



# Abstract

Moving from the large debate concerning the role of innovative data sources for social policies and official statistics on the one side, and the ethical concerns involving the use of big data for social research on the other side, this study aims at investigating the evolution of migration research in the big data era. It presents findings from a thematic analysis of transcripts obtained from semi-structured in-depth interviews exploring the experiences and perspectives of seven research and data experts involved in different levels on initiatives that aim to employ innovative data sources and techniques for social good projects unfolding in the field of migration. Interviews had a focus on questions of data quality, regarding both “old” and “new” data sources. Findings from the current study suggest that, in the *datafied* world, research is invested with an operational role oriented towards the formulation of politically relevant insights for which more timely and granular data are demanded, and current official statistics are consequently considered inadequate. The ethical risks associated with the use of big data in a field as sensitive as the one of migration and the necessary involvement of private companies (that hold the greatest part of such a digital knowledge) not only as data providers but, more importantly, as statistics makers, raises concerns about the risks of big data (mis)use, distribution of responsibilities between the public and private sector, and what role is left to official statistics, traditionally entrusted with providing society with a solid, accessible, and enduring complex of knowledge on its wellness.

## Keywords

Official statistics, big data for good, research ethics, research conceptions, data quality, migration, in-depth interviews, economics of convention, solutionism.



# Table of contents

Acknowledgements.....	i
Abstract.....	i
Keywords.....	i
Table of contents .....	iii
<b>Chapter 1</b> <b>Introduction.....</b>	<b>5</b>
1.1 Big data for public good initiatives.....	9
<b>Chapter 2</b> <b>Theoretical background .....</b>	<b>13</b>
2.1 The world of big data and its impact in sociology .....	13
2.2 Data quality in official statistics and big data from a convention theory perspective.....	16
2.3 Solutionism, big data, and migration.....	18
<b>Chapter 3</b> <b>Methodology.....</b>	<b>21</b>
3.1 Data collection mode .....	21
3.2 Interview participants .....	22
3.3 Interview strategy .....	22
3.4 Limitations of the method and ethical considerations .....	25
<b>Chapter 4</b> <b>Findings and discussion.....</b>	<b>27</b>
4.1 Research (on migration) is operational and action-oriented.....	27
4.2 Researchers' mixed feelings when working with big data.....	40
4.3 Shrinking space for official statistics (as we used to know it).....	43
<b>Chapter 5</b> <b>Conclusions .....</b>	<b>47</b>
5.1 Research objectives and findings.....	47
5.2 Research value and suggestions for future development .....	48
5.3 Recommendations .....	48
<b>References .....</b>	<b>51</b>





# Chapter 1 Introduction

In 2014, an article from the Harvard Business Review<sup>1</sup>, has used the adjective ‘good’ to define what big data are not or, better, are not *per se*. After describing Google Flu Trends<sup>2</sup> resounding failure, the analytics expert Kaiser Fung warned against the belief that “*more* data lead to *better* analysis” and put forward data quality argumentations about the overstated validity, completeness, and accuracy of big data. Nevertheless, as the same author states, their value and promises cannot be denied, and the increasing enthusiasm in such innovative sources confirms that quality concerns downsize when huge data volumes immediately available are at stake.

The ever-growing amount of individuals’ digital footprint available worldwide creates in fact new sources of ongoing analytics for social indicators and social policy research that are attracting the interest of actors from diverse sectors. Some government agencies are using these data to obtain measures on urban traffic, tourism, and pricing, and even some statistical offices are exploring the benefits of these innovative data sources compared to survey data (OECD, 2019). As a matter of fact, in the last decade, many new initiatives have emerged aiming at offering a deeper and more efficient analysis of social matters – traditionally investigated through the research instruments most used in social science (surveys, censuses, administrative statistics, etc.) – by drawing on modern – technologically based – data sources.

The observation of the unprecedented capacity of data mining techniques to provide detailed information on individual behaviours and decision-making processes has fuelled the solutionist thought<sup>3</sup> according to which technology can offer a way out to any social problems and, as a consequence, that tech companies might potentially be “the world’s greatest do-gooders” (Nachtwey and Seidl, 2020). The idea that business data may have a social value has been visibly applied during the COVID-19 pandemic. Not only high standing tech corporations like Apple, Google, and Experian have devised and launched apps tracking the spread of the virus<sup>4</sup> but, in Europe, 14 mobile network operators agreed (for the first time ever) to provide data on their clients’ mobile phone location for free to the European Commission’s Joint Research Centre to observe the relationship between human mobility and the

---

<sup>1</sup> [Google Flu Trends’ Failure Shows Good Data > Big Data \(hbr.org\)](https://hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data) <https://hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data> (Accessed 26 March 2021)

<sup>2</sup> [When Google got flu wrong: Nature News & Comment](https://www.nature.com/news/when-google-got-flu-wrong-1.12413) <https://www.nature.com/news/when-google-got-flu-wrong-1.12413> (Accessed 27 March 2021)

<sup>3</sup> See paragraph 2.3 to know more about Solutionism.

<sup>4</sup> [Tech Giants Like Apple, Experian, And Google Fight Pandemic with Digital Tools \(forbes.com\)](https://www.forbes.com/sites/neiledwards/2020/05/21/apple-experian-and-google-fight-pandemic-with-digital-apps/) <https://www.forbes.com/sites/neiledwards/2020/05/21/apple-experian-and-google-fight-pandemic-with-digital-apps/> (Accessed 12 June 2021)

spread of coronavirus and evaluate the effectiveness of mobility restriction measures to contain the pandemic<sup>5</sup>.

It is therefore strengthening the practice to build private-public partnerships to harness and apply data innovations in the resolution or management of those societal issues where an immediate and effective political response is needed. Among these is human mobility and migration. The most precious data sources in this field come in fact from telephones and smartphones, in addition to other sensors recording individuals' locations, social media posts and internet searches, pictures and videos, and satellite data. The amount of more or less structured data extracted from these devices is considered as a valuable resource for the formulation, implementation, and evaluation of migration policies.

In the context of migration research, data have long been seen as a flaw in the scientific efforts to analyse the phenomenon. The UN Population's Division Migration Section offers «the world's most complete set of empirical statistics on the international migrant stock» (Henning and Hovy, 2011), with data for more than 200 countries disaggregated by age, sex, and origin<sup>6</sup>. It is the major collector of information to investigate migration trends. Nevertheless, due to the discrepancies among the numerous data sources used, researchers are invited to be cautious in its usage and go through a preliminary cross-check of references.

The primary sources of migration statistics are border data collection, registries, and field inquiries (UNFPA, 2019). The collection of data at border is composed of the entry and exit statistics collected at all points of entry or departure in a country, including airports and seaports. Field inquiries include censuses and sample surveys, the most comprehensive and detailed sources of granular data on the migrant population, that can be complemented by population registers, employment registers, and information provided by the public offices in charge of granting residence permits. All those are generically called “administrative data sources”<sup>7</sup>.

The inherent complexity of migration, whose causes and dynamics are object of a fascinating area of study<sup>8</sup>, makes very complicate (if ever possible) to have a complete and accurate description on the evolution of the phenomenon. The differences in countries' recording methodology and policy structures, the high costs of fieldworks' data collection processes, and borders' porosity are some of the elements that make migration statistics uncertain and difficult to compare.

The technologic discoveries on the huge potential of digital data in accessing the ‘social’ along with the raise in the global significance of international migration<sup>9</sup> have therefore paved the way for the demand and research of a renovated, immediate, and deeper overview of the phenomenon in its manifestations that would allow to obtain accurate estimates for planning purposes and for informing

---

<sup>5</sup> [Coronavirus: Mobility data provides insights into virus spread and containment to help inform future responses | EU Science Hub \(europa.eu\)](https://ec.europa.eu/jrc/en/news/coronavirus-mobility-data-provides-insights-virus-spread-and-containment-help-inform-future) <https://ec.europa.eu/jrc/en/news/coronavirus-mobility-data-provides-insights-virus-spread-and-containment-help-inform-future> (Accessed 12 June 2021)

<sup>6</sup> <https://www.un.org/en/development/desa/population/migration/data/estimates2/estimates19.asp> (Accessed 19 May 2021)

<sup>7</sup> For more detailed information on the migration data collection instruments and difficulties in the European Union, see Singleton (2016).

<sup>8</sup> See Massey et al. (1993) for a review of the main contemporary theories on international migration.

<sup>9</sup> See Taylor and Meissner (2020) about how the narration of Europe's 2015 migration influx as a crisis fuelled demands for new ways of tracking, mapping, and predicting human mobility.

policies<sup>10</sup>. It is claimed that data can help in describing and predicting people movements and, therefore, in informing more effective policies to discipline the flows, evaluate integration measures, and enforcing cooperation between origin and destination nations (Zagheni, Weber, and Gummadi, 2017).

Migration is maybe one of the social issues that has showed in the most flagrant and, unfortunately very often, dramatic way the limits of current dedicated policies and the fragility of the existing inter-states alliances. The pressure experienced by European country states following to 2015 migration influx pushed international and domestic agendas towards the intention of making the incoming flows of migrants more controllable, if not predictable. The phenomenon of migration, that has generally been considered as “one of the most unpredictable and uncertain population processes”<sup>11</sup>, became associated in international governmental plans, and not only, to terms like “ordered”, “manageable”, “regular”<sup>12</sup>, “predictable”. The ambition to discipline and control irregular migration would be made possible thanks to the massive deployment of new technologies offering tracking, mapping, and forecasting systems of people on the move or willing to move. Some researchers are exploring, for example, the possibility to estimate immigrant stocks by using Facebook advertising platform (Zagheni, Weber, and Gummadi, 2017) or study the link between short-term mobility and long-term migration with geo-tagged Twitter data (Fiorio et al., 2017).

While some of the data that make this kind of studies possible can be accessed freely<sup>13</sup>, the greatest (and easiest to use) part of this high-tech knowledge has a proprietary nature: records from credit card and mobile phone companies, credit agencies and hospitals, able to provide structured data, already stored in databases or that can be easily transferred into one. Gaining access to such data is one of the first obstacles to the use of these alternative sources and, to overcome it, new multi-stakeholders’ partnerships and initiatives have built up with the purpose to activate and facilitate data sharing practices among business companies and governmental and non-governmental actors (including the civil society and the academic community)<sup>14</sup>.

The plan, particularly pursued by inter-governmental agencies and humanitarian organisations, is to make this kind of collaborations and mechanisms structural in any situation where data are necessary for designing a social policy and develop a legal framework guaranteeing access to private data safely and responsibly. One of the most problematic issues – discussed in this study – when sharing and using private data is in fact the risk to erode individual’s privacy and lose track of whom and for what purpose data are used. Partnerships with governments are particularly controversial as these could potentially lead to problematic intrusions into citizens life for purposes of surveillance and control. The issue assumes a more sensitive dimension when it comes to migrants’ and refugees’ data.

Another concern, object of this thesis as well, relates to the quality and reliability of the underlying data or resulting statistics, especially for inference purposes. Although internet coverage has

---

<sup>10</sup> <https://data4migration.org/>

<sup>11</sup> Event “Uncertainty and complexity of migration” (2018) – University of Southampton <https://www.southampton.ac.uk/news/events/2018/11/11-complexity-of-migration.page> (Accessed 14/06/2021)

<sup>12</sup> Consider for example, the Global Compact for Safe, Orderly and Regular Migration | International Organization for Migration (iom.int) (<https://www.iom.int/global-compact-migration>) (Accessed 10 June 2021)

<sup>13</sup> These are mainly data coming from social media posts and internet searches, which are mainly semi- or unstructured and hence require substantial efforts to be transformed into a shape that can be analysed.

<sup>14</sup> Such as, for instance, the “Big data for migration Alliance” (BD4M) declared by the European Commission and the International Organization for Migration in 2018 (see the paragraph below for more details) and Social Science One.

impressively improved and widespread across the world and age-generations in the last years, many are the countries and societal groups affected by a digital gap that might concern a lack of either material or educational means to properly access and utilize technology. The use of social media, moreover, vary significantly according to age and culture. Making use of analytics derived from social networks platforms, credit cards, or communication devices entail a risk of deficiencies in terms of accuracy and representativeness that might make the compiled data not illustrative of the population that researchers and statisticians would like to study.

Although the above mentioned ethical and technical concerns lack a comprehensive solution yet<sup>15 16</sup>, policy makers, operational agencies, and researchers are entrusting the digital traces that migrants leave along their way to compensate the shortcomings of traditional data collection systems and, to this end, are creating more or less formal partnerships preparatory to a structural development of the methodology. These initiatives, in the field of migration, may entail interesting and unprecedented opportunities but also worrisome challenges and risks. Risks concerning individual and group privacy protection (Sandvik & Raymond, 2017), discrimination and perpetuation of colonial asymmetries (Taylor and Meissner, 2020), and human rights violations<sup>17</sup> (Beduschi, 2018).

The structural use of innovative data sources in the field of international migration is a complex process involving different aspects. It requires a deep reflection on the way we want to continue to do research<sup>18</sup>, take decisions, involve diverse stakeholders, and manage risk. Of the multiple dimensions that these reflections entail, this thesis focuses on **how research is engaging in (and affected by) the use of big data for studying migration and what are the consequential effects with respect to research and data quality conceptions and the role of official statistics**. To respond to this research question, a qualitative study was realized based on a review of the most recent literature on the use of big data in social science research (with a particular attention on migration) and a series of semi-structured in-depth interviews to a group of researchers and experts who deal with innovative data sources for studying human mobility and migration. The resulting transcripts were used to extract and analyse underlying conceptions and perceptions on the work of research with data technologies, data quality issues, and space/role left to official statistics in the field.

---

<sup>15</sup> As far as it concerns ethics, a great debate is running at a supranational level and in the humanitarian and human rights sector on data responsibility and protection, but an official legal framework has not been issued yet. Discussions in this direction can be found in: OECD Policy Study on Health Data Governance – Privacy, Monitoring and Research (2015), the Council Recommendation concerning Guidelines covering the Protection of Privacy and Transborder Flows of Personal Data [C(80)58/FINAL], the Inter-Agency Standing Committee Operational Guidance on Data Responsibility in Humanitarian Action.

<sup>16</sup> Statisticians may be able to adjust for the lack of representativeness up to a degree as these adjustments are necessarily harder and require more assumptions than creating analytical weights for surveys with probability-based sampling. Among other statistical techniques, Bayesian statistical models are currently used to make probabilistic assumptions on the biases (such as undercounting, overcounting, etc.) and include them in the model. These probabilistic assumptions can be made either on the basis of previous trends or on elicited expert opinion to create *a priori* distributions. The evidence that each data source has its own problems, the tendency in statistics is to create a synthetic database that integrates the different sources and considers their inconsistencies and tries to harmonize them. Some of the several projects currently active on this front and dealing with harmonizing different sources to produce data and projections on migration are [FUME - Future Migration Scenarios for Europe](https://futuremigration.eu/about-us/) (https://futuremigration.eu/about-us/) and [Quant MiG](https://www.quantmig.eu/) (https://www.quantmig.eu/).

<sup>17</sup> Such as the denial of protection to vulnerable groups of migrants.

<sup>18</sup> In this regard, this research study welcomes and joins the work of Meissner and Taylor (2021) in wondering “How to research migration using data technologies” (UNPUBLISHED)

The current study is structured as it follows. After a short presentation of the most common initiatives and projects where big data play a role in analysing and finding a solution to humanitarian issues (included in the “Introduction” chapter), the theoretical framework guiding the discussion on data quality conceptions and research approach is illustrated: this opens with a presentation of what the advent of big data represented for sociological research, followed by a focus on the economics of convention and solutionism perspectives. The thesis continues with a description of the research methodology adopted and its main limitations and ethical concerns. The last two chapters regard the presentation and discussion of research results, followed by the conclusions that include a summary of the study and suggestions for future research on the topic.

## 1.1 Big data for public good initiatives

We are used to picture *big data* as huge amounts of data collected and exploited mostly by the giants of the Web. Companies like Google, Amazon, and Facebook that use data produced by us, more or less consciously, to extract profitable value. Concretely, this means increasing turnover, expanding the customer base, creating new services and products and, ultimately, boosting profit.

But what if big data were more than that? What if big data were not just an obscure matter contributing at further increasing the resources of wealthy capitalists but a valuable and unprecedented source of social value? Many are the public and humanitarian actors wagering that big data can be transformed into a heritage at the service of global community. Disclosing the information contained in the digital deposits would allow to measure the quality of our life more accurately, create new jobs and business models, but above all acquire a greater understanding of the most complex phenomena concerning us, from environmental to economic and social ones, including migration.

In 2009, at the height of the global financial crisis, the then-Secretary-General of the United Nations (UN) Ban Ki-moon initiated the UN Global Pulse (UNGP) initiative, as a research and development laboratory to find out if and how big data and real-time analyses could contribute to the definition of more agile and effective policies. Since then, at the UN level, in other intergovernmental contexts, and in the civil society many have been the collaborative laboratories and networks created with the explicit goal of “harnessing big data technology for the public good”<sup>19</sup>. Among these, the most famous and influential are the UN Committee of Experts on Big Data and Data Science for Official Statistics (UN-CEBD)<sup>20</sup>, Big Data Europe<sup>21</sup>, DataKind<sup>22</sup>, Data-Pop Alliance<sup>23</sup>, MIT Connection Science<sup>24</sup>, and the Flowminder Foundation<sup>25</sup>. The idea behind all these initiatives is using new data sources and new technical methodologies to inform more targeted policies and enable decision making when addressing the most difficult human challenges, like inequality, poverty, hunger, education, health, and disasters.

---

<sup>19</sup> <https://www.unglobalpulse.org/about/>

<sup>20</sup> <https://unstats.un.org/bigdata/about/index.cshtml>

<sup>21</sup> <https://www.big-data-europe.eu/tag/eurostat/>

<sup>22</sup> <https://www.datakind.org/about>

<sup>23</sup> <https://datapopalliance.org/about/vision-and-members/>

<sup>24</sup> <https://www.media.mit.edu/>

<sup>25</sup> <https://www.flowminder.org/>

Below, it is reported a list (from Ali et al., 2016) offering an idea of how ample the employment of big data analysis techniques (mostly, machine learning) may be in development programs.

**Figure 1 : Big Data projects for different development areas and tools for big data analytics**

Project	Reference	Type	Open Data	Description
<i>Humanitarian emergencies</i>				
Ushahidi	[71]	Online Service	✗	Crowdsourcing platform, merging data from various sources for various humanitarian emergencies.
Digital Humanitarian Network	[72]	Online Service	✗	Network of IT volunteers to leverage the technology to fight human crises.
Trace the Face	[73]	Non-profit	✗	Crowdsourced online platform to help find separated migrants.
GNUcoop	[33]	Online Service	✗	Network of IT professional to deploy information and communication technologies for development (ICT4D).
Open Street Map	[74]	Non-profit	✗	Real-time, online and crowdsourced map service to map a natural crisis of a region.
Pakistan Body Count	[75]	Online Analysis	✗	Makes use of publicly available data to map the casualties caused by suicide bombing and drone attacks in Pakistan.
Services Advisor (UNHCR)	[34]	Web App	✗	Real-time, interactive map services for migrant to find aid agencies' operations and locations quickly.
<i>Healthcare</i>				
Personal Genome Project (PGP)	[76]	Non-profit	✓	Publicly shared genomic data for research purposes.
1000 Genomes Project	[77, 78]	Non-profit	✓	Human genome sequencing to study the relationship between phenotypes and genotypes.
Google Flu Trends	[43]	Non-profit	✓	Inactive now, but provides data about flu and dengue trends for different regions.
Data.gov (Health)	[79]	Governmental	✓	Open health data, tools and applications from the US Government.
UNGP (Health Projects)	[80]	Non-profit	✗	Provides health case studies over different regions of the world.
Health Data (The World Bank)	[81]	Non-profit	✓	Publicly available health data from The World Bank's projects and studies.
Individualized Health Initiative (JHU)	[82]	Academic	✗	Research initiative to promote individualized healthcare by combining and analyzing patients' data from various sources.
Human Connectome Project	[83, 84]	Non-profit	✓	Accurate mapping of human connectome.
<i>Education</i>				
PSLC Data Shop	[85, 86]	Non-profit	✗	Open data repository of learning data for educational data mining.
Educational Data Mining	[48]	Non-profit	✗	Community dedicated for R&D based on learning data mining of different education data.
Data.gov (Education)	[87]	Governmental	✗	Open data, tools and apps related to education at all levels from the US Government.
Education Data (The World Bank)	[88]	Non-profit	✓	Open data from The World Bank's projects and studies related to education.
Alltuition	[89]	Online Service	NA	Makes use of open education data to provide students' with best possible educational financial opportunities.
Simple Tuition	[90]	Online Service	NA	Provides information related to different financial opportunities provides by the educational institutes.
Mastery Connect	[49]	Service Provider	NA	Tool provides teachers with the real-time understanding level of each student of his/her class.
Knewton	[50]	Service Provider	NA	Adaptive tool for to enable individualized learning and teaching.
ThinkCERCA	[51]	Service Provider	NA	Personalized online teaching tool, enable teachers to design and teach according to the changing standards.
UNGP (Education Projects)	[80]	Non-profit	✗	Educational case studies over different regions of the world.
<i>Miscellaneous</i>				
Billion Prices Project	[91]	Academic	✓	Data from many online retailers are analyzed for near real-time economic research.
Predictify.me	[92]	Service Provider	✗	Data collection and predictive analytics for business growth and market understanding.
Be Data Driven	[93]	Service Provider	✗	Development of products for various organization based on their data.
First Mile Geo	[94]	Service Provider	✓	A platform for data collection, visualization and analysis.
Data.gov	[6]	Governmental	✓	Provides open data for different field, e.g., health, education, agriculture etc.
UNGP	[95]	Non-profit	✗	Dedicated to deploy technology and especially big data for development.
The World Bank (Data)	[96]	Non-profit	✓	Provides open data from its various studies and projects for different fields.
<i>Tools for big data analytics</i>				
Open Data Kit (ODK)	[97]	Open Source Tool	NA	Provides online tools for data collection and analysis.
RapidMiner	[98]	Open Source Tool	NA	Predictive analytics platform.
Hadoop	[99]	Open Source Tool	NA	Open source tool for distributed computations on large amounts of data.
Weka	[100, 101]	Open Source Tool	NA	Open source tool to deploy machine learning algorithms for analytics on big data.
Elastic Map Reduce (Amazon)	[102]	Processing Tool	NA	Data processing service for large amounts of data.

Source: Ali et al. *Big Data Analytics* (2016) 1:2. pp. 9-10

A data-driven culture to manage humanitarian issues has also been favoured by the indicator framework (currently comprising 231 unique indicators) adopted to monitor the implementation of the 2030 Agenda for Sustainable Development<sup>26</sup>. Its 17 Sustainable Development Goals have been thought and structured by leveraging the momentum of the data revolution and to demonstrate the centrality of data for development.

The idea that something positive can come out of the overload of data and digitalization was labelled by the same UN Global Pulse as “data philanthropy”, implying a form of cooperation among data-

<sup>26</sup> In 2015, the United Nations (UN) General Assembly adopted Resolution 70/1, Transforming our world: The 2030 Agenda for Sustainable Development, with its 169 targets clustered into 17 Sustainable Development Goals (SDGs) (United Nations 2015).

holding private companies, humanitarian organisations, and/or governments for the benefit of the public. Today, more and more humanitarian and development organisations crave free access to digital data in the attempt to compensate, in this way, the scarcity and inefficiency of national statistics when it comes to deliver data to contrast humanitarian crises and inform timing policy responses. Nowadays, data donation is for corporate enterprises, and not only, one of the best moves in charitable giving. The knowledge that can be gleaned from the massive and incessant amounts of diverse data daily produced is considered immediately *actionable* and can be easily applied to ordinary and emergency aid programs.

UNICEF, for instance, one of the largest and most recognizable humanitarian organizations in the world, has been partnering since 2014 with tech giants like Google and IBM to feed a real-time data software platform with the objective to monitor critical situations and “inform life-saving humanitarian responses”<sup>27</sup>. The software is named “Magic Box” and it works as a collector of anonymized data from different sources (both of public and private nature) and is “composed of multiple github repositories designed to ingest, aggregate, and serve data”<sup>28</sup>.

In the field of migration, the call for an initiative specifically thought to harness big data analytics to track migrants’ flows and stocks and inform relative policies has been fulfilled by the European Union and the UN jointly that, in 2018, launched the “Big Data for Migration Alliance”<sup>29</sup>. Co-convened by The European Commission’s Knowledge Centre on Migration and Demography (KCMD)<sup>30</sup> and the IOM’s Global Migration Data Analysis Centre (GMDAC)<sup>31</sup>, the Big Data for Migration Alliance (BD4M) aims at bridging the gap between data demand and data supply by gathering different stakeholders and creating stable collaborations among them. Data experts, data scientists and academics (investigating the potential of big data to study migration), policy makers informing about the important migration issues at national/regional level, national statistical offices (struggling to satisfy the demand of data from the public and politics), and the private sector (for which the interest is not natural but needs to be solicited) are connected to advance discussions on how to harness the potential of big data sources for the analysis of migration and to treat the phenomenon politically. Despite the large attention devoted in such initiative to ethical and privacy concerns, the way in which the resulting scientific outcomes will be used to design, monitor, and justify political measures is still unclear. The operative application of results derived by deeply granular and algorithmic - based studies into a field as sensitive as the one of migration policy may have unexpected and unintentional implications on the way migration is observed and vulnerable migrants are considered. These implications are investigated and taken into account in this current research work which aims at offering food for thought on how research on the argument can be developed in a way in which scientific enthusiasm does not obscure human beings’ dignity.

---

<sup>27</sup> <https://www.unicef.org/innovation/Magicbox>

<sup>28</sup> <https://mikefabrikant.medium.com/the-magic-box-wiki-a69e20a1dcfe>

<sup>29</sup> <https://data4migration.org/>

<sup>30</sup> <https://ec.europa.eu/jrc/en/migration-and-demography>

<sup>31</sup> <https://gmdac.iom.int/>





## Chapter 2 Theoretical background

This study explores the raise in the use of big data to analyse social issues (i.e., migration), the consequent tensions with official statistics, and the associated research and data quality conceptions by embedding the discussion in the sociological framework provided by the pragmatic approach of the “economics of convention” and the ideas of solutionism.

These two approaches, respectively based on a situationist perspective and focused on considerations of normative beliefs, offer the suitable theoretical context where to analyse critically the social demand for better data and new technologies in the field of migration research and to investigate its methodological implications.

The economics of convention setting is considered relevant for the current discussion because among its extensive considerations about contemporary problems of economic, cultural, and sociological nature, it offers valuable contributions in the field of statistics, one of the fields from where the movement has originated. Some of the key contemporary topics object of its analysis include, in fact, the social construction of qualities and the connection between norms, values, and practical action.

Solutionism ideology provides, in a complementary way, the historical and cultural background ideal to explain the trust and progressive use of new technologies also in the field of migration, and the increasing influence that new actors play in the sector.

The illustration of both approaches is preceded and introduced by a paragraph presenting the main methodological challenges and paradigmatic shifts that data revolution has provoked in sociology, whose perspective is considered and adopted in this thesis.

### 2.1 The world of big data and its impact in sociology

The research problems entailed in the utilization of big data start with their definition. As the use of a vague adjective to characterize the category lets presume (*big* – data), the identification of this kind of figures and its distinction from the universe of possible data seems to be left to our own free interpretation. How “big” must be a data set to be considered belonging to the category of *big data*?<sup>32</sup> In our collective imagination, the term “big data” evokes images of long and shining flows of codes symbolic of a digital era where the humankind has recently landed on. However, in German sociology, there are scholars claiming that big data have always existed or, at least, existed and have been used for science purposes for more than two-hundred years. Baur et al. (2020) see in the censuses, newspapers, and other forms of data collections used by modern administrations an early form of big data, implying

---

<sup>32</sup> This article from “Towards data science” (<https://towardsdatascience.com/how-big-is-big-data-3fb14d5351ba>, accessed 12 June 2021) tried to give a measure which, however, is very dependent on the historical moment taken into consideration.

in this way that it is not the size marking the innovative character of such form of data, but rather its enhanced chances of variety, access, and analysis.

The first documented use of the term “big data” dates to 1997 and referred to the problems detected by scientists at NASA with the visualization of large amounts of data and computers limited capacity<sup>33</sup> to memorise and process an unprecedented size of information. Its introduction in the world of sociology had to wait, instead, for a longer time.

Back in 2007, Savage and Burrows addressed the proliferation of ‘social transactional data’ and ‘digital by-product data’, whose access was granted on a routinely basis, and the impact that these would have provoked in the field of sociology. They did not explicitly use the term “big data”, but the visionary perspective of their work made it become “one of the most cited papers in the discipline of sociology” (Burrows and Savage, 2014) in the 2010s. Although the full potentialities of new digital technologies were not clear yet, the two authors successfully made a sense of the threat that their introduction in the social science research could represent for empirical sociology<sup>34</sup>, whose data collection and analysis systems would have soon appeared worth of a reconsideration in their jurisdiction. In their view, the methodological resources mastered in sociology, such as sample surveys, in depth interviews, and mindful questionnaire designs, would have not represented any longer instruments for a privileged access to societal dynamics but, quite the opposite, dated research methods not justifying the supremacy of sociology in the field among other disciplines. Such new data allow for a proper “metricization of social life” (Burrows and Savage, 2014) revealing most of the forms of human engagement with the social world, and offering, therefore, a new access door to what sociology yearns to describe.

Their comparison and merge with public sources (such as information coming from censuses, electoral rolls, and others) could make possible a match between sociology and transaction data and bring, in this way, knowledge together. However, the authors did not acknowledge the opportunities for a better understanding of the different social dimensions that big data could offer. They rather prospected the creation of a new space of social analysis, made of information easily embedded in time and context and mergeable with a broader spectrum of details. All these elements, along with the engagement of new (private) actors in the routinely collection and analysis of data, precluded in Burrows and Savage’s view the “crisis of empirical sociology” and an upcoming “commercial sociology” not less sophisticated than the first. It happens in fact that sociologists are not the only ones producing sociology now, but a variety of professionals (i.e., statisticians, economists, communications experts, data scientists, etc.) can claim an expertise on the collection and analysis of data generated by the daily transactions among technology consumers and by the social media. It is what Thrift (2005) called the age of *knowing capitalism*, interested by a proliferation of data sources and a renovated interest in – and innovative analysis of – human dynamics and interactions.

In 2014, Rob Kitchin addressed the epistemological challenges posed by big data in social sciences after defining their most peculiar dimensions. The 3 Vs of big data – *volume*, *variety*, *velocity* –, along with their characteristics of being *exhaustive* (potentially covering an entire population), highly

---

<sup>33</sup> See “12 Big Data Definitions: What's Yours?”, published by Forbes on Sep 3, 2014

<https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/> (Accessed 15 June 2021)

<sup>34</sup> Mulvaney (1942) defined Empirical Sociology as “the inductive study of how people appreciate and deal with each other”. Empirical sociologists are therefore engaged in the research and collection of data on the “way people deal with each other”.

detailed in terms of resolution and identification, comparable with different data sources (*relational*), and *flexible* because expandable and scalable in their production. This impressive list of resources brings along, however, also important challenges. As a matter of fact, the timeliness and abundance of data imply serious management and analysis difficulties solvable only with the very powerful computation techniques detecting the model that best explains the findings of the analysed data set. Such approach entails the adoption of a very different perspective with respect to the one used in traditional social research techniques where the analyst has the final say.

According to Kitchin, it is in fact the same significance of *empirical research* that has changed. Before the overwhelming presence of large amounts of data surprised the world of social science, researchers used to plan, design, and run their projects bearing always in mind their research questions. Questions determining the nature, amount, and quality of the collected data. When big-data-based research is conducted, the questions come after receiving the data and with the awareness that, whatever the use is, data will keep coming in great quantity and on a regular basis. This is what big data did and are doing. They are offering a new empirical and epistemological approach where the theory does not guide the process, but it is rather from the data themselves that underlying laws and principles are pulled out (Radermacher, 2019).

The large and repeated use of big data has provoked indeed a new breaking point in the continuum of paradigms alternating in the history of knowledge production. By adopting Kuhn's theory (McLeod, 2020), we can argue that big data have challenged the traditional paradigm of science, based on strict assumptions and scientific samples, and opened the way towards a revolutionary approach for producing knowledge where data are not only an instrument but the origin of the research as well.

In this frame, there is room for the development of a tension between those demanding for a jurisdiction of the traditional scientific methods (although in an extended and renovated way) and who thinks that data can reveal findings by themselves, without any theoretical guidance. The latter is the position of those who see in big data a rebirth of empiricism and the obsolescence of pre-formulated hypotheses and models to be tested. The identification of patterns and correlations inside data is sufficient to produce conclusions and no further experimentations are necessary. It is indeed a shift from a deductive to an inductive approach based on the beliefs that data can provide a full overview on a phenomenon, that preliminary theories and hypotheses become useless, data can offer findings without any human mediation (and, therefore, with no human bias or error), and their insights are not embedded in any specific context or domain.

These argumentations are exposed to the criticism of those pointing, oppositely, that big data are actually subject to sampling bias (given that they are both a representation and a sample) and that the same algorithms used to analyse them are fruit of a theorisation and scientific validation. The combination of these two approaches overcomes the traditional distinction between deductive and inductive to embrace a thinking where concepts, theories, and methodological expertise are used to direct the process of knowledge discovery and, in turn, data are used to orient the further investigations as well as the process of interpretation and theoretical development (Halford & Savage, 2017).

As data-driven science will likely become the prevailing paradigm of scientific method, a re-education on big data appears beneficial and necessary. This should not be limited to the study of computational techniques but of rather the infrastructures through which data are originated, used, and shared. The

development of a larger sensibility on data, embracing also their sociological, political, and cultural aspect, may preserve the individuals from the risk of being exposed to manipulation attempts and would create more opportunities to creatively use data in the different activities of engagement with society.

## **2.2 Data quality in official statistics and big data from a convention theory perspective**

*I'm well aware that my statistics result from conventions,  
yet I believe in a reality that I am asked to quantify (Desrosières, 2009)*

*L'économie des conventions* (EC), or convention theory, has established as a research program and movement of both socioeconomic and sociological character in the mid-1980s, under the lead of a small group of economists, some of them formerly employed at the INSEE, the French national institute for statistics and economic analysis (Desrosières, 2011a).

Moving from a deep research of the foundations of quantification and categorization practices in the history of statistics as a social science<sup>35</sup> (Diaz-Bone, 2019), EC has the great merit of revealing how groundless it is to claim for a “realist” information about a social phenomenon (i.e., unemployment rate)<sup>36</sup> and explains the inevitable entanglement of data production and evaluation with the political and social norms/values prevailing in the relative historical and geographical context<sup>37</sup>.

Conventions, in this approach, raise up to guiding principles and logics that actors refer to coordinate their expectations and evaluate the situations they experience every day. By helping in distinguishing wrong and right, good and bad, relevant and irrelevant, conventions – that often emerge endogenously from the accumulation of many precedents (Young, 1996) – represent the basis to support any qualitative evaluation of social processes and interactions. Under the lens of conventions, EC treats contemporary problems of economic, cultural, and sociological nature.

In the field of statistics, it is appropriate to talk of a “conventionality of statistics” by considering the number of conventions that statisticians need to agree on before measuring a concept (Desrosières, 2009). Exactly as the economics of convention states, conventions represent in this regard the propaedeutic step to each measurement and to the execution of any comparison, coordination, and testing. Measurement in this view is seen as a social process that, in order to be considered valid, requires the acceptance of the adequacy of its processes and outcomes by the concerned community<sup>38</sup>.

---

<sup>35</sup> The economics of convention has also contributed to the definition of the emerging field of “sociology of quantification” in the social sciences (Diaz-Bone and Didier, 2016)

<sup>36</sup> The same reflection can be extended to the idea and demand of measuring the human phenomenon of migration.

<sup>37</sup> In this sense, EC's theoretical support led to the presentation of statistics not as a mere technical tool external to reality and instrumental to the justification of certain political programs or the production of empirical evidence, but rather as an indispensable factor on which economy functioning relies and States evaluate their performances accordingly. In his excursus of the history of statistics, Desrosières (1998) offers numerous examples showing that the transformations of the forms of government and their instruments are directly linked to the appearance (and disappearance) of statistical tools. In the same way, he shows how the same concept of data quality is instrumental to the political needs of the globalized and interconnected European governance order (see the paragraph below).

<sup>38</sup> This concept will return and be contextualized in the field of migration in the conclusions of this research work.

Measurements, uses and the conventions defining the qualities expected of them are co-constructed (Desrosières, 2000). This means that not only conventions determine the resulting statistics, but they are also at the basis of the prevalent common evaluation that is done of a (statistical) information as “trustworthy and relevant” (Diaz-Bone and Horvath, 2021). From this point, it becomes clear the influence that the norms and rules prevailing in a certain societal context have on, among other things, the assessment of data quality.

*We know very little about how users perceive  
and use information about quality dimensions [...]*

*it seems as if producers of statistics have tended to put on hypothetical user hats and tried  
to imagine what they would like to know had they been users of statistics.  
(Groves and Lyberg, 2010)*

The current prevailing framework for defining official statistics data quality (Eurostat, 2000) is considered, for instance, the fruit of the harmonization process that has concerned the European continent in the last forty years (Desrosières, 2000).

Data are collected to the end of detecting phenomena of our society and all the coordination systems that we use to give an order to something that otherwise would be unmanageable and unproductive (Diaz-Bone, 2016). In this frame, the discipline of statistics has been called, in twentieth century, to play the uneasy role of being the repository of the highest and most credible empirical knowledge and, at the same time, the provider of a universal language translating the societies of multiple contexts and connecting the main sources of information to feed an international debate on social issues. Such a task became more intense in the 1980s, when the fast pace of integration running in the European continent stimulated the dissemination of the “quality movement” and, consequently, the need to harmonize European statistics<sup>39</sup> (Desrosières, 2009). The development and diffusion of this movement brought to the affirmation of six quality criteria<sup>40</sup> set out in official European statistics documents which provide for, among other things, attempts to quantify their application level. One of the main interests of European institutions at that time was indeed the definition of measurable tools enabling the comparison among different actors and products. In this sense, the notion of “quality” can be seen as a set of specifications judged to be sufficiently stable over time to provide a foundation for comparison (Desrosières, 2000).

This paradigm, based on a “context of intertemporal comparison and equivalence” (Desrosières, 2009), meets today against other ways of producing data, involving new actors and epistemic values – new “data worlds” (Diaz-Bone and Horvath, 2021). Exactly like official statistics have their own referral order of methodological procedures, quality standards, and working approach, the advent of innovative

---

<sup>39</sup> The homogeneity of this new-born convention space contrasts with the heterogeneity of the national statistics sources where the procedures used to build these standard products remain confined. A demonstration of this contrast is offered by the difficulties that still now are met when trying to measure in a homogenous way a social phenomenon such as migration (see paragraph 4.1.1.3 for more details).

<sup>40</sup> These criteria are relevance, accuracy, timeliness, accessibility and clarity, comparability, and coherence. EC moves against the set of six quality criteria the objection of being explicitly orientated towards a realist perspective that does not bring justice to the influence of judgement conventions and pragmatic methods in the field of data indicators. The need to have a set of criteria responding to the social demand of realist statistics, guaranteed by accountable recognized institutions, overcame any claims for intellectual honesty that experts other than statisticians could arise (Desrosières, 2009).

data production systems (i.e., big data), brings along a new system of values, a digital culture challenging traditional data production systems as based on different premises and concretized in alternative (and evolutive) outcomes. From the standpoint of convention theory many are in fact the limits and data quality problems in the big data era.

In the Economics of Convention perspective, data are considered as “social artefacts and social constructions” (Radermacher, 2019, as cited in Diaz-Bone and Horvath, 2021) fruit of the interaction and mutual transformation produced by different actors and conventions. An integration of official statistics with innovative data would entail the inevitable conflict between those (mainly “methodological statisticians”) who are aware of the conventional nature of data with those who want data to be easily comparable (independently of the referring context) and reality bearer<sup>41</sup>.

The admission of big data in the world of statistics is seen as a fruit of the affirmation of neoliberalism, following to which the state has lost a great part of its authority and control in many sectors, one of these being data production (Diaz-Bone, 2016). Power, in neoliberalism, is widespread, not centralised, and along with power, also control and quality guaranties are diffused. A part of both data collection and data analysis gets ceded to private agencies and other organisations delegitimizing, in this way, state centred official statistics, and making quantification and classification practices hard to access and observe<sup>42</sup>. The expected outcome of this process is a gradual loss of relevance of official statistics and their relative conventions system, that will leave space to new data production systems whose system of values and relative conventions, although largely unexplored yet, are strictly connected with a society regulated and harmonised not through governmental standards and benchmarks but in a common context whose rules and values are disciplined by technology.

### **2.3 Solutionism, big data, and migration**

The book “To Save Everything Click Here”, by E. Morozov, invites to look critically at our technology-driven society and consider the effects that the massive use of IT devices is producing in our own understanding of humanity. He refers to this perspective with the term “technological solutionism”, identifying a logic (or mentality) according to which for every social problem or human “imperfection”, technology can offer a solution.

He finds its best concretisation in those Silicon Valley corporations promoting a new capitalist culture (or “capitalist spirit”, as comprehensively described by Nachtwey and Seidl, 2020) where the entrepreneur has the ambition and means to fix the hardest social problems and, in doing that, pursues and finds benefits for her own business. It is the complementary image of the phenomenon called “philantrocipitalism”: if Bill Gates and Warren Buffet taught us that philanthropy can be a business, solutionists show how business can be philanthropic. If the first has its foundation in the personal wealth and influence that business success allows to accumulate (and in a more or less implicit desire to see what

---

<sup>41</sup> One of the main challenges connected to conventions and their influence in statistics is in fact that, while their impact is well known to indicators and statistics producers, people who use them may not be aware about it in the same way and use data and indicators as expressions of a universal reality, unequivocally true (Desrosières, 2009).

<sup>42</sup> This process is even more evident with the practice of Trusted Smart Statistics in Eurostat (see paragraph 4.3 for more details).

money can really do), the latter is rather based on the instruments offered from the business itself and make philanthropy not as a matter pursued in private sphere, apart from the working life, but as a specific commercial area developed inside the company. It is not a sort of corporate greenwashing operation, but rather the fruit of a deep belief in owning the means to fix and improve the world, and the awareness that by chasing the company's sake, people's sake will follow.

As Morozov reports in his book, Mark Zuckerberg, back in 2008, declared: “there are a lot of big issues for the world to get solved and, as a company, what we are trying to do is to build an infrastructure on top of which to solve some of these problems”. And here it is. Little by little, but more rapidly than many other radical social change processes on earth, a new problem-solving infrastructure has decisively taken place in the world, involving public institutions, and carrying along a series of values and mental paradigms determining our own evaluation criteria.

In such solutionist worldview, there is no space for inefficiency, the status quo can and must be ameliorated, the old must rapidly be replaced by the new. The way towards self-improvement is right there, brightly marked at the sound of “likes”. The internet offers the way and space to designate the perfect society, where a solution is offered to make everything and everyone better. It is a mentality having large applications, not for the single individual only but for the entire society. “If only the right algorithms are in place” (Morozov, 2013), a solution can be found to any problems (climate change, urban traffic, resource distribution), regardless the level of complexity, the dimensions involved and fluidity.

Such logic of easy solutions to complicate problems finds fertile field of application in politics, where problems are notoriously too many and solutions very few. Despite the concerns that the totally new operational environment of these companies has raised and the call for more regulations, whenever a crisis or state of emergency blasts, technology appears as the most immediate and comfort remedy to draw on before (if ever) finding a long-term solution. This has been soundly clear with the COVID breakout, for example, where technology has been massively used to facilitate communication and remote working operations, as well as to contain the contagion. The pandemic has thus offered tech companies a renovated legitimacy to stand along governments in the management of public affairs, making the border between their activity in service providing and governance blurrier (Nachtwey et Seidl, 2020).

The commonality of interests between private companies and governmental institutions towards the realisation of a common good was manifested even before the current corona outbreak, after another alleged “crisis”. In 2015, the rise of mixed migrant flows to Europe has showed the limits of current migration communitarian policy<sup>43</sup> and the fragility of the existing inter-states alliances. In this case as well, new technologies have been called upon to reassure governments that the political and accountability crisis (the real one that the facts of 2015 have manifested in my opinion) that the unexpected inflows of people provoked would have been an isolated case, as the use of digital and satellite devices would have helped in tracking human mobility and, therefore, making the phenomenon more

---

<sup>43</sup> Such as the European Commission's Global Approach to Migration and Mobility (GAMM), articulated in 2011. For more information see European Commission 2011:4-7, and for a critical analysis about how the GAMM contributed to connote migration as a “risk”, see Meissner and Taylor, 2020.

predictable (Meissner and Taylor, 2019). This is made possible thanks to the exploitation of technology true power:

*«Technology is not really about hardware and software anymore. It's really about the mining and use of this enormous data to make the world a better place».*  
*(Eric Schmidt in 2011, when he was still Google's executive chairman)*

The events of 2015 and other extraordinary situations happened in the last decade (i.e., the 2010s Haiti cholera outbreak, the Venezuelan emigration throughout the Americas and Southern Europe, and the post-hurricane Maria exodus from Puerto Rico<sup>44</sup>) demonstrated that the incessant amount of data produced by telephone cells and social media is a valid support to make 'nowcasting' and inform emergency responses in situations of exceptional human mobility. Not only private companies and governments, but also the academics started being attracted by the potentialities and undiscovered use of new data sources to analyse migration and demand more technology-based data to progress with their studies (see the chapter "Findings and discussion" below).

If we evaluate data in an economics of conventions perspective according to which our evaluation criteria are shaped and driven by the system of values, norms, and also political strategies prevalent in the corresponding historical context, then the modern conception of data quality and its principles (described among the findings of this study), in its conventional nature, reflect today the values and standards of a solutionist world, where technology rules our lives and determines what is desirable and what is inadequate. Everything, also research methodology, assumes in this frame the characteristic of a computer-made product designed to be fast, immediately available, in line with current needs, unlimited. The new requirements demanded in social research of an ongoing flow of data always downloadable and fit for the purpose is the fruit of a solutionist culture that app by app has taken place in the palm of our hand and changed the way we see at things and what we want<sup>45</sup>.

---

<sup>44</sup> See, for example. Alexander et al. (2019) and Bengtsson et al. (2015)

<sup>45</sup> This consideration will result clearer after reading the results of this current research project.



## Chapter 3 Methodology

This research is based on a qualitative study conducted through a review of the most recent literature on the use of big data in social science research (with a particular attention on migration) and a series of semi-structured in-depth interviews to experts approaching the subject on different ways and levels. The purpose of this investigation is not to generalize its findings or to measure a concept, but to explore the possible effects that a data collection mode based on big data can have on the approach to migration research. More explicitly, the research question of this study is: *How does the use of big data affect the research on migration? Investigations on research and data quality conceptions on the use of big data to study migration.* In this paragraph, the research design, participants, and strategy will be presented.

### 3.1 Data collection mode

In consideration of the complexity of the topic and the need to stimulate an open conversation, semi-structured in-depth interviews were chosen as data collection method for this study. «In-depth interviewing is a qualitative research technique that involves conducting intensive individual interviews with a small number of respondents to explore their perspectives on a particular idea, program, or situation» (Boyce and Neale, 2006). As a matter of fact, it is common in a constructivist approach to use interview data to let interviewees' beliefs and conceptions emerge (Halldén et al., 2007).

In-depth and semi-structured (as well as unstructured) interviews have the advantage to let the interviewer probe freely and go into depth (Blair et al., 2014), create a relaxed atmosphere favouring participants' will to speak, adapt the questions to the tone, rhythm, and arguments as they raise during the conversation, and allow to follow up on new and unexpected topics. On the other side, in-depth interviews present significant pitfalls, among which the fact of being prone to bias, time-consuming, results difficult to infer, and a strong dependency on the interviewer' skills to determine the success of the research (Hughes et al., 2002).

In order to minimize interviewer effects in this research, interview questions have been designed to facilitate a standard interviewer behaviour and the tone of the conversation was kept informal but neutral. As both the preparation to the interview (specifically targeted according to each participant's background and research projects) and the interview itself proved to be very time-intensive, the number of units to include in the analysis could not be large. In fact, no measurement or generalization about the results of this study is going to be advanced: it is not the objective of the research, and it would be unfeasible due to the small size of the sample and the fact that no random sampling method was used in the selection of the group of participants. However, from the collected interviews emerged a common trend of themes, issues, and topics which let deduce that the sample size was appropriate for the research objective.

Interviews were realized in a time span of one month, from March 12th to April 7th, 2021, and were all conducted from the same author of this study.

## **3.2 Interview participants**

The group of interviewees has varied years of working experience and was drawn from a variety of international projects using innovative data sources with statistical purposes. Potential participants were chosen in consideration of their publications and the relevance of these ones with the object of the study. Seven out of ten people, contacted at first via email, accepted to participate in the interview. The remaining three were not available.

In order to embrace the different fields that innovative data initiatives on migration can involve, there was an effort to invite participants with different affiliations. Among them are academics and computer scientists, statisticians devising models to integrate different data sources, members of international statistical offices involved in projects of official statistics innovation, and representatives of eminent private companies as well as humanitarian organisations.

Interviewees were informed of the purpose of the study, assured of anonymity and confidentiality, and voluntarily consented to participate.

## **3.3 Interview strategy**

Interviews were designed and planned according to three stages (before, during and after the interview) disciplining the correspondent instructions and protocol to follow. Gubrium et al. (2012) and Boyce and Neale (2006) were used as a guide to develop the process.

### **3.3.1 Before the interview – preparatory work**

Once the stakeholders of concern were identified (see the paragraph “Interview participants”), the related interview modalities and instruments were selected and developed: namely, the interview instructions and guide, communication tools, and invitation email.

The experts invited to take part in the research received a detailed email including a short presentation of the author, information on the nature and purpose of the project (with explicit indication of the affiliation with the Universities of Lausanne, Neuchatel, and Lucerne), objectives of the interview, and explanation of the reasons why the expert was selected as a participant. It was also clarified the informal character of the interview, its expected duration, the preferred means of communication, and the confidentiality of the responses.

The email address of Doctor Kenneth Horvath, supervisor of this study, was included in the text as additional contact available to provide any eventual further information on the research. The experts who accepted to join the study were asked to express preferences on the day and time when to conduct the interview.

### 3.3.2 The interview – approach and questions

Interviews were realized remotely and conducted by using the communication platform Microsoft Teams<sup>46</sup>. Interviews opened with a brief introduction on the research project, description of its purpose, and a disclaimer of confidentiality. Participants were provided with further details on the expected development of the interview, were offered the opportunity to make questions, and were finally asked for permission to record. Everyone accepted.

During the interview, questions were formulated to cover the expected list of issues. Although the interview had the style of an open discussion and some questions differed according to the interviewee's expertise and working area, 5-6 key queries led the conversations.

---

#### **Interview's guiding questions**

1. *Can you tell me more about your projects using innovative data sources to study migration?*
  2. *How has your work changed following to the introduction of big data for research purposes?*
  3. *What are the biggest challenges when using big data and other innovative data sources?*
  4. *What are the main data quality issues affecting traditional data sources, in your opinion?*
  5. *What are the main data quality issues affecting innovative data sources, in your opinion?*
  6. *How do you feel about using big data for migration research purposes?*
- 

Questions were designed with an open-ended structure to elicit longer answers on respondents' knowledge, opinions, or feelings. *Instead of questions like "How do you deal with...", formulations like "What experiences with big data come to your mind ...?", "From your own professional perspective, how do you feel about using big data in your own work context ..." were preferred.* Probes were included when needed to ask for concrete examples or more detailed explanations.

Question order was devised to motivate respondents' interest and availability to provide rich of details and honest answers. To this end, the considerations expressed by Dillman et al. (2014) on questionnaire design were taken into account, and factual questions were made before opinion questions. For example, questions such as *"What projects with big data and official statistics you were involved in?"* were made before asking, *"What do you think of the problems that notoriously affect big data?"*

Active listening techniques were carefully applied (facilitated by the possibility to record instead of taking notes), and any new information relevant to the research object was followed up without losing sense of time and progress of the interview. Some longer or more complex replies were summarized to check against the interviewer's interpretation.

At the end of the interview, appreciation for the time dedicated was expressed and it was reiterated that any information included in the thesis does not identify the respondents and that a copy would be shared in case of interest.

Interviews had a duration ranging from 45 to 70 minutes.

---

<sup>46</sup> Video Conferencing, Meetings, Calling | Microsoft Teams <https://www.microsoft.com/en-ww/microsoft-teams/group-chat-software>

### 3.3.3 After the interview – data analysis

As said above, the interviews were conducted via Microsoft Teams and recorded under participants' consensus. The material used for the analysis, therefore, consists of transcripts of the seven interviews.

Interview responses were collected into a unique document (the entire collection of transcripts) and read through more times to familiarise with the available data and look for common arguments, patterns, and differences among participants. The coding scheme adopted at this stage was the result of a reflexive process where opinions, descriptions of experience, and conceptions were extrapolated following to multiple rounds of coding. After reading the transcripts all together, each transcript was examined individually. Both the codes and verbatim extracts of the transcripts were reported onto a sheet of paper and divided in groups according to the topic.

The resulting main topics were the following:

- Perceived problems and limits of current knowledge on migration (disadvantages of official statistics)
- Opinions on the quality of big data
- Opinions on the quality of official statistics
- Motivations to use big data
- Descriptions of big data applications
- Approach to research
- Perception of the work with big data
- Relationship between official statistics and big data
- Perceived risks in terms of privacy and ethics
- Opinions about private sector engagement in data initiatives on migration

The aim was to produce a map of all the topics brought up in the interviews including key sentences expressed by each participant. This is the stage where the individual identities in the data got lost and information were gathered only according to the argument of concern.

A following deeper analysis of the verbatim extracts helped to better identify expressed knowledge and opinions on the different topics, as well as underlying conceptions and perceptions that were used to recognise the most original contributions extractable from the interviews and to group the above listed sub-topics into major, meaningful, and distinguished themes coherent with the research question:

- **A more operational approach in the research on migration**
- **Researchers' mixed feelings when working with big data**
- **Shrinking space for official statistics**

The definition used to name the themes reflect a cautious work of synthesis, interpretation, and abstraction from the verbatim content. In the chapter "Findings and discussion", the results of the above-mentioned qualitative analysis are presented and commented in terms of the theoretical literature evolving around the economics of convention theory and solutionism.

### **3.4 Limitations of the method and ethical considerations**

The impossibility to compare this study with others examining data quality conceptions through surveys or interviews make the theoretical basis of the empirical analysis audacious but also more vulnerable. It is hoped that the approach taken in this research will help to stimulate new analyses and broaden our understanding of data quality and the relative conceptions that data experts and researchers hold, with particular concern to their implications when big data are used to analyse human phenomena.

Another limit of this research may concern the small size of the sample. It could be interesting, in particular, to collect more perspectives from representatives of the private sector whose companies are involved in data sharing practices to study migration. The current study lacks a good coverage of the business sector.

In the data analysis stage, the progressive and repetitive work of grouping and categorization of the different themes and concepts may cause a loss of information. For the scope of the research and for keeping consistency with the research questions, a part of the interviews' contents was excluded from the current analysis.

The partially structured and free style of the interviews entailed that the topics were not all treated with the same level of details by the participants. Some of the experts cared to engage in a deep presentation of big data applications in the field of migration, others in debating the trade-off between big data benefits and risks. This is the reason some participants' quotes are reported more frequently than others' ones.

As described in the previous paragraph, measures were adopted to guarantee confidentiality and participants' private information protection.

- The interviews were kept anonymous.
- During the analysis stage, an identification number was assigned to every participant, and it is used for quoting their answers all over the findings' presentation.
- Any detail that could be linked to any participant was deleted.
- The interview transcripts were used only for research purposes, not shared with any third party, and stored in compliant with GDPR rules.



## Chapter 4 Findings and discussion

The qualitative analysis illustrated above yielded to the following three themes that will be further described and commented individually in the paragraphs below.

- i. Research (on migration) is operational and action-oriented
- ii. Migration researchers' mixed feelings when working with big data
- iii. A shrinking space for official statistics (as we used to know it)

In the next sub-paragraphs participants' data quality conceptions on official statistics and big data emerged in the interviews are commented and framed in consideration of the debate that has developed in the last decades around the concept of data quality and its guiding principles. The objective is to contextualize and explicate participants' answers while pointing out those controversial aspects showing that, on the one side, the expectations with regard to traditional sources are fruit of a new market-and-technology-oriented culture that is investing also the world of academic research and statistics, provoking official statistics' loss of relevance and marginalization. On the other side, it is shown that the integration of innovative data sources into the world of official statistics still necessitates a deep reflection on how their inclusion would change the same domain of statistics (as we used to know it) and how this new hybrid sector (half public and half private) could be made safe and sustainable for everyone.

### 4.1 Research (on migration) is operational and action-oriented

In this paragraph, the term "conception" is used to mean the complex of ideas and beliefs that people may have about a subject. As reported by Meyer et al. (2005), it is common in the literature on "conceptions" theorizing that «a "conception" of some phenomenon reflects (variation in) individuals' experiences and development of an understanding of that same phenomenon in specific contexts». Specifically, for the purposes of this investigation, experts' conceptions of research and data quality are assumed to be associated with the opinions and expressions used while describing their research projects and experience in big data research initiatives.

While studies on research conceptions can be found in the literature (see for example Brew 2001; Meyer et al 2005; Åkerlind 2008, Kiley and Mullins 2005), data quality conceptions, to the best of the author knowledge, have not previously been investigated through qualitative or quantitative studies. The current content analysis focuses therefore on those criteria and relevance that certain aspects may have when using data to investigate human phenomena such as migration.

The elements leading to the conception of a more operational approach in research are based on the consideration of the following aspects emerged in the interviews and discussed in more details in the paragraphs below:

- **Great relevance attributed to the criterium of “timeliness” (to take decisions)**

Among the six quality principles that interviewees considered in the evaluation of strengths and weaknesses of traditional and innovative data sources, fast periodicity and the immediate availability of data appeared to be the main reasons why the use of big data is desirable and official data sources – with their slow validation and release processes – appear as inadequate to today’s research needs on migration.

- **Work with big data to advance policies and resolve problems**

Interview transcripts provide the image of a research activity that aims at being more engaged in the field and that wants to give its contribution by providing practical and actionable answers to politically relevant and useful research questions.

- **Greater attention on migrants’ individual characteristics**

The unprecedented level of details that new technologies offer allows to obtain nano-data about the single facts characterising the migration experience of a person. The modalities used to migrate and migrants’ identity become identifiable and used as explanatory variables within nowcasting and forecasting models. The more granular is the information, the more actionable are the insights that can be formulated.

**4.1.1 Great relevance attributed to the criterium of “timeliness” (to take decisions)**


During the interviews, some questions aimed at investigating the main limits of today’s knowledge on migration and the reasons why drawing on big data to study the phenomenon may be considered a valid alternative. It came out that the instruments traditionally used to extract key numerical information and characteristics of migration (namely administrative registries, censuses, and surveys) are perceived as inadequate, or appropriate for limited purposes, and affected by several quality issues<sup>47</sup>. See the table below.

---

<sup>47</sup> The analysis and comment on the interviews’ outcome offer the context to give more details, updates, and a background on the topics of concern.



**Table 1: Main data quality problems in traditional data sources according to interviews' participants**

<b>F R E Q U E N C Y</b> 	4	A. Official statistics are not timely
	3	B. Admin. sources not designed to produce migration statistics
		C. Surveys and censuses hardly reach the migrant population
	2	D. Low periodicity of statistics
		E. Very few questions on migration in censuses
		F. Secondary movements are not considered or distinguished
	1	G. Scarce level of details and disaggregation
		H. Inconsistency among national data sources
		I. Comparability problems in the data collected across countries
		J. Slow statistics evaluation and publication processes
		K. Very scattered data sources

*Synthesis of participants' answers with indication of the number of people reporting each issue (Source: from the author)*

The items above represent a synthetic collection of answers received to the question:

- *What are the main data quality issues affecting the traditional data sources used for studying migration, in your opinion?*

As can be noticed, despite almost all participants have a statistical background, they reported, as data quality problems, issues with a nonstatistical nature.

For long time, the field of research and statistics has strictly associated the goal of quality to the maximization of *accuracy* in the estimation of statistics. *Accuracy* is generally defined as the proximity between the estimated value (in some cases calculated on a sample) and the (unknown) “true” value for the total population (Grais, 1998). In survey methodology, for example, its assessment is framed in the paradigm of “total survey error”<sup>48</sup>, which sees in the mean squared error (MSE)<sup>49</sup> a specific metric measuring the *accuracy* of survey data. By taking into account the two different groups of errors that may affect a survey - representation errors (or sampling errors) and measurement errors (or non-sampling errors) -, the total survey error offers an indicator of the data quality of the estimate (Groves et al., 2009).

While, in the past century, it was therefore common to associate a small error (and great *accuracy*) with the broader concept of data quality, in the last twenty years, other researchers and statistical

<sup>48</sup> Total Survey Error is a theoretical framework guiding the design of a survey to the end of maximising accuracy by minimising the possible sources of errors: the errors resulting from using a sample to represent a larger population (representation errors or sampling errors) and the errors that occur in survey responses, related to the data collection and processing procedures (measurement errors or nonsampling errors), (Groves et al., 2009).

<sup>49</sup> The Mean Squared Error (MSE) of a survey estimate is the average squared difference between the estimates produced by many hypothetical repetitions of the survey process and the population parameter value. Each estimate computed from survey data has a corresponding MSE that summarizes the effects of all sources of error on the estimate: a small MSE corresponds to a small and controllable TSE, contrarily, a large MSE indicates that one or more sources of error are affecting the accuracy of the estimate (Biemer, 2010).

organizations (Eurostat 2000 and OECD 2003) broadened the set of data quality indicators by including not only *accuracy* but also *relevance*, *timeliness*, *accessibility*, *coherence*, and *comparability*. Groves and Lyberg (2010) care in stressing the nonstatistical nature of these additional indicators that, as such, are hard to measure and difficult to match with the total survey error paradigm. They state that, in a design situation, these dimensions are generally rather considered as the constraints to the realisation of a high-quality survey.

Another interesting aspect observable in the table is the large importance attributed by participants to the scarcity and slowness of traditional migration data: four people remarked the lack of timeliness in official statistics, two pointed out the low periodicity, and one complaint about the slow evaluation and publication processes (all aspects contributing to determine scarce timeliness performances). The large attention given to timeliness and the fact that it is reported by interview participants as a data quality problem offer two interesting causes of reflection. On the one side, their relevance in the sample of the several reasons why official statistics might not be an optimal quality source of data on migration suggests that research has experiencing a shift in the data quality priorities attributed to its products (from accuracy to timeliness). On the other side, it gives an idea of how successfully the quality framework disseminated by international institutions (i.e., Eurostat, UN, IMF, and OECD) have managed to become the reference for quality reporting to their users.

On the basis of this evidence, in the examination (following this paragraph) of official statistics' data quality problems reported by participants, the data quality principles listed in the Code of Practice for the National Statistical Authorities and Eurostat<sup>50</sup> – relevance, accuracy, timeliness and punctuality, accessibility and clarity, comparability and coherence – are used as a reference, that is part of the regulatory framework according to which EU-wide migration and asylum data are collected<sup>51</sup>. The limits listed in table 1 have been attributed by the author of this study to the principles of relevance, timeliness, accuracy, and comparability.

#### **4.1.1.1 Relevance (points B and G in table 1)**

According to the interviewees, official statistics do not provide enough or appropriate material to advance studies on migration.

*Censuses and administrative data are not oriented to the study of migration [3]*

*Do not allow us to understand the short-term variation of migratory flows [3]*

*Administrative sources are not used to produce statistics on migration but to regulate administrative processes... They follow processes and do not include information at an individual level [4]*

*There is no great availability of data [5]*

They refer, in particular, to those public registers presenting the disadvantage of not being specifically designed to collect information on migration but being rather focused on the facts needed for specific administrative purposes.

<sup>50</sup> European Statistics Code of Practice (2017) <https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf/e7f85f07-91db-4312-8118-f729c75878c7?t=1528447068000>

<sup>51</sup> Eurostat Demography and Migration Database: <https://ec.europa.eu/eurostat/web/population-demography-migration-projections/data/database>

Failing in meeting users' needs (in this case migration researchers) and in predicting their research demands and priorities indicates a lack of *relevance*.

As Desrosières (2000) illustrates very well in his analysis of statistics history, *relevance* is one of the criteria most debated in the disciplines of economics and sociology. Moreover, in the years of the construction of the European Union, was key in the public legitimation of its institutions: to the end of strengthening European Union's authority as quality guarantor and source of unified policies in a political context that is still struggling to achieve a complete unification, *relevance* would contribute to offer a structure considered meaningful and effective by its users.

In its definition, it finally prevailed a customer-oriented perspective<sup>52</sup>: «A statistic is *relevant* when it meets user' needs. Identifying users and their expectations is thus essential» (Grais 1998, p. 32). *Relevance* is therefore about the progressive reduction of the gap between data users' needs and data providers' capabilities.

In a conventionalist perspective, the persistence of such a gap, still noticeable in the statistics used for migration studies, is attributed to the fact that, although *relevance* issues are frequently discussed at a national level (within the domestic statistics institutes) by scientist experts on the subject for which statistics should be relevant, at a regional level the discussion is brought by methodologists who can give deeper insights on *accuracy* concerns and who inevitably underestimate the *relevance* aspect of each discussion (Desrosières, 2000).

#### **4.1.1.2 Timeliness and Punctuality (points A, D, J in table 1)**

While acknowledging the great value of official statistics in the formulation of long-term policies (see paragraph below for more details), traditional sources are not considered sufficiently timely by the most of participants. The necessary statistical checks on numbers and validation process from issuing countries on the one side, and the long time-lags between consecutive rounds of data collection on the other side, are perceived as a strong limit to the commencement of new research opportunities and that narrow down the field of data analyses application.

*Timeliness is one of the greatest limits of official statistics [2]*

*It takes too much time to collect and release this kind of data [3]*

*Analyses always come later [4]*

*Eurostat data on asylum request have a three-month delay because of the data quality checks and the need to be validated by the issuing countries [5]*

Among the practical data applications obstructed by the low periodicity of statistics publication mentioned during the interviews are the analysis of variations in migrants flows during short intervals of time – such as, for example, the seasonal migratory movements generally happening for working

---

<sup>52</sup> A customer-oriented standpoint prevailed in fact in the entire preparation of the list of quality criteria dating back to the 1990s. The close focus on the needs of “statistics users” is the reason of the exclusion from this list of important criteria concerning ethics (i.e., independence, neutrality, etc.). Today, in the European Statistics Code of Practice, the five above mentioned groups of criteria are named “Principles of Statistical Output” and, although, they are not changed with respect to the formulation used in the 90s, they have been incorporated in a wider framework including ethical issues and good practices norms connected to “Institutional Environment” and “Statistical Processes”.

reasons – and the study of the changes in the stocks of migrants in a country after a certain shock, such as weather calamities or facts provoking economic instability.

*These [innovative] data allow us to understand the short-term variation of migratory flows... which we cannot do with official data [3]*

*Traditional data sources are collected in a time frame that does not allow for studies on the interval [5]*

The primary purpose of censuses, namely, obtaining data on a given population at a given moment in time, and the long interval between consecutive rounds of data collection (typically, 10 years) are perceived as a limit because do not permit an assessment of the flow of migration on a continuous basis. This is indeed the reason why population censuses are better suited for collecting data on immigrant stocks that, unlike migration flows, are static measures of migration restricted to net residual immigrants in the population at the time of the inquiry (UNFPA, 2019).

#### **4.1.1.3 Coherence and Comparability (points H and I)**

The items “H” and “I” in the table concern a longstanding issue in the history of cross-national research and surveys: coherence and comparability. As examined above (see paragraph 2.2), comparability is probably the quality criterium most representative of the big and controversial efforts of harmonization in statistics accompanying Europe unification process from the 1980s. A homogenization process contrasting with the heterogeneity of the national statistics sources where the procedures used to build standardized statistics remain confined.

*National connotations and definitions make hard to have comparisons [2]*

*There is inconsistency among traditional data sources [3]*

Despite migration and asylum data have become increasingly comparable in recent years in the EU (Singleton, 2016), statistics assessments across countries appear still problematic due to the variety of national data sources, dissimilar definitions used to indicate migration events, and different methods used to collect data.

In a context where statistics are requested to be timely, relevant for the customer, and fit for purpose, the requirements of coherence and comparability<sup>53</sup> reveal the utopian expectation of achieving quality by combining all these principles at the same moment and in the same way. How can statistics be comparable in time and space (original purpose of statistics) if their same definition, classification, and methodological construction must be adapted to customer evolving needs and urgent requests?

#### **4.1.1.4 Accuracy and Reliability (items C, E, F, and L)**

In the history of censuses and survey research, migrants’ participation is traditionally perceived as a major challenge. Tourangeau at al. (2014) rightfully include migrants among those “hard-to-survey” segments of the population due to the numerous difficulties that are usually encountered in sampling and reach them out. As a matter of fact, a part of the migrants living in a country are generally in an

---

<sup>53</sup> Desrosières states that comparability is about the “permanence in time, or the identity in space, of objects whose existence logically precedes measurement procedures”.

irregular situation (people with no residency permit elude any kind of administrative registration), change frequently their place of residence, speak a language different from the official ones, require additional precautions due to vulnerable situations, and may be reluctant to participate in an investigation on their personal life details. These characteristics make difficult (if not impossible) to have up-to-date population registers and demand for the adoption of adapted survey methods. As a result, the biases that most frequently can raise in this field are undercoverage and non-response. Groves et al. (2009) define undercoverage as the weakness of sampling frames<sup>54</sup> failing to include parts of the target population<sup>55</sup> in any survey research.

In their evaluation of the problems affecting official statistics quality and making them less accurate, participants advanced in fact several motivations why coverage is not reliable.

*Changes in personal events are not communicated to the relevant authorities [3]*

*Surveys are conducted at a household level and do not include persons living in collectives [3]*

*Representation is another problem as in censuses, especially, migrants are rarely included [4]*

When the population registers do not include the entire immigrant population, the samples extracted for research purposes are biased or not representative. In most cases, as said before, population registers exclude irregular immigrants – a part of the immigrant population particularly important for the investigation of integration policies and the political debate surrounding migration – and individuals who stay in a country for a period shorter than 12 months<sup>56</sup>. Moreover, in the case of fieldwork investigations, the consideration of household units, as a basis for survey sampling, usually<sup>57</sup> excludes the institutionalised population and other individuals not living in households but in other units, such as communal living arrangements, where is very common to find a component of the migrant population, both for convenience or for cultural reasons. Another complaint recorded in the interviews concerns inaccuracies in the population registers, for which registration and de-registration, in particular in the EU context, is not mandatory, incentivized or facilitated<sup>58</sup>.

*There is no incentive in the countries of origin to register those who leave the country [3]*

This entails that the only people who can be included in administrative sample frames are regular and resident migrants (who usually benefit of a stable housing situation), and that the capacity to study migration in a country and/or across different contexts is significantly limited.

---

<sup>54</sup> Sampling frames are lists or procedures intended to identify all elements of a target population. (Groves et al., 2009)

<sup>55</sup> The target population is the group of elements for which the survey investigation wants to make inferences using the sample statistics. (Groves et al., 2009)

<sup>56</sup> [Sampling immigrants in Europe | Migration data portal](https://migrationdataportal.org/blog/migrant_sampling_using_population_registers) [https://migrationdataportal.org/blog/migrant\\_sampling\\_using\\_population\\_registers](https://migrationdataportal.org/blog/migrant_sampling_using_population_registers) (Accessed 29 May 2021)

<sup>57</sup> The French National Institute for Statistics and Economic Studies carried out surveys of the homeless in 2001 and 2012. Moreover, the Institute's 2009 and 2019 longitudinal surveys on the integration of newcomers (ELIPA and ELIPA2) collected information on recent immigrants who had signed an integration contract, including on people living in collective housing (OECD 2019)

<sup>58</sup> Spain seems to be an exception in this regard as in there the registration is not only mandatory but also incentivized for all types of population, including irregular immigrants, and is needed to access basic public services (for additional information, see [Sampling immigrants in Europe | Migration data portal](https://migrationdataportal.org)) (See footnote 34)

One of the participants denounced another shortcoming in the official data source most commonly used in the field of migration.

*Censuses present only two or three questions on citizens' migration history [4]*

By focusing the attention on the data collection instrument used in censuses, the participant reports here a measurement error, belonging to the group of non-sampling errors. The key components of measurement errors are the respondent, the interviewer, and the questionnaire (Groves et al., 2009). In this case, the interviewee evaluates as insufficient the number of questions that in national censuses are dedicated to the investigation of citizens' (or resident immigrants) migration history. This implies an insufficient level of details on the phenomenon object of investigation, to which other participants have referred to when complaining about the scarce level of details and disaggregation offered by official statistics (item G).

As a matter of fact, although population censuses should offer the occasion for obtaining extensive details not typically available in registers or among the data collected at the border, most of the standard census questionnaires used worldwide present a core of three questions aiming at investigating on individuals' migration history: country of birth, citizenship, and year/period of arrival. As reported by UNFPA (2019), in 2010 Round of Censuses: «Of the 149 countries for which data are available in the United Nations Statistics Division database, more than 87 percent integrated a question in their census about country of birth; 75 percent asked for citizenship; and 50.3 percent asked for immigrants' year or period of arrival. When looking at the combination of core questions included in the questionnaire, 66 percent of all countries asked both questions on country of birth and citizenship and only 34 percent asked all three core questions in their latest census».

In some European countries larger attention is given to the investigation of second-generation immigrants. For this reason, census' questions on migration request for information about parents' country of birth (see UNSTAT National Census Questions Repository<sup>59</sup>).

The United Nations have been advocating for the inclusion in the censuses core of questions about migration of a query on the reasons for migrating, considered fundamental in a period when migratory movements are characterized by mixed flows including, among others, asylum seekers, economic migrants, migrants seeking to reunify with their families, or people without documentation.

The accuracy problems denounced above do not find a well-suited solution in alternative and tech-based data sources where the most serious quality issue concerns indeed accuracy, as confirmed by participants. Interviews' outcomes on this point are illustrated and commented in the paragraph below.

#### **4.1.1.5 Data quality evaluations on big data and other innovative data sources**

In consideration of the analysis developed above about the data quality problems of data sources traditionally used for studies on migration, it is interesting to examine the positive and negative aspects that interview participants presented with regard to innovative data sources.

---

<sup>59</sup> [National census questions repository \(un.org\)](https://unstats.un.org/unsd/demographic-social/sconcerns/migration/census/index.html#/data) <https://unstats.un.org/unsd/demographic-social/sconcerns/migration/census/index.html#/data> (Accessed 30 May 2021)

**Table 2: Main data quality advantages in big data according to participants**

Category	Frequency	Feature
Accessibility		Accessibility is improving
Accuracy		FB has a good coverage all over the age population experiencing migration
		High level of granularity and detail in the information
Comparability		Guarantee of uniformity
Timeliness		Timeliness is one of the greatest advantage
		Data are immediately available
		High frequency data collection

Source: from the author

**Table 3: Main data quality disadvantages in big data according to participants**

Category	Frequency	Feature
Accessibility		Complicate to access
		Different access to media platforms
Accuracy		Representation errors and selection biases
		Unclear definition of concepts (black boxes)
		False information
		Not good to estimate absolute values
		Accuracy unknown
		Missing values
		Biases in the tool impossible to find
		Impossibility to distinguish the units of interest from the others
Clarity		Need to be identified, redesigned and understood
		Insufficiently transparent data production methodology
Comparability		Instable data sources
		Resulting studies can be valid in some countries and not others
		Hystorical analysis are not possible
		Not reliable

Source: from the author

From the tables above, where the frequency of each item is marked by the coloured bars, it clearly emerges that *accuracy* is largely perceived as a shortcoming affecting big data according to interviewees. It was largely recognized that big data is often selective, incomplete, and erroneous. The “exhaust” form of this kind of data worries statistical researchers with respect to the representative nature of the data. As Groves (2011a) has pointed out, big data «rarely offer well-defined coverage of a large population» and researchers using big data sources need to work with what they find in terms of the population that is represented by the data.

*Big data are selection biased as these data are not sampled but self-generated [1]*

*Data coming from Facebook are useful to study and analyse information concerning Facebook users only [3]*

*Data coverage is a problem as many people do not have access to digital devices and credit cards [5]*

Another problem in this regard concerns the control exerted by researchers on the concepts and definitions used to identify a target group. In the case of Facebook advertising platform, for example, that has been tested to estimate or measure mobility (see, for instance, Gendronneau et al. 2019, Spyrtatos et al. 2019), the number of short-term moves or the most recent movers are deduced on the basis of the number of users who live in a country and used to live somewhere else. The terminology currently adopted by Facebook is “away from hometown”. If, from the one side, this guarantees a homogenous classification for any country of origin (differently from what happens with official sources), on the other side no details are provided about what “away from hometown” means. Using social media data entails in fact the adoption of some assumptions (i.e., “black-boxes” definitions) and make these work for the purposes of the research.

*Another important [advantage of big data] is the common definition they have. The national connotation of some terms (i.e., “migrant”) makes difficult to make international comparisons. Although they use black boxes definitions (“according to FB”), this is still quite a guarantee of uniformity. [2]*

*Facebook works like a black box, so to use their data you must accept the default definition of migrant... the algorithm after a while sees that a user is no longer in the previously registered location. There are documents in which Facebook says that this information is calculated by an algorithm that considers both the information provided by the user himself and the IP address which acts as a counterproof of the information provided by the user; it also considers the language. But there is no document that clarifies how a migrant / expat is identified and therefore it is not even stated which information among those considered prevails. This is another data quality problem. It is a black box, and we must trust it as it is. [3]*

*An AI algorithm is generally not transparent and the link from the algorithm to the output is something that only who has done the algorithm knows. And the algorithms produced by big companies are a black box, as intellectual property they can be protected and not declared. [5]*

In terms of *comparability*, a largely complained cause of unreliability is given by the instability of innovative data sources. If the purpose of statistics is to generate harmonized categories allowing for comparisons about the social world in the long term, it is necessary a presumption of enduring existence of those mechanisms (and technologies) producing data and categories. In the case of social media and other innovative data sources this presumption cannot be given for granted, and for several reasons. Social media follow, of course, economic logics according to which their massive use and existence are dependent on the (temporary) success of the platform/tool in the market. Data by-products of an instrument that after a few years gets replaced by a more fashionable one, are not very useful<sup>60</sup>. Moreover, in a same and existing tool, access regulations, working algorithms, and application programming interface (API) keys change over time and sometimes in a more restrictive way.

---

<sup>60</sup> This happened, for example, with Google Plus.



*I started analysing data from a platform that, after a while, didn't exist anymore. And this showed me one of the limitations of these data sources: they are very dynamic and not very stable. They exist today and tomorrow, who knows? They are evolving over time. [2]*

*The algorithms used to categorize a person as a migrant also change frequently. In 2018, 2019 and perhaps even 2020, it was found that stocks of migrants in the UK changed suddenly and significantly, which leads to the conclusion that the change is not actually due to a mobility of migrants but to a change in the algorithm used by Facebook to register them, which is probably aimed at improving it, but which makes the consideration of the data unstable for the purpose of researching trends and identifying small changes within the migrant population [3]*

Despite the inaccuracies, biases and errors characterizing innovative data sources, the undeniable advantages offered in terms of immediate availability, frequent periodicity, and high granularity (see right column in table 2) make still worth their use for research purposes and their eventual introduction in the world of official statistics.

*It is definitely in the future. It is our intention to bring this new world into official statistics, although it's not going to happen overnight. [1]*

The great importance given to these requirements, despite the costs these entail in terms of quality loss, makes clear that users' requirements have changed also in the field of research and that timeliness is not considered any less important than accuracy. Among the six quality indicators internationally adopted, timeliness is probably the one that best symbolizes the market-oriented character that statistics needs to assume to keep being relevant and that all the discourses about the difficulties of producing a high-quality product in an almost-real-time manner are pushed into the background when the logic according to which having some data prevails on the one of having no data at all until when a high level of quality can be guaranteed. The same Eurostat Code of Practice<sup>61</sup> grants the possibility to publish instantly “preliminary results of acceptable aggregate accuracy and reliability when considered useful”. It is sign of a renovated concept of data quality that needs to respond to large numbers' policies, where data and figures are more commonly used not for their specific absolute value but for their capacity to offer a trend or to confirm an observed (or supposed) tendency.

*According to my personal experience, I think that these [innovative] data are good if we want to catch some trends: not looking at absolute values but looking at variations. Although they might be distorted (for different distortion sources), they can catch very well the trends [5]*

*To find trends, we do not need very precise data, but an immediate confirmation in data of the phenomenon we are observing [5]*

The specific values can be corrected later, but some evidence expressed in the shape of numbers is immediately needed and cannot wait for the necessary checks. It is for this precise kind of needs that the use of big data becomes particularly desirable, if not necessary.

---

<sup>61</sup> [European Statistics Code of Practice \(2017\)](https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf/e7f85f07-91db-4312-8118-f729c75878c7?t=1528447068000) <https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf/e7f85f07-91db-4312-8118-f729c75878c7?t=1528447068000>

### 4.1.2 Data research to advance policies and resolve problems (motivations to work with big data)

The perception of timeliness as a major concern emerged from the reflections advanced by participants on the data quality aspects of big data and official statistics is accompanied by a certain consideration of the role of data and their use in research.

If research is commonly conceived as an insightful exploration process aiming at providing a deep understanding of a particular topic, at re-examining existing knowledge (Meyer et al. 2005), or establishing ground truth (Meissner and Taylor, 2021), many interviewers' answers suggest another idea of research strictly connected with the adoption of new working tools: research as instrumental in solving problems and major crises.

*Data can tell you where you need to act... only the migrants who are measured and detected get assistance [2]*

*Timeliness is very important if you need to make decisions. We can't wait for the next census! [2]*

*Official sources will need to change and adapt in order to be relevant to today's policies [4]*

*Research in general and academic can be part of a framework/structure that work in emergency situations (such as COVID) and includes already ethical considerations, principles, and risks. [4]*

*In order to make long term policies, traditional data are very good (they are even the most used), but in order to respond to a crisis we need timelier data [5]*

*[...] I wanted to bring the results of those research to the real world [6]*

The quotes above provide the image of a research activity that aims at being more engaged in the field, a research aware of the relevance of its topic of investigation in the current political agenda and of the potentiality of a newly discovered knowledge, and that wants to give its contribution by providing practical and actionable answers to politically relevant and useful research questions. A research work that can and wants to provide “actionable insights” [2] and, in a speculative solutionist perspective, that can find solutions in data.

The illustration of the reasons pushing participants to draw on big data for their research work helps in enhancing this conception of research as concrete-purpose-oriented. Although not explicit question was made on this subject, the stories that participants shared on their working projects and involvement in the field of migration research manifested the convenience of working with big data from their point of view and their innovative applications.

#### Motivations to use big data and applications extrapolated from the interviews

- Produce evidence for policy making
- Compensate for the lack of official statistics in those countries where there is a scarcity of administrative data and no money to realise surveys, and allow to discover how mobility occurs in those regions where there are no data as migration often escape borders control
- Observe human movements
- Estimate movements for work purposes and for humanitarian situations

- There is so much information in Facebook data with no need to wait for the user to be active (differently from Twitter)
- Understand the short-term variation of migratory flows
- Recommended for studies where timing is crucial, such as following an unpredictable shock in the population
- Observe secondary movements in the continent (sim cards, credit cards)
- Detect in process migrant flows.

The motivations above and underlying perspectives seem to match well with the calls that many international institutions (EU and the UN in particular) have been making in the last years wishing for a structural use of big data analytics to track migration flows and, explicitly, support policy (UN Global Pulse, 2017). The creation of alliances and coalitions (such as for example the Big Data for Migration Alliance and Social Science One, among others) bringing together academics, politics, and industry entails a convergence of interests towards the most pressing and influential requests, making blurrier the distinction among the single original missions.

From these precise observations emerge, moreover, the tension currently experiencing (and well documented in the interviews) migration researchers and statisticians with respect to official statistics, whose qualitative aspects and construction procedures are debated. The discipline of statistics and the same notion of “quality” are strictly entangled with the idea of harmonized categories and specifications that thanks to their stability and reliability over time and space, allow to describe the social world and provide a foundation for comparison. How can statistics be comparable in time and space if their same definition, classification, and methodological construction must be adapted to customers evolving needs and situational requirements? This is one of the points in the debate around official statistics and big data that needs to be further analysed in consideration of the (new) role that society expects from statistics.

### **4.1.3 Greater attention on migrants’ individual characteristics**

In the description of the benefits from the use of innovative data sources and disadvantages of traditional data sources, great relevance in the transcripts is given to the extraordinary disaggregation opportunities offered by digital data.

*Characteristics [about migrants] that you can access [with official statistics] are limited. From social media we do not only know where they are connecting from but also qualitative aspects, like the device they are using. [2]*

*Researchers can use these data to monitor the trend of active users and get an idea of the amount of a subset of migrants [3]*

*Facebook gives the chance to observe the users base thanks to the level of disaggregation [3]*

*It is still impossible to distinguish the migrants from those who are not migrants [4]*

Smartphones, social media, advertising platforms, satellites, sensors, and drones provide indeed an unprecedented level of details not only about people’s movements and activities, but also about individual qualitative characteristics. These tools allow to go deep and obtain microdata about the single facts characterising the migration experience of a person: place of origin, countries crossed, purchases done

in the crossing and destination country, people travelling with. The modalities used to move, and migrants' identity become identifiable, describable, and used as explanatory variables within forecasting models.

Many studies<sup>62</sup> have expressed concern over the effects that such increased level of characterisation of “the migrant” as an individual to study rather than as representative of a human phenomenon may produce. If the use of data technologies aims precisely at improving detection of movements, the expected result is an increased level of differentiation both between “the migrants” and the rest of society, and within the same groups of migrants (distinguishing desirable migrants from undesirable ones), (Taylor and Meissner, 2020). In the first case, the risk is the intensification of labeling practices and of the diffusion of a dangerous and xenophobic narrative across society, in the second one it is easy to fear that control and security concerns may prevail over categorization needs. Moreover, the greater granularity that big data analytics can provide could help answer concerns over the changes in migratory movements and “shape the ways migration becomes relevant” (Taylor and Meissner, 2020) for society and politics.

The three aspects described above (great importance attributed to timeliness, the political utility of research, and the higher level of details on the individual expecting from data) contribute to support a conception of research as an operational discipline that needs to be strictly attached to the current socio-political context, attentive to the needs, and able to provide actionable insights. In a field, such as the one of migration research, where international migration has been traditionally conceptualized as a neutral and “normal”, although complex, human process involving a diversity of motivations (from the economic causal factors considered in the neoclassical theory to the affirmation of migrants as agents evaluating autonomously their convenience to move), it become legitimate to wonder about questions like: who does need timely statistics? Why migration studies must be useful to inform policies? And useful to whom, exactly? To what end?

The resulting perception is the one of a problem-based conception of research, where the objective of the research process is to study the problem and solve it (Meyer et al 2005, 2007). The arguments brought against the adequacy of traditional data sources and the necessity to justify the use of big data in research with the argumentation of pursuing a public good enhances in fact the formalisation of migration as a problem to whose resolution everyone needs to contribute. The risk in this frame is that an unclear specification of where the problem exactly stands and how it wants to be “resolved” might bring to a politicization of the research activity in an unexpected and unwanted sense.

## **4.2 Researchers' mixed feelings when working with big data: concerns and pressure to innovate**

The ethical concerns briefly mentioned in the previous paragraph are not new to the people involved in the data for good initiatives related to migration interviewed for this study. Along with the enthusiasm and excitement manifested in relation to the exceptional potentialities and applications of

---

<sup>62</sup> See, for example, Taylor and Meissner, 2020 – Scheel, 2013 – Didier, 2005 – Bonditti, 2004 – Guild et al. 2008.

innovative data sources in the study of human mobility, a sense of discomfort has in fact emerged in some conversations when more subjective questions were made about the possible implications of using highly disaggregated, large, and granular data in the field. An attempt to reproduce the tension perceived between these two opposite moods is offered in the words cloud below that contains the most common words and expressions used by the interviewed experts to describe their research work with big data. The adjectives and nouns marked in blue present a negative nuance, in red are those expressing positive impressions, and those shown in violet stand for more neutral positions. Although, a higher number of neutral and positively connotated words was recorded, some negative terms (black boxes and mixed feelings) presented a higher number of occurrences<sup>63</sup>.

**Figure 1: Adjectives used by interview participants to define their work with Big Data**



Source: from the author

If from one side, in fact, it was manifested a general sense of trust in the statistical techniques precisely designed to protect individual personal information and anonymity (interview participants described methods such as *differential privacy* [7], and *multi-party computation and privacy-preserving computation* [1]), on the other side it is noticeable a kind of frustration and worry about the uncertainty on the use that is done of the data analyses produced and who is actually benefiting of this work.

<sup>63</sup> The bigger is the word in the cloud the higher is its number of frequencies in the interviews collection.

*[One of the major obstacles when using big data is the] lack of a formal adoption of principles and limits in their use [1]*

*I often wonder who is benefiting of this. Are migrants benefiting? Or these data are rather used to issue stricter policies on migration and build new borders? [2]*

*I have mixed feelings about the work on forecasting asylum requests [2]*

*The problem with innovative data is not that it is produced by private companies, but the fact that there may be political control over these data. [3]*

*The greatest risk and in which there is a shared responsibility [...] is the ethical aspect of using these data in the migration field because there are risks that go beyond the rights of privacy and use of data for surveillance... individual and civil liberties, but also indirect risks that data are used to formulate policies at national level against the populations in need of protection. This risk does not exclusively concern the world of big data, it concerns all data, also those collected with traditional sources, but with big data it is amplified... very granulated because it can lead to the identification of the individual, and this cannot be underestimated. [4]*

*At an academic level, all publications about these methods are made available and transparent in order to be replicated. Therefore, some governments could learn from these techniques and use them for not ethical purposes. [5]*

*Data sharing with governments for good is still embryonal but if this scales the ethical questions will become more important and this is a hugely important issue to address in the near future [6]*

As made clear in the sentences above, one of the first reasons of concern lies in the lack of a legal and accountability framework regulating the production of analyses and official statistics with privately held data, guaranteeing their use in an ethical and rights-based way.

*To overcome this problem, various initiatives are emerging but there is so much to do, and this must be a priority! Guidelines will have to be formulated with principles that must not be violated when using these data [4]*

*It is necessary to be aware of the risks and they need to be overcome in a systematic and responsible way [4]*

In the great field of social research, new data sources availability and the tools to exploit them have widespread at a path faster than any ethical and legal standards could develop regarding the use of such data. In an area such as the one of migration research where data have often been perceived (as illustrated above) as a major constraint, the new wave of opportunities offered by big data brings data collectors and analysts to face new moral dilemmas between the legitimate intellectual curiosity on a new research frontier and the challenges that the proliferation of personal and highly granular data poses to those traditional assumptions about individuals' privacy protection and autonomy.

*I mean... big data is new sources of data; this is what's really valuable! It's not even the size that's the most important. It's new sources of information that we never had before. It's like a new telescope. I don't know what it is. But I'm certainly going to look through it now! [7]*

In the very last years, studies such as Weinhardt (2020) and Meissner and Taylor (2021) are investigating how to combine the ethical rigor demanded in research with the incessant call to exploit any new source of information available. In this impasse, the feeling is that the researcher hesitates to take a clear stance on the issue and keeps doing its job using any available information in the hope that, in the meantime, a regulatory framework is agreed along with a clear attribution of responsibilities and establishment of working principles.

*Statisticians have to come with a solution to respect ethics, otherwise there is no reason for not using any kind of datum we get in contact with, providing that there are some principles [1]*

*The ethical theme is important but cannot represent a limit to the development of big data techniques to make analyses [3]*

In the entire group of participants, it was in fact made clear that ethics cannot represent a limit to innovation: in the frame of a solutionist culture that clearly manifests its effects, the problem is not identifiable in the fast development of new technologies, but in the slow progress of law. Everyone is aware that a big-data-based knowledge is already structural in many realities and whether to innovate data collection methods should no longer be an issue. Some of the interviewed researchers, in fact, especially those engaged in official data projects for governmental or intergovernmental institutions, demonstrated to feel the need and responsibility to include big data in their work:

*In a situation of abundance of data, politicians may draw on controversial sources in order to produce policies [1]*

*It is an obligation and great responsibility [...] to investigate and exploit the dynamics of these new data sources [1]*

*Neglecting the existence of these new data and keep using the old data collection paradigm is not even an option [1]*

*Given the widespread use of big data in various social sphere, we wondered how the use of big data in the field of migration and human mobility was – they have already been used systematically in other areas [4]*

*Data innovation is one of the priorities [4]*

Leaving innovative data exploitation outside of academia or official statistics production centres is not even an option and the knowledge compound on migration will surely have a significant technology-based component. In such a context, the pressing questions relate therefore to what outcome is prospected for those traditional data sources that have been used until now to inform migration policies, and whether to new sources will correspond a new political migration strategy.

### **4.3 Shrinking space for official statistics (as we used to know it)**

The affirmation of big data and new analysis techniques has posed many different challenges in the field of sociology and for the research instruments traditionally used. Some of these challenges were being analysed in their deeper implications before the same term “big data” was coined (see Savage and Burrows, 2007). The loss of appeal of those traditional methodological instruments (such as in-depth interviews or high-quality sample surveys investigations) legitimising sociologists’ expertise in

accessing society knowledge revealed to be the beginning of a deeper crisis investing not only traditional data collection resources but – and more worrying – all the measures that on those methods build their effectiveness and reliability, such as official statistics.

The rise of automatically produced digital data offers a new way of quantifying and investigating the population that not only makes traditional statistics marginal, but it also questions the relevance and necessity of those statistical logics and orthodox scientific practices that, until recently, were considered the fundament of a high-quality and unbiased information. First among all, the normalization of the inductive practice to capture data first and elaborate adaptative research questions later, in the place of “top-down design of classifications and variables to be surveyed” (Radermacher, 2019).

Interview transcripts contain several references to practices and techniques alerting about a reshaping (if not transformation) of official statistics’ role – as we have known it so far – in the near future, and a progressively increasing involvement of private actors in the production of statistics. From the explicit reference to the necessity to produce official statistics with privately held data, to the suggestion of practices devolving a part of the statistics production process to third external actors (see quotes below).

*A good idea could be to work directly with these companies [3]*

*We try to understand together with the private sector how big data could satisfy policy questions [4]*

Such a new way of doing research and issuing statistics is finding a fertile land where to expand its roots in a peculiar historical moment characterized by a popular loss of trust in governments and official institutions, in general (see Davies, 2017<sup>64</sup> and Radermacher, 2019), and a contemporary gain of public influence and confidence in those exponents of the private sector driving the change of society with strong technological forces and responsible for that same data revolution challenging traditional statistics (Nachtwey and Seidl, 2020). High-tech companies and goodwill millionaire enterprises are affirming themselves as easy solutions finders to difficult social problems and, as such, are effectively attracting the interest of decision-making players. In particular, those companies benefiting of a privileged and quantifiable access into people’s feelings, identities, intentions, affiliations and (of course) movements have become an important actor to involve in the world of information supplier traditionally populated either with slow and expensive statistical institutes or argumentative journalists. In the development of statistical processes and measures, new and difficult-to-legitimate-actors are brought on board in the frame of opaque multi-stakeholders’ socio-political initiatives where responsibilities and competencies mix unclearly between governmental and non-governmental entities.

The participation of private actors in the world of statistics and, above all, the input of their data into official statistics – that have always been embedded in countries’ public administration – entail a large discussion about the regulation and use of privately owned data by the public sector. The access to private granular data<sup>65</sup>, that mobile phone or credit card companies make possible, has opened a great thinking about how to implement data sharing and transferring data from private companies to public institutions without undermining citizens’ privacy and private business interests. The European

---

<sup>64</sup> See William Davies (2017) “How statistics lost their power – and why we should fear what comes next”: [How statistics lost their power – and why we should fear what comes next | Government data | The Guardian](https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy) <https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy> (Accessed 3 June 2021)

<sup>65</sup> Eurostat has introduced the term “nano-data” to refer to data records at sub-individual level.



Statistical System has been working on overcoming this obstacle and the serious associated risks<sup>66</sup> by introducing in the statistical system a fundamental paradigm change that brings important implications among which the sharing of competencies and responsibility between statistical offices and private companies.

*We do not ask private companies to share their data but use their data [1]*

*We ask to the business operators to elaborate statistics (not data) according to the criteria established by the “customer/buyer (an institution)”, so we don’t have access to the data but obtain the statistics we want. This is another approach that would allow to overcome the privacy obstacle. [5]*

The one mentioned above is the paradigm of “Trusted Smart Statistics”: «a computation model based on distributing the computation outwards (to the data sources) as opposite to concentrating the data inwards (from the data sources) during the statistics production phase» (Ricciato et al., 2020). In other words, through this model, private companies would share with statistical offices no row aggregated data concerning their customers (or business processes). Any datum would be kept in its private and protected sphere. It is rather the statistical office that shares a part of its computational methodology and technology to the data-sources-company in order to obtain directly the desired resulting statistics.

This process would present the advantages of guaranteeing a stronger protection of data confidentiality, minimizing the costs and burden in terms of technical resources and storage, and sharing (instead of delegating<sup>67</sup>) the control on the process to produce statistics (Ricciato et al., 2020). It is a practice that, on the one side, protects business companies interest in having a full control of “their data” in all the stages of statistical production (this means that companies are confident that data are not breached or used for purposes other than those agreed beforehand), on the other side it externalizes a highly significant part of the statistics production process (computation) that would not be any longer executed by the public statistical office conventionally and historically entitled to do this, but by an uncontrollable profit-driven corporation. This outcome would prefigure a full automatization of the statistical production process (Ricciato, 2020) implying, in a conventionalist perspective, the disappearance of any track of the conventions-building process that has brought statistics to the stage it is now. Not only, but the translation of current statistical methodology encoding into a software program brings with it all the convention-based practices and make them less mutable.

It is part of a statistics innovation process entailing a reinvention of the way of doing statistics and of the same responsible institutions that, in order to survive the challenges and risks that data revolution pose, need to reshape their area of influence. In this “datafied world”, statistical offices have not only processed the idea of losing the monopolist role of statistics producers that used to cover in the past but, in this self-reflection process about the sense of their existence in a crowded data ecosystem, have been painfully carving out their sphere of legitimacy and activity. After abandoning the conception of

---

<sup>66</sup> Among them are breach of individual privacy, potential exertion of mass surveillance and social control, and the collateral introduction of forms of injustices, power asymmetries and racism. See the growing field of Critical Data Studies (Boyd and Crawford, 2012; Iliadis and Russo, 2016) and Data Justice ([www.datajusticelab.org](http://www.datajusticelab.org)).

<sup>67</sup> Here – it is worth explaining – the point of view would be the one of private companies that would totally leave to statistical offices the control over their data with the agreed purpose of producing statistics. A part of the agreement, anyway, nothing would prevent different a usage of data, and this is a reason of high vulnerability for business companies.

a statistics factory bringing questions in and information out in a uniquely internal working procedure, statistical offices are rather mutating in a client of opportunistically different enterprises entitled to hold a resource that for them has always been scarce and expensive to achieve. Statistical offices are now the one who ask and not anymore those who are asked. Their role in this new paradigm is of guarantor of a methodological quality that needs to be translated into a code executable by machines and adaptable to different data sources.

*In the past, when data had to be collected, everything (or almost everything) was done internally in the statistical office: the collection, processing and publication. It was a fine paradigm where policy making questions were transformed into questionnaire items in order to give back answers. [...] But now, in this new world of digital data, the process is not internal and not even managed by public institutions. It is way more complicated and gaining access to data is not easy even for an international statistical office. Data are only in the private domain. [...] So, for a statistician producing official statistics, the old paradigm shifts, as pulling data in does not work any longer. [1]*

Private companies, on the other side, have the occasion to exploit a new market opportunity while carrying their business as usual. Data analysis and statistics production could become in fact a transversal activity whose main clients might be not only statistical offices, but also governments and humanitarian organisations. Business companies, that have notoriously more opportunities and freedom to invest in new activities than intergovernmental public institutions, will have the chance to specialise in the field, making their services increasingly more competitive, and progressively reduce the gap between data users' needs and data providers' capabilities. In such a scenario, it is easy to imagine that the space left to old-fashioned statistics is destined to become smaller and harder to justify.

# Chapter 5 Conclusions

## 5.1 Research objectives and findings

The analysis of how conceptions on research and quality are varying in the data revolution represents an important step towards the understanding of technology's impact in the way we approach social problems and in our own evaluation system. This thesis represents an attempt to investigate this variation by means of a qualitative study based on the contributions of the most recent literature on the use of big data in social science research and a collection of seven in-depth interviews to experts active in the field in different way. The discussion of the resulting findings was approached in a sociological perspective and guided by the paradigm of the economics of convention and the concepts of solution-ism.

The study has a focus on the use of innovative data sources in the field of migration (in particular, international migration) because some significant initiatives, involving an increasing number of actors from different sectors, are attempting to gain a deeper and more immediate level of details on – if not measuring – the phenomenon, but the final purpose, actual benefits, and narrative driving this process appear still opaque and spark uncertainty. Due to the extremely complex nature of migration, these initiatives may entail in fact interesting and unprecedented opportunities on the one side but also worrisome challenges and risks on the other.

In the attempt to make a step towards a clearer overview on the interest behind the demand for an improved knowledge of migratory flows and stocks, the thesis wants to investigate the underlying conceptions of migration research and data quality intrinsic in this kind of initiatives and deduce how migration research is changing in this regard, and what tensions with the existing knowledge may arise. The resulting analysis of the interview transcripts revealed that (1) research on migration is moving towards a more operational purpose and aims at producing results useful to the political management of the phenomenon. That means providing, among other things, insights on the design and evaluation of migration policies and responses to emergency situations. While the hope, in this regard, is that an analytic and deeper understanding may contribute to improve the protection of migrants at risk and integration programs, the doubt that such powerful knowledge would rather be used to enforce more restrictive policies and extreme border securitization is not excluded. The lack of a clear and binding system consistent with human rights protection provokes in this regard (2) mixed feelings in researchers and other experts who acknowledge the necessity to innovate the data-based-information system governing migration knowledge but, at the same time, fear and look for solutions to its unrestrained and uncontrolled use. Finally, the comparison of traditional and innovative data sources along with the illustration of the implications of the innovation process that official statistics are going through (at least in the European context) pointed out that (3) the influence and role of the official statistics system as custodian of evidence-based information on society, is shrinking to share (if not leave) its space with, mostly, private actors demonstrating to be a valid “out-source” not only for their ownership on data but for the wider comparative advantage they benefit of in terms of computing techniques and data

protection. This outcome would be framed in a full automatization of the statistical production process that would make traditional research methods (such as sample surveys, censuses, and interviews) dated and unutilized data collection systems.

## 5.2 Research value and suggestions for future development

As evidenced by the summary above, this study presents a double added value: on the one side, it complements the work of those researchers who wonder how it is possible to do “migration research right in times of big data” (Meissner and Taylor, 2021) and, on the other side, enriches the discussion on the role of official statistics and traditional data sources in a world where technology has been transforming deeply the system of values and reference conventions, and not-institutional actors claim for smarter solutions to change the world. It offers, besides, interesting occasions of further analysis.

A possible expansion could be, for instance, the exploration of other conceptions on big data research and data quality achievable by collecting more numerous and diverse interviews. It could be particularly interesting, for example, to involve different representatives from the private sector willing to express their views on data sharing projects with no-business actors and how these partnerships can be sustainable in the future. This would be particularly helpful to understand if it actually makes sense to talk of data philanthropy (in the sense of data shared and or analysed for free) or if it is rather more realistic that tech companies will extend their data production and analysis business area to the commercialisation of statistics on social phenomena and that data will be another commodity apaanage of those who can pay.

A second future development of this work could be a qualitative analysis of the conceptions that technicians, researchers, and representatives of the business and academic world, engaged in the study of migrant flows using innovative data sources, have on migration itself. The large political interest and incentives to measure and order the phenomenon (consider, among others, the attention raised by the Global Compact for Safe, Orderly and Regular Migration<sup>68</sup>), along with the high priority given to the regular, widespread, and timely collection of migration data, arouse reflections on the underlying consideration of migration as a problem to solve or a risk to manage (Taylor & Meissner, 2020). The exploration of how migration is perceived at an individual level could add up to the knowledge on the popular perception of the phenomenon (that is object of regular investigation through international surveys, like the European Social Survey<sup>69</sup>) and if new dynamics will affect its narrative.

## 5.3 Recommendations

In consideration of the risks and uncertainty that still accompanies the use of new data sources to study social issues and, more generically, to find operational applications in our daily life, it would be desirable the adoption of a prudent approach towards the introduction of digital data in the decision-making

---

<sup>68</sup> Global Compact for Safe, Orderly and Regular Migration [https://www.un.org/en/ga/search/view\\_doc.asp?symbol=A/RES/73/195](https://www.un.org/en/ga/search/view_doc.asp?symbol=A/RES/73/195) (Accessed 19 May 2021)

<sup>69</sup> European Social Survey <https://www.europeansocialsurvey.org/data/themes.html?t=immigration> (Accessed 19 May 2021)

process and keep referring to that expertise and knowledge that has allowed to achieve a certain level of sensitivity with respect to delicate human matters. This is particularly recommended when dealing with complex phenomena like migration, where some practices could develop a number of unwanted damages to already vulnerable populations.

As Desrosières well expressed back in 2000, “the fact of quantifying a phenomenon in a given way is already a social choice”. The attempts to measure and quantify migration by using big data marks a choice whose social implications and underlying motivations are still opaque. Such uncertainty, along with the disconnection between the big-data-branded-knowledge possible on migration and the knowledge obtained in the past decades by pursuing a neutral and detached research interest on the complex dimensions of the phenomenon, generate frustration and concern in the people – researchers, statisticians, analysts, businessmen – that approach the field with good-intentions. “Quantification is frequently impossible or relies on conventions for which no consensus exists” (Desrosières, 2000). If will ever exist a way to make migration quantifiable, it will need to pass for a comprehensive and inclusive quest of a consensus on measuring techniques, applications, and accountability.



# References

- Alexander M., Polimis K, and Zagheni E. (2019) The impact of Hurricane Maria on out-migration from Puerto Rico: Evidence from Facebook data University of Toronto *Max Planck Institute for Demographic Research*
- Åkerlind G. (2008). A phenomenographic approach to developing academics' understanding of the nature of teaching and learning. *Teaching in Higher Education* 13(6): 633-644. DOI: 10.1080/13562510802452350
- Ali A., Qadir J., Rasool R., Sathiaseelan A., & Zwitter A. (2016). Big data for development: applications and techniques. *Big Data Analytics*, 1, 1-24.
- Baur N., Graeff P., Braunisch L., and Schweia M. (2020). The Quality of Big Data. Development, Problems, and Possibilities of Use of Process-Generated Data in the Digital Age. *Historical Social Research* 45 (3): 209-243. doi: 10.12759/hsr.45.2020.3.209-243
- Beduschi A. (2018) The big data of international migration: Opportunities and challenges for states under international human rights law. *Georgetown Journal of International Law* 49(4):981-1071
- Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., Rebaudet, S., & Piarroux, R. (2015). Using mobile phone data to predict the spatial spread of cholera. *Scientific reports*, 5, 8923. <https://doi.org/10.1038/srep08923>
- Bernard G. Mulvaney, C. S. V. (1942). The Place of Empirical Sociology. *The American Catholic Sociological Review*, 3(4), 225-230. doi:10.2307/3707460
- Biemer Paul P. & Lyberg Lars. (2003). Introduction to survey quality. Hoboken, NJ: *Wiley-Inter-science*, <http://www.loc.gov/catdir/toc/wiley032/2003544863.html>
- Biemer P.P. (2010), Total Survey Error: Design, Implementation, and Evaluation, *Public Opinion Quarterly*, Volume 74, Issue 5, Pages 817-848, <https://doi.org/10.1093/poq/nfq058>
- Blair, J., R.F. Czaja, and E. Blair (2013). Designing Surveys: A Guide to Decisions and Procedures. 3rd ed. Los Angeles: *SAGE Publications, Inc.*
- Boltanski L., Thévenot L. (2006) On justification. Economies of worth. *Princeton: Princeton University Press.*
- Bonditti, P. (2004). From Territorial Space to Networks: A Foucauldian Approach to the Implementation of Biometry. *Alternatives*, 29(4), 465-482. <https://doi.org/10.1177/030437540402900405>
- Bosworth M., Guild M. (2008). Governing Through Migration Control: Security and Citizenship in Britain. *The British Journal of Criminology* 48(6): 703-719. DOI: 10.1093/bjc/azn059
- Boyce C., Neale P. (2006) Conducting In-Depth Interviews: A Guide for Designing and Conducting In-Depth Interviews for Evaluation Input. *Pathfinder International Tool Series*
- Boyd D., Crawford K. (2012). Critical Questions for Big Data, Information, Communication & Society 15(5): 662-679. DOI: 10.1080/1369118X.2012.678878
- Brew A. (2001). Conceptions of Research: a Phenomenographic Study. *Studies in Higher Education* 26(3): 271-285. DOI: 10.1080/03075070120076255
- Burrows R., & Savage M. (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data & Society*. <https://doi.org/10.1177/2053951714540280>
- Davies W. (2017). How statistics lost their power – and why we should fear what comes next. <https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy> (Accessed 14 June 2021) *The Guardian*.
- Desrosières A. (1998) The politics of large numbers – A history of statistical reasoning. Cambridge Massachusetts: *Harvard University Press.*

- Desrosières A. (2009): How to be real and conventional. A discussion of the quality criteria of official statistics. *Minerva*, 47(3), 307–322
- Desrosières A. (2011a). The Economics of Convention and Statistics: The Paradox of Origins. In: *Historical Social Research*, 36(4), 64–81
- Desrosières A. (2000). Measurement and its Uses: Harmonization and Quality in Social Statistics 68(2): 173–187. DOI: 10.1111/j.1751-5823.2000.tb00320.x
- Dewey, John (1938): Logic. The theory of inquiry. New York: Holt
- Diaz-Bone R. (2016) Convention Theory, Classification and Quantification. *Historical Social Research* 41 (2): 48–71. doi: 10.12759/hsr.41.2016.2.48-71.
- Diaz-Bone R. and Didier E. (2016) The Sociology of Quantification – Perspectives on an Emerging Field in the Social Sciences. *Historical Social Research* 41 (2): 7–26. doi: 10.12759/hsr.41.2016.2.7-26.
- Diaz-Bone R. (2019) Convention Theory, Surveys and Moral Collectives. In: Joller S., Stanisavljevic M. (eds) *Moralische Kollektive. Wissen, Kommunikation und Gesellschaft (Schriften zur Wissenssoziologie)*. Springer VS, Wiesbaden. [https://doi.org/10.1007/978-3-658-22978-8\\_7](https://doi.org/10.1007/978-3-658-22978-8_7)
- Diaz-Bone R. and Horvath K. (2021) Official statistics, big data, and civil society. Introducing the approach of “economics of convention” for understanding the rise of new data worlds and their implications. *Statistical Journal of the IAOS*, vol. 37, no. 1, pp. 219–228, 2021 DOI: 10.3233/SJI-200733
- Didier B. & Elspeth G., Introduction: Policing in the Name of Freedom’, in *Controlling Frontiers: Free Movement into and Within Europe*, eds *Didier Bigo and Elspeth Guild* (Aldershot: Ashgate, 2005), 1.
- Dillman D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method* (4th ed). Hoboken, NJ: John Wiley & Sons, Inc.
- Eurostat (2000). *Assessment of the Quality in Statistics*. Eurostat General/Standard Report, Luxembourg, April 4–5
- Fiorio L. et al. (2017), Using Twitter Data to Estimate the Relationship Between Short-term Mobility and Long-term Migration, *ACM*, New York, NY, USA, <http://dx.doi.org/10.1145/3091478.3091496>
- Gendronneau C., Wiśniowski A., Yildiz D., Zagheni E., Fiorio L., Hsiao Y., Stepanek M., Weber I., et al. (2019). *Measuring Labour Mobility and Migration Using Big Data*. European Commission, Brussels
- Grais B. (1998). Harmonisation statistique et qualité : le cas des statistiques sociales. Paper at Eurostat Seminar in Mondorf on “The Future of European Social Statistics” (4th session, March 26–27, 1998).
- Groves, R.M., F.J. Fowler, Jr., M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau (2009). *Survey Methodology*. 2nd ed. Hoboken, NJ: Wiley.
- Groves R. and Lyberg L. (2010) Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly* 74(5):849–879 DOI: 10.2307/40985408
- Groves R. (2011a). Three Eras of Survey Research. *Public Opinion Quarterly* 75:861–71.
- Gubrium J., Holstein J., Marvasti A., & McKinney K.D. (2012). *The SAGE Handbook of Interview Research: The Complexity of the Craft*, second edition. DOI:10.4135/9781452218403
- Halford S., & Savage M. (2017). Speaking sociologically with big data: symphonic social science and the future for big data research. *Sociology*, 51(6), 1132–1148. <https://doi.org/10.1177/0038038517698639>
- Halldén O., Haglund L., and Strömdahl H. (2007). Conceptions and Contexts: On the Interpretation of Interview and Observational Data. *Educational Psychologist* 42 (1): 25–40. doi:10.1080/00461520709336916. [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
- Henning S. and Hovy B. (2011) Data sets on international migration. *International Migration Review* 45(4):980–985
- Iliadis A, Russo F. (2016). Critical data studies: An introduction. *Big Data & Society*. DOI: 10.1177/2053951716674238



- Kiley M., Mullins G. (2005). Supervisors' Conceptions of Research: What are they? *Scandinavian Journal of Educational Research* 49(3): 245-262. DOI: 10.1080/00313830500109550
- Kitchin R. (2014), The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences. *Big Data & Society*. DOI: 10.1177/2053951714528481 bds.sagepub.com
- Löfgren K. [kentlofgren]. (2013, May 19). Qualitative analysis of interview data: A step-by-step guide [Video file]. Retrieved from ([Qualitative analysis of interview data: A step-by-step guide for coding/indexing - YouTube](#))
- Massey D., Arango J., Hugo G., Kouaouci A., Pellegrino A., & Taylor J. (1993). Theories of International Migration: A Review and Appraisal. *Population and Development Review*, 19(3), 431-466. doi:10.2307/2938462
- McLeod S. A. (2020). Thomas Kuhn - science as a paradigm. *Simply Psychology*. <https://www.simplypsychology.org/Kuhn-Paradigm.html>
- Meissner F., & Taylor L. (2021). *How to Research Migration using Data Technologies? Re-visibilising Migration Information Infrastructures*. Unpublished Manuscript.
- Meyer J., Shanahan M., Laugksch R. (2005). Students' Conceptions of Research. I: A qualitative and quantitative analysis. *Scandinavian Journal of Educational Research* 49(3). DOI: 10.1080/00313830500109535
- Morozov E. (2013). To Save Everything, Click Here: The Folly of Technological Solutionism. *Public Affairs*
- Nachtwey O. and Seidl T. (2020) The Solutionist Ethic and the Spirit of Digital Capitalism. *SocArXiv*. pp. 1-51.
- OECD (2003). Quality Framework and Guidelines for Statistical Activities, Version 2003/1, <http://www.oecd.org/dataoecd/26/42/21688835.pdf> Google Scholar
- OECD (2019). Harnessing New Social Data for Effective Social Policy and Service Delivery <https://search.oecd.org/social/soc/Workshop-NewSocialData-16Oct2019-BackgroundNote.pdf>
- Pulse UG (2012). Big data for development: Challenges & opportunities. Nueva York, mayo: Naciones Unidas.
- Radermacher W J. (2019) Governing by the numbers. Statistical governance: Reflections on the future of official statistics in a digital and globalized society. *Statistical Journal of the IAOS*; 35(4): 519-5
- Ricciato F., Wirthmann A., Hahn M. (2020) Trusted Smart Statistics: How new data will change official statistics. Available from: [https://www.researchgate.net/publication/342311589 Trusted Smart Statistics How new data will change official statistics](https://www.researchgate.net/publication/342311589_Trusted_Smart_Statistics_How_new_data_will_change_official_statistics) (accessed 04 June 2021).
- Salganik M. J. (2018). *Bit by bit: Social research in the digital age*. Princeton: Princeton University Press.
- Sandvik K. and Raymond N. (2017) *Beyond the Protective Effect: Towards a Theory of Harm for Information Communication Technologies in Mass Atrocity Response*, *Genocide Studies and Prevention*, 11/1: 9-24.
- Savage M., & Burrows R. (2007). The Coming Crisis of Empirical Sociology. *Sociology*, 41(5), 885-899. <https://doi.org/10.1177/0038038507080443>
- Scheel S. (2013) *Autonomy of migration despite its securitisation? Facing the terms and conditions of biometric rebordering*. *Millennium* 41(3):575-600
- Singleton A. (2016) *Migration and asylum data for policy-making in the European Union. The Problem with Numbers*. Bruss. CEPS Pap. Lib. Secur. Eur.
- Spyrtatos S., Vespe M., Natale F., Weber I., Zagheni E., Rango M. (2019). Quantifying international human mobility patterns using Facebook Network data. *PLoS ONE* 14(10): e0224134. DOI: 10.1371/journal.pone.0224134
- Spyrtatos S., Vespe M., Natale F., Iacus S.M., Santamaria C. (2020) *Explaining the travelling behaviour of migrants using Facebook audience estimates*. *PLoS ONE* 15(9): e0238947. <https://doi.org/10.1371/journal.pone.0238947>
- Taylor L. (2017). *What is data justice? The case for connecting digital rights and freedoms globally*. *Big Data & Society*. <https://doi.org/10.1177/2053951717736335>

- Taylor L., & Meissner F. (2020). *A Crisis of Opportunity: Market-Making, Big Data, and the Consolidation of Migration as Risk*. *Antipode*, 52(1), 270–290.
- Thrift N.J. (2005): *Knowing Capitalism*. *SAGE Publications Ltd*.
- Tourangeau R., Edwards B., & Johnson T. P. (2014). *Hard-to-survey populations*. Cambridge: Cambridge University Press.
- United Nations Global Pulse (2017) *Social Media and Forced Displacement: Big Data Analytics and Machine-Learning*. UN Global Pulse and UNHCR Innovation Service
- United Nations Population Fund (UNFPA) and U.S. Census Bureau (2019) *Measuring Migration in a Census - Select Topics in International Censuses*. *United Nations Publications*, New York.
- Weinhardt M. (2020). Ethical Issues in the Use of Big Data for Social Research. *Historical Social Research* 45 (3): 342–368. doi: 10.12759/hsr.45.2020.3.342-368
- Young, H Peyton (1996). The Economics of Convention. *Journal of Economic Perspectives*, 10 (2): 105–122. DOI: 10.1257/jep.10.2.105
- Zagheni E., Weber I., and Gummadi K. (2017), Leveraging Facebook’s Advertising Platform to Monitor Stocks of Migrants, *Population and Development Review*, Vol. 43/4, pp. 721– 734, <http://dx.doi.org/10.1111/padr.12102>