InGRID

Supporting expertise in inclusive growth

www.inclusivegrowth.eu

Deliverable 6.11

SPOTLIGHT REPORT BIG DATA AND WORK 2.0

Zachary Kilhoffer

May 2021

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 730998

Abstract

In recent years, big data has become a transformative tool for labour research. Big data can supplement official data sources like labour force surveys, which often fail to adequately represent vulnerable people. This report builds on desk research and two workshops (Special Interest Groups) to discuss the opportunities and challenges of applying big data in labour studies. Big data is more granular and timelier than traditional data sources, and researchers continue to discover new data sources and ways to use them. Novel studies use big data to research migration and non-standard employment, explore labour market supply and demand with online job advertisements, and audit online advertisements for potential discrimination against protected groups. Despite its potential, big data presents theoretical, methodological, and practical challenges for researchers. Many challenges relate to non-probability sampling techniques, which require special attention to correct for biases. Big data can be very noisy, difficult to obtain, and problematic for data privacy. Labour researchers ought to leverage big data and remain aware of its limitations. This requires careful selection and consideration of data sources, and investment in requisite skills and technology.

Acknowledgements

I would like to thank Simon Lenau, Ralf Münnich, and Karolien Lenaerts for their helpful review of this report. I would also like to thank all those who organised, presented, and attended both workshops: 'Use of non-probability surveys in a modern information society' and 'Big data and labour markets: Understanding precariousness'.

This report constitutes Deliverable 6.11, for Work Package 6 of the InGRID-2 project.

May 2021

General contact: inclusive.growth@kuleuven.be
p.a. InGRID
HIVA - Research Institute for Work and Society
Parkstraat 47 box 5300, 3000 LEUVEN, Belgium

For more information zachary.kilhoffer@ceps.eu

Please refer to this publication as follows:
Kilhoffer, Z. (2021). Spotlight report big data and work 2.0, Deliverable 6.11,Leuven, InGRID-2 project 730998 – H2020

This publication is part of the InGRID-2 project, this project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 730998.

The information and views set out in this paper are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

InGRID
Supporting expertise in inclusive growth

Contents

1. Introduction

Big data has become one of the most transformative tools for empirical research in a number of domains. It is ubiquitous within articles and reports by information technology researchers, and increasingly widespread in a variety of disciplines such as sociology, medicine, management, and economics (De Mauro et al., 2016).

While big data is closely tied to information technology, its expansion into new fields merits a closer look. In recent years, big data, and the types of analysis required to use it, have become increasingly relevant in labour economics for two overarching reasons. First, big data (alongside other trends connected to digitalisation) changes the way that people work. Second, big data changes the way researchers study the world of work. This report is focused on the second point, attempting to provide an overview of how big data has, and will continue to, impact labour market research. It builds on desk research and two workshops - Special Interest Groups (SIGs) - held in 2020 as part of the InGRID-2 project.[1]

The remainder of this paper begins with background, discussing the emerging developments of big data. I then discuss the application of big data in the field of labour economics, with particular focus on precariousness in the labour market. Next, I present a brief look at some of the challenges of big data, both generally and specific to labour economics. I then discuss a selection of novel applications of data science, before drawing conclusions and recommendations for future research.

This report does not aim to describe any particular big data source or methodology in great detail. Rather, it attempts to give a selection of topics where big data is changing the way that labour market studies are performed, and especially where it changes our understandings on people facing precarious situations in the labour market.

[1] These SIGs were called ‘Use of non-probability surveys in a modern information society’ and ‘Big data and labour markets: Understanding precariousness’. For more information, see https://www.inclusivegrowth.eu/special-interest-groups.

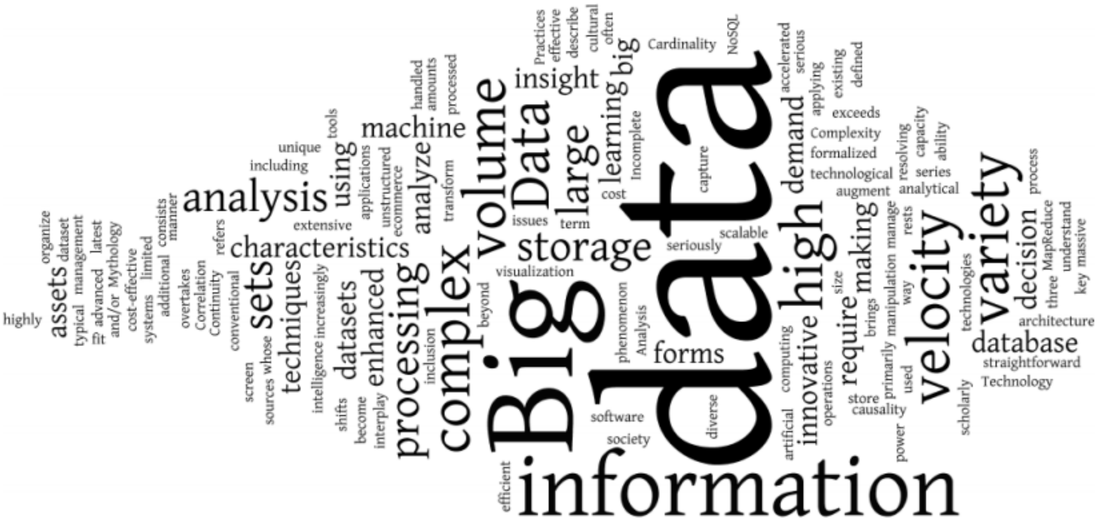
2. Background

Big data is a buzz word and subject to a variety of meanings. Given the different understandings of big data, it is useful to clarify what it means with a more formal definition (De Mauro et al., 2016: p. 1):

> *Big Data is the Information asset characterised by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value.*

This definition was built on a systematic literature review of publications concerning big data. Based on the 1,437 abstracts, the most prevalent words can be visualised as follows.

Figure 1. Key words appearing in abstracts of papers related to big data



Source De Mauro et al., 2016: p. 2

Incidentally, Figure 1 was created using a key process connected with big data – digitisation of analogue information into digital, then searching for patterns. It is a representation of big data, which is itself built from big data.

For De Mauro et al., the essential components of big data are understood as:

1. information;
2. technology;
3. methods;
4. impact.

Perhaps the first sources of **information** in big data involved mass digitisation (Coyle, 2006), such as the Google Books Library Project. The project began in 2004 and aimed to digitise more than 15 million printed books in several large college libraries. By digitising all this information, the process of 'datafication' became possible. For example, the corpus created by the Google Books Library Project was converted into sequences of contiguous words, or n-grams, and it became possible to analyse the occurrence of n-grams over centuries. Researchers in a number of social science fields used Google Books' datasets to generate new insights (Michel et al., 2011). The digitisation of physical books formed an important step in the development of big data, and it also hints at the **close relationship between big data and digital libraries**, **and big data and the field of library and information science**. New sources of information include networks of physical devices containing sensors, software, and other technologies, commonly known as the Internet of Things (IoT) (Atziori et al., 2010).

The **technology** component of big data is essential because of the generation, storage, and computational requirements of large amounts of data. Moore's law suggests that the number of transistors that fit on a silicon chip doubles every 18-24 months, and implies that data storage grows exponentially (Moore, 1965). New frameworks were designed to deal with the computational requirements of big data. For example, HDFS (Hadoop Distributed File System) allows multiple machines located anywhere to cooperate on a single computational task (Shvachko et al., 2010). Cloud computing offerings like Amazon Web Services (AWS), Google Cloud Platform, Microsoft Azure, Alibaba Cloud, and others offer a variety of technology services required in every step of research concerning big data. Finally, communication networks allowing larger and faster transfer are essential to facilitate the movement and computation of big data.

Methods refer to the new techniques required for processing big data, which tend to be more complex than traditional statistical strategies. The most frequently identified and discussed methods include (Chen et al., 2012; Manyika et al., 2011):

- natural language processing;
- machine learning;
- neural networks;
- predictive modelling;
- regression models;
- social network analysis;
- sentiment analysis;
- signal processing;
- data visualisation.

Not all of these are new additions – notably regression models and data visualisation are core components of more traditional economics. However, even these have evolved a great deal as big data progressed.[2]

Impact refers to the ways that the utilisation and management of big data impact society. Among the more pressing concerns, impact concerns the way that big data is harnessed for value creation, used by organisations, and impacts individual privacy. Moreover, the accessibility of big data modulates its impact and contribution to innovation, driving concerns about anticompetitive business practices, and the creation of a new digital divide among companies, driven by differing levels of access to data (Boyd & Crawford, 2012).

One point of note is that big data can be grouped into two main categories: data generated from 'online activities', and administrative data. Administrative data is generally described as data derived from the operation of administrative systems, such as data collected by government agencies for the purposes of registration, transaction, and record keeping (Elias, 2014). While administrative data is

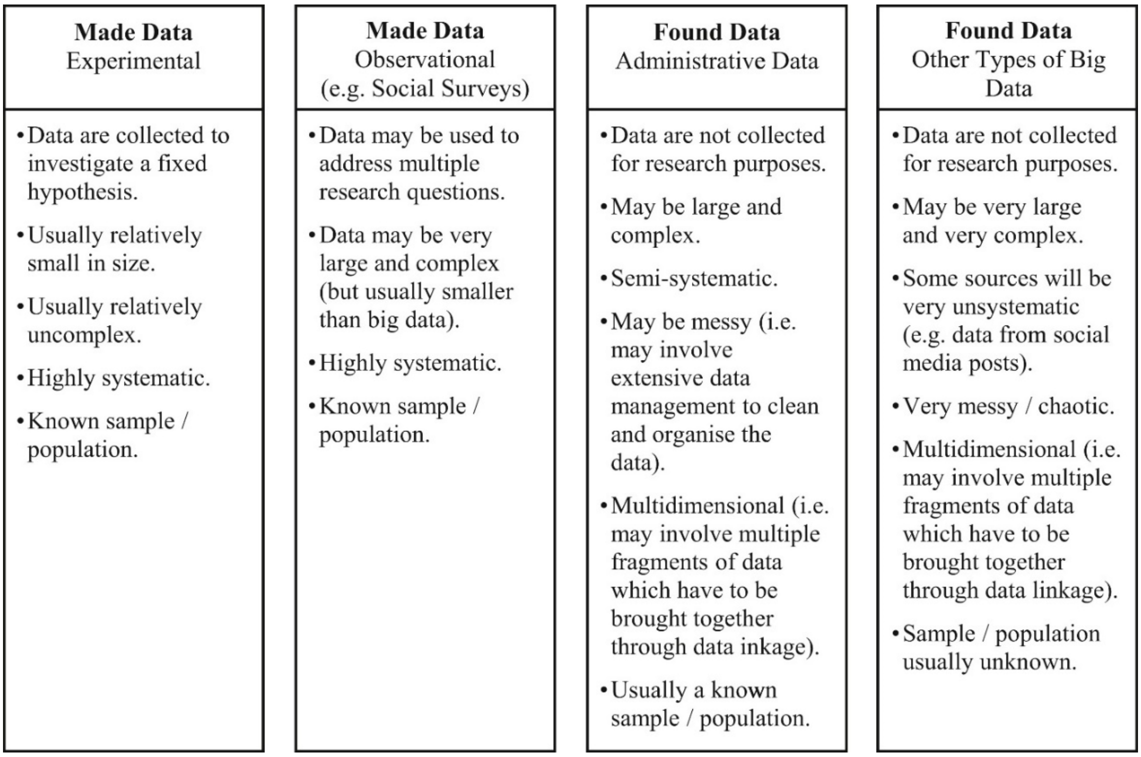
[2] See discussion below on linear regression techniques.

often neglected in mainstream discussions of big data, it may be particularly valuable to addressing questions in the social sciences, especially regarding social inequality (Connelly et al., 2016).

A further point of clarification is that big data should not be considered synonymous with data collected through the internet. Big data can result from commercial transactions, sensors (i.e. satellite and GPS), genome data, and administrative data such as education records, medical records, and tax records (Connelly et al., 2016).

In Figure 2, the two left-most pillars show traditional data sources, while the right-most pillars show administrative data and other types of big data. Some of the key distinctions are that big data are not collected for research purposes, may not be systematic, tend to be messy, tend to involve multi-dimensional data linked together, and may not be a known sample or population. Administrative data, compared to other forms of big data, tend to be more systematic and represent more specific populations.

Figure 2. Data sources in social science research



Made Data Experimental	**Made Data** Observational (e.g. Social Surveys)	**Found Data** Administrative Data	**Found Data** Other Types of Big Data
• Data are collected to investigate a fixed hypothesis. • Usually relatively small in size. • Usually relatively uncomplex. • Highly systematic. • Known sample / population.	• Data may be used to address multiple research questions. • Data may be very large and complex (but usually smaller than big data). • Highly systematic. • Known sample / population.	• Data are not collected for research purposes. • May be large and complex. • Semi-systematic. • May be messy (i.e. may involve extensive data management to clean and organise the data). • Multidimensional (i.e. may involve multiple fragments of data which have to be brought together through data inkage). • Usually a known sample / population.	• Data are not collected for research purposes. • May be very large and very complex. • Some sources will be very unsystematic (e.g. data from social media posts). • Very messy / chaotic. • Multidimensional (i.e. may involve multiple fragments of data which have to be brought together through data linkage). • Sample / population usually unknown.

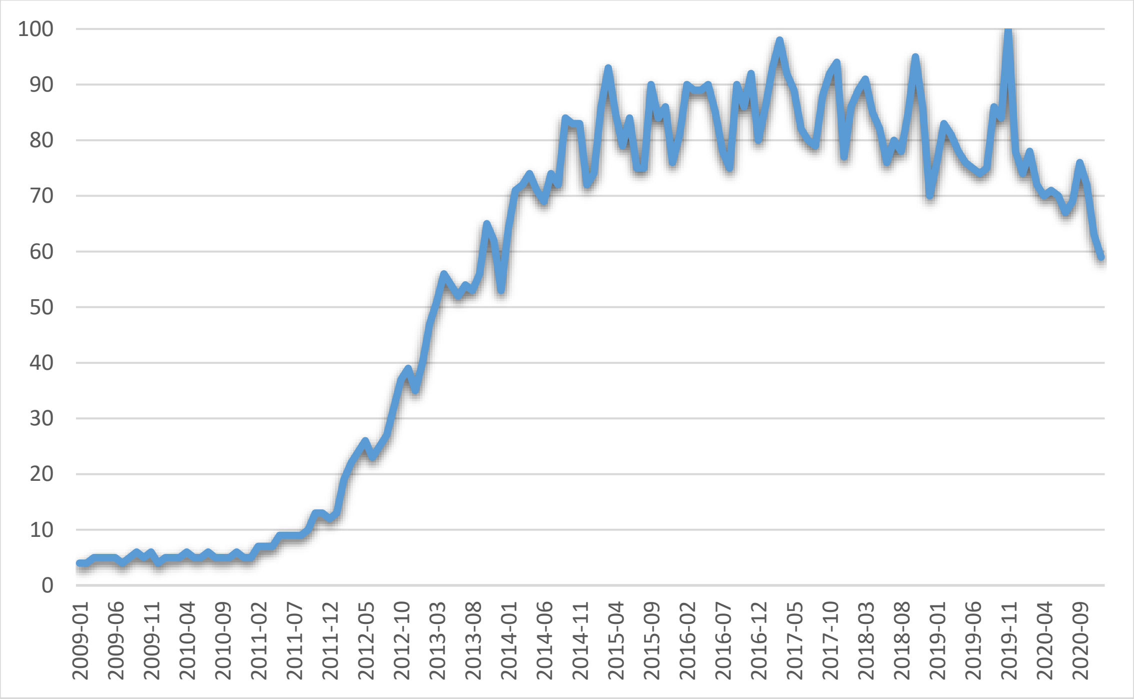
Source Connelly et al. (2016)

One unambiguous point is the growth in big data and its prevalence. The amount of raw data being produced daily has grown astronomically. According to the World Economic Forum, the accumulated digital universe of data was 4.4 zettabytes[3] in 2013, and within seven years, had grown ten-fold to 44 zettabytes in 2020. By 2025, this figure should nearly quintuple to 163 zettabytes (Reinsel et al., 2018).

The growth in data has coincided with growing interest in big data and the field of data science – essentially a hybrid of traditional statistics and computer science. Clearly, the supply and demand for big data analysis is on an upwards trajectory.

[3] One zettabyte is $1,000^7$ bits, or $1,000^4$ gigabytes.

Figure 3. Google searches for 'big data' worldwide



Note: 100 is equal to the maximum number of relative searches.

Source Google Trends (accessed 6 January 2021).

3. Big data and labour research

Big data is of interest to labour researchers due to the ‘granular, population-level data with multiple dimensions that allow researchers to analyse cases along many variables’ (Taylor et al., 2014: p. 5). Due to the IoT and other sources, observational data sets are now available that are much larger and of much higher frequency than traditional surveys (Harding & Hersh, 2018). This allows big data to act as a supplement or substitute to other data sources, notably survey data from governments, such as the Labour Force Survey at EU-level. Furthermore, big data allows social scientists to test certain research questions and hypotheses previously limited to theory. By using huge datasets, and searching them for signals of correlation, researchers can find ‘bits of gold’ in a sea of sand.

The apparent first paper using big data and economics came from Ettredge et al. (2005), on US unemployment rate (Choi & Varian, 2012). Ettredge et al. (2005) assumed that people’s internet behaviour reveals useful information about them. In this instance, job-related information gathering - such as looking for online job portals - could predict the unemployment rate in subsequent weeks. While the study was quite small in scale, it demonstrated that web searches may have more explanatory power than official unemployment insurance claims data in modelling unemployment rate.

In subsequent years, researchers used big data to explore other macroeconomic issues such as inflation. Several papers were published in 2009-2010 using Google search data to predict/forecast unemployment. Choi and Varian (2012), for example, found that adding relevant Google Trends variables to simple seasonal AR models outperforms models excluding these variables from 5 to 20%. These reports demonstrated the potential for big data to supplement traditional data sources, strengthening the predictive power of models.

In the past decade or so, big data has found many more applications in labour studies. The data sources have expanded beyond search data, and the research questions explored have grown in number and specificity. The following section discusses a few such applications.

4. Selected applications of big data in labour research

The advantages of big data can enable better research on economically vulnerable people. This is especially the case when considering people who may not be well-represented by traditional data sources. Examples include migrants, NEETs, and people in non-standard employment relationships[4] - especially casual or informal workers, domestic workers, or platform workers.

Based on the presentations of the SIG 'Big data and labour markets: Understanding precariousness', hosted as part of this report, and additional literature, a selection of applications are discussed below: migration, non-standard employment, online job boards, and online advertisements. The first two represent demographics where traditional data sources are often insufficient for policymakers.[5] Online job boards and online ads are two sources where big data offers important new insights into job opportunity and discrimination.[6]

Afterwards, under Challenges of big data for labour researchers, findings from the SIG 'Workshop on non-probability samples' and additional literature inform a discussion of where big data still poses challenges for social science researchers.

4.1 Migration

Mobility and labour migration are ideas at the core of the EU. Mobility primarily refers to workers posted abroad and cross-border commuters, whether within a Member State or between them (Eurofound, 2021), whereas labour migration means the movement of persons from one state to another, or within their own country of residence, for the purpose of employment (International Organization for Migration, 2015).

The migration crisis (or refugee crisis) in 2015 and 2016 exposed significant flaws in the EU's asylum policy (European Parliament, 2017). While migration has become a less significant issue for Europeans' voting decisions, it remains an important topic where traditional data sources are insufficient to guide policymaking.

The traditional sources of population censuses, household surveys, labour force surveys, administrative sources, and other statistical sources primarily measure accumulated entry or immigration visas, accumulated permission to work in a country, estimated stocks of undocumented foreign citizens, recruitment costs, and remittances. However, Bircan notes that 'coherence, consistency and comparability in national and international migration statistics may still be the exception rather than the standard' (Bircan et al., 2020: p. 1). In spite of multiple attempts by national governments, international and regional organisations, and private institutions to improve data collection on migration, the data suffer from inconsistencies in data collection methodology, a lack of adequate statistics, and limited comparability.

To work towards comprehensive, accurate, and timely data on migration, and improve policymakers' ability to create migration policy, big data can play an important role. For example, Böhme et al. (2020) used Google Trends Index (GTI) for migration-related search terms to measure migra-

[4] Non-standard employment refers to anything other than indefinite (open-ended) and full-time employment between a natural person (employee) and a natural or legal person (employer) (Eurofound, 2017).

[5] Derived from the presentations of Tuba Bircan and Fabian Stephany, InGRID-2 Webinar, 26 October 2020.

[6] Derived from the presentations of Fabio Mercorio and Sara Kingsley, InGRID-2 Webinar, 26 October 2020.

tion from a certain country and predict subsequent emigration. This strategy allowed for more short-term predictions ahead of official data releases, which may take several years. Ultimately, the models utilising GTI performed better than those without.

Others have used geolocation data from Twitter and Facebook to analyse movement within and between countries, and produce estimates for stocks of EU movers and EU mobility by Member State (Zagheni et al., 2014; Gendronneau et al., 2019). Still another strategy is to use mortality data to extrapolate hidden populations, including asylum seekers and undocumented migrants (Houttekier et al., 2011). This technique may be especially useful as mortality is such a stable and consistent demographic parameter, and marginalised populations can be difficult to pinpoint with other data sources.[7]

One particularly large and ambitious project is from Horizon2020 and called the HumMingBird Project: Enhanced Migration Measures from a Multidimensional Perspective.[8] HumMingBird covers 10 European countries and consists of 16 partners: researchers in a number of disciplines from different research institutes and universities, SMEs, NGO networks, and a European Research Infrastructure Consortium (ERIC). HumMingBird utilises air-traffic data, airline top-up transfer, remittances, mobile recharge records, and other sources to find insights.

One of the novelties of HumMingBird is cooperation with the private sector. For example, researchers cooperated with Turkey's primary telecommunications operator, Turkcell, to receive an exhaustive dataset of all phone-based activity in 2019 for a subset of the population, including refugee and migrant groups. These data were anonymised and aggregated in accordance with the Data For Development Challenge and the Data for Refugees Challenge setups (Bircan et al., 2020). The granularity of these data are remarkable. Antenna traffic data allows a full year of precise location data three times per day for difficult to reach groups, including Syrian and Afghan refugees. These data allow not only for observation of migration from Turkey, but also creating profiles for different regions, and categorising migrants' behaviour within those regions.

4.2 Non-standard employment

Non-standard workers can be difficult to capture with normal labour force surveys for a number of reasons. For example, those who work sporadically may not be working at the moment the survey occurs. The Eurostat Labour Force Survey lacks categories for certain individuals, like platform workers or temporary agency workers. More generally, such non-standard or informal workers may be very difficult to define and categorise.

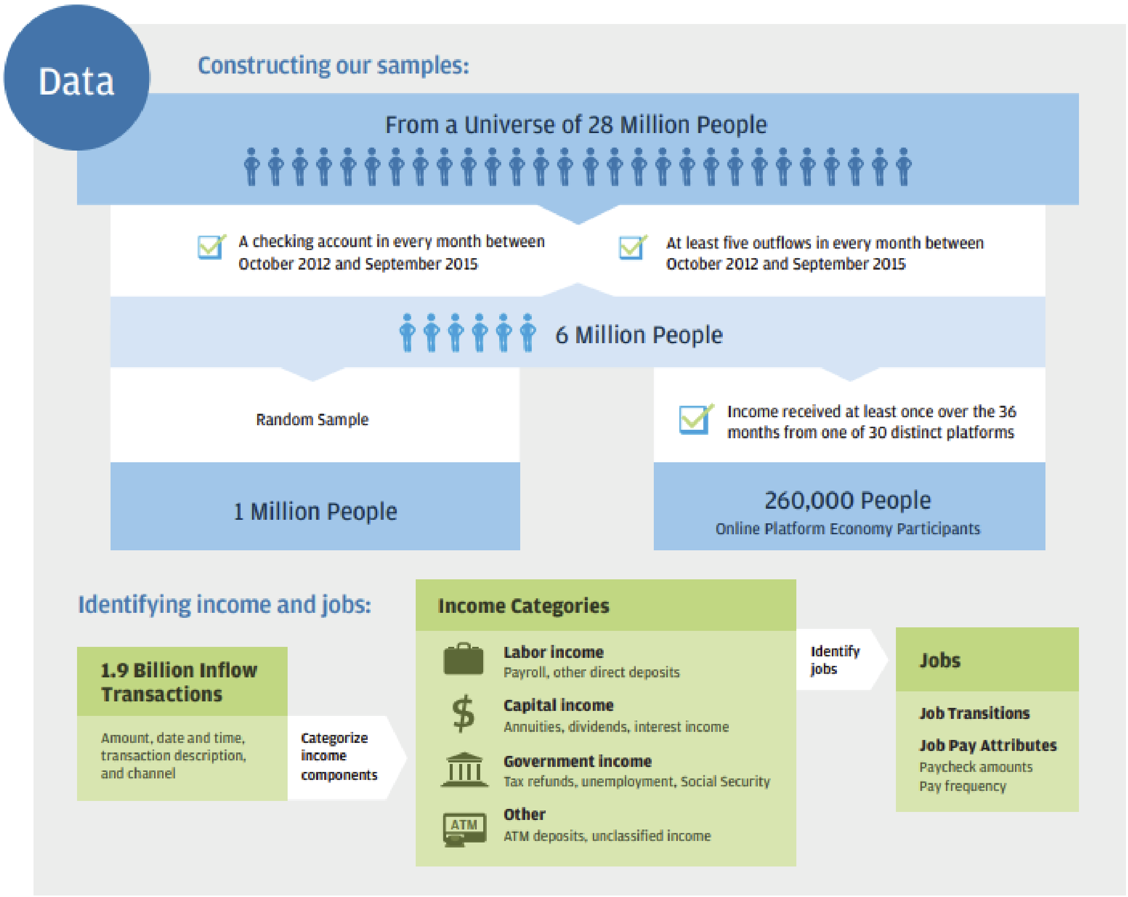
Nevertheless, many non-standard workers have certain socioeconomic vulnerabilities, such as difficulty accessing social protection and steady income. Non-standard workers often face an intersection of vulnerabilities, such as being young and of migrant background. This points to the need for better understanding of such individuals.

One example where big data has been of help is in measuring platform workers, and how platform work interacts with income volatility. As a first example, Farell and Greig made use of administrative data of 1 million customers from JPMorgan Chase, America's largest bank, between October 2012 and September 2015 (2016). These data were high-frequency and from a randomised, anonymised sample, as shown in Figure 4.

[7] See also the MISAFIR project, available at https://interfacedemography.be/project/misafir/.

[8] See current publications here: https://hummingbird-h2020.eu/publications.

Figure 4. Estimating platform workers with JPMorgan Chase administrative data


Data
Constructing our samples:
From a Universe of 28 Million People
A checking account in every month between October 2012 and September 2015
At least five outflows in every month between October 2012 and September 2015
6 Million People
Random Sample
Income received at least once over the 36 months from one of 30 distinct platforms
1 Million People
260,000 People
Online Platform Economy Participants
Identifying income and jobs:
1.9 Billion Inflow Transactions
Amount, date and time, transaction description, and channel
Categorize income components
Income Categories
Labor income
Payroll, other direct deposits
Capital income
Annuities, dividends, interest income
Government income
Tax refunds, unemployment, Social Security
Other
ATM deposits, unclassified income
Identify jobs
Jobs
Job Transitions
Job Pay Attributes
Paycheck amounts
Pay frequency

Source Farrell and Greig (2016: p. 2)

In addition to estimating the number of American platform workers, the study found that income volatility is most prevalent among youth, with more than 70% of people aged 18-24 experiencing more than a 30% month-to-month change in total income. Findings also suggest that platform work helps offset dips in other earnings, as well as when people are between jobs. These results provided empirical evidence not only on the scale of platform work, but also on its potential role in improving income security for people struggling with low or sporadic income, or unemployment.

An additional example using data from platform sources is from the Oxford Internet Institute, which set up the Online Labour Index (OLI). The OLI tracks all tasks posted to the five largest English-language OLPs, representing at least 60% of the market by traffic. The Oxford Internet Institute has reached an agreement with these platforms to gather data through API calls or periodically webcrawl vacancies. The result is an easy-to-use tool to view supply of tasks by country, time period, type of occupation, and growth trends (Kässi & Lehdonvirta, 2018). For example, the OLI generally indicates overall growth in total tasks around 25% annually, but found a temporary slump in March-April 2020 associated with the COVID-19 pandemic. The OLI illustrates the value of big data provided in near real-time, showing how non-standard workers are impacted by greater economic trends.

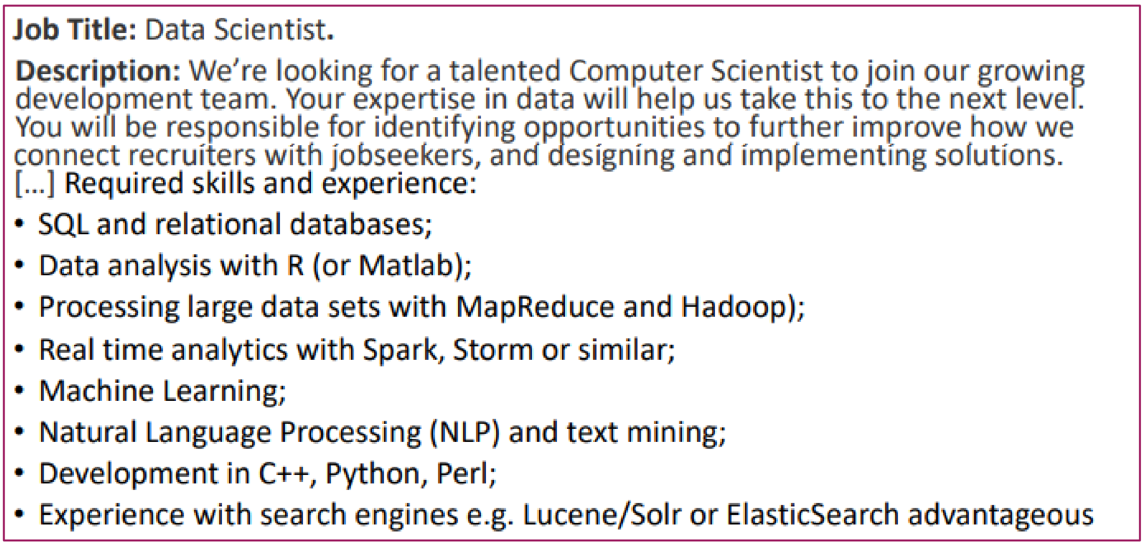
Furthermore, the OLI gives a great amount of detail on the types of tasks requested, and what skills are required to perform them. In certain respects, this is much more detailed than the Eurostat Labour Force Survey, which does not capture the content of work or skill requirements beyond formal degree requirements.

4.3 Online job boards

Online job boards (i.e. Indeed or Monster.com) have grown to address both the supply and demand sides of job search. Already in 2014, Carnevale et al. found that more than 80% of jobs requiring a Bachelor's degree or higher, and between 60% and 70% of all job openings, were posted online (2014). These numbers are expected to have grown substantially since then, but still lean towards higher-skill occupations.

The data in job ads are comparable to those related to the OLI, discussed above. However, rather than describing a discrete task, job ads typically contain information such as the job title, location, and description of responsibilities, as shown in Figure 5. For this reason, job ads can describe which jobs are demanded in which quantity, as well as provide detailed descriptions of job content, which can help anticipate skills demand. Furthermore, many of the most prominent online job boards are publicly accessible, creating a huge and timely resource to better understand job supply.

Figure 5. Example of online job ad



Job Title: Data Scientist.

Description: We're looking for a talented Computer Scientist to join our growing development team. Your expertise in data will help us take this to the next level. You will be responsible for identifying opportunities to further improve how we connect recruiters with jobseekers, and designing and implementing solutions. […] Required skills and experience:

- SQL and relational databases;
- Data analysis with R (or Matlab);
- Processing large data sets with MapReduce and Hadoop);
- Real time analytics with Spark, Storm or similar;
- Machine Learning;
- Natural Language Processing (NLP) and text mining;
- Development in C++, Python, Perl;
- Experience with search engines e.g. Lucene/Solr or ElasticSearch advantageous

Source Presentation of Fabio Mercorio, InGRID-2 Webinar, 26 October 2020

On the other hand, taking unprocessed raw data like the job ad shown, and transforming it into useful insights on emerging occupations and skills, requires an intensive process of data transformation and cleaning, classification, and extraction, largely driven by natural language processing and other forms of machine learning. This can take place with two general strategies. The simpler approach uses existing thesauri or ontologies (i.e. WordNet), but this is domain and language dependent, and results can only be as good as the reference taxonomy. The second approach uses co-occurrences (n-grams) of words to discover terms that are likely to occur together. While the insights may be better and more novel, the resulting data are noisier, a human expert is required to review the AI's suggestions, and the process is more computationally expensive.[9]

Several EU-level initiatives aim to leverage big data for labour market analysis by retrieving online job data via webscraping or API calls, with examples including Eurostat and Cedefop (2021-2024), Horizon2020 (2020-ongoing), and Cedefop II (2016-2021). These projects have a number of goals, including new insights into: (1) occupational and skill discovery; (2) soft, digital, and hard skill rates; (3) newly emerging occupations and skills; and (4) extending existing occupation taxonomies (e.g. ISCO, ESCO, O-NET).

[9] Derived from the Presentation of Fabio Mercorio, InGRID-2 Webinar, 26 October 2020.

At present, these programmes are not primarily focused on learning about vulnerability in labour markets. However, minimal adjustments would allow finding additional information, such as the types of contracts offered.[10] This could help understand who is most at risk of non-standard employment contracts. Additionally, insights into skills demand can help to empower educators to provide in-demand skills to future job applicants, reducing the risk of job mismatch, low-income, and unemployment.

4.4 Online advertisements and discrimination

While online job boards serve as a type of advertisement, the more conventional online advertisements can also help understand vulnerability in the labour market. A few particular types of advertising (e.g. jobs, employment, credit, and housing) can be strongly linked to economic opportunity, making them a good starting point for labour market researchers to identify potential discrimination.

As an illustration of this idea, in some jurisdictions, laws seek to prevent discrimination against certain protected groups. In the US, these include discrimination by gender, age, and race/ethnicity for job and credit advertisements. Discrimination need not be explicit. For example, targeting by zip code can be a proxy for race, given how people of different ethnicities in the US are not evenly distributed, but often quite segregated.

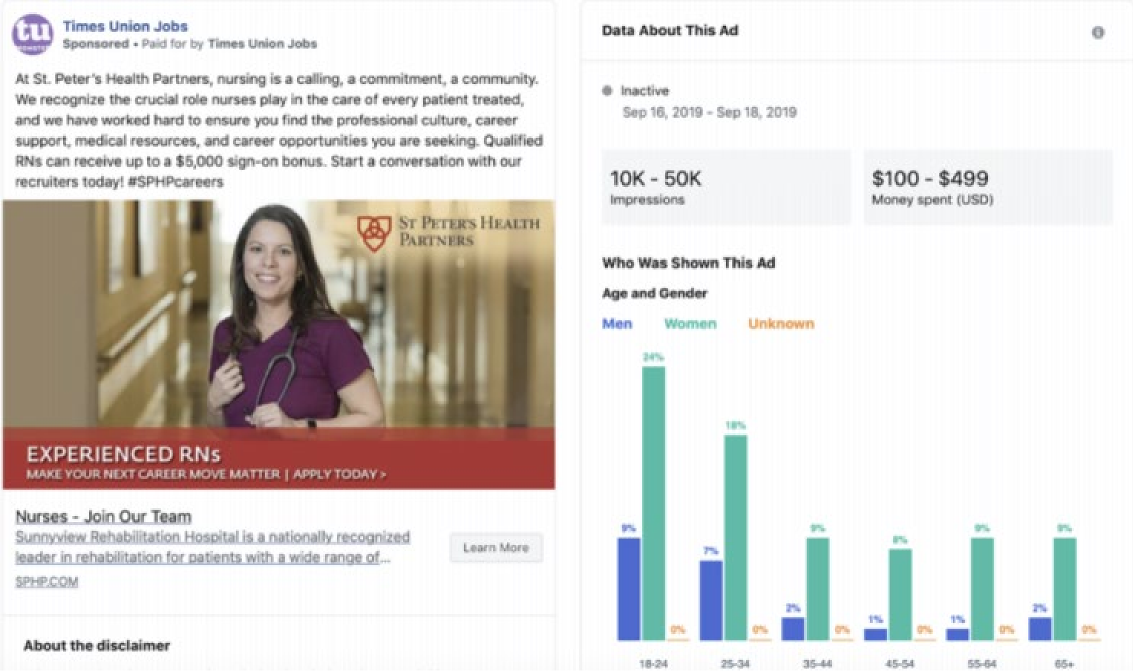
A 2019 lawsuit against Facebook sought to restrict the ability of advertisers to target or exclude protected demographic groups (Jenner & Block, 2019). As a result of the settlement, 'special ad categories' for credit, employment, housing, social issues, elections, or politics cannot target individuals by gender, age, or zip code (Facebook, 2019). However, similar to how zip code can be a proxy for race, advertisers can quite precisely target or exclude protected groups in new ways. In particular, Facebook's Audience Insights Tool enables advertisers to select demographics by devices used. Bilingual Hispanics living in the US are much less likely to access Facebook on their desktop or iPad, for example, meaning ads on these mediums are very rarely shown to this group.[11]

Some researchers have begun auditing big data from social media advertisements to better understand how gender and race discrimination can occur (Kingsley et al., 2020), as is currently underway using Facebook's advertisement API. Initial results of the audit of Facebook's ad API shows that advertisements may distribute economic opportunity on the basis of stereotypes. Furthermore, discrimination can also occur against transgender and non-binary people, as Facebook allows personal profiles many more options than just men and women, and classifies those who are not labelled men or women as 'unknown' for the sake of advertising outreach. For example, some job advertisements for registered nurses were shown to many more women than men, and no transgender or non-binary people (Figure 6). Similarly, the audit found one sheriff's office placed an ad hiring for several positions, but the ad was shown exclusively to men.

[10] Derived from the Presentation of Fabio Mercorio, InGRID-2 Webinar, 26 October 2020.

[11] Derived from the Presentation of Sara Kingsley, InGRID-2 Webinar, 26 October 2020.

Figure 6. Facebook job ad shown primarily to women



Times Union Jobs
Sponsored • Paid for by Times Union Jobs

At St. Peter's Health Partners, nursing is a calling, a commitment, a community. We recognize the crucial role nurses play in the care of every patient treated, and we have worked hard to ensure you find the professional culture, career support, medical resources, and career opportunities you are seeking. Qualified RNs can receive up to a $5,000 sign-on bonus. Start a conversation with our recruiters today! #SPHPcareers

Nurses - Join Our Team
Sunnyview Rehabilitation Hospital is a nationally recognized leader in rehabilitation for patients with a wide range of...
SPHP.COM

About the disclaimer

Data About This Ad

● Inactive
Sep 16, 2019 - Sep 18, 2019

10K - 50K
Impressions

$100 - $499
Money spent (USD)

Who Was Shown This Ad

Age and Gender

Source Presentation of Sara Kingsley, InGRID-2 Webinar, 26 October 2020

At least two issues are relevant here. First, online platforms may empower advertisers to place ads in a manner that discriminates against protected groups. Second, online platforms use algorithms that may entrench existing biases, like which jobs are more suitable for women or men, and which race is more likely to take advantage of access to a line of credit for a house or car. This suggests that additional research is necessary to audit online advertising, ensuring that both the design of the platform, and the outcomes of the platform's algorithms, do not further discrimination and unfairness in socio-economic opportunity.

5. Challenges of big data for labour research

Up to this point, a number of benefits of big data in labour research have been discussed, while the challenges have been mostly glossed over. This section aims to break these challenges down into succinct categories, then briefly present them.

5.1 Theoretical

One theoretical challenge concerns data mining, a process between computer science and statistics, involves extracting and discovering patterns in big data. The advent of data mining means some economists are developing research questions based on the dataset, rather than seeking out data to answer research questions. Economists may suspect a dataset has great analytical value without being able to say in advance what that is. This diverges from the more traditional hypothesis-led approach in economics, but may yield unexpected and useful insights.

Relatedly, interpreting findings from big data often relies heavily on human experts of a particular domain. **Noisy data requires more human intervention** to separate the signal from the noise, so to speak. One example was already mentioned above in the discussion of online job boards. Suggestions for new occupations, or updates to existing occupational categories, cannot be left exclusively to AI algorithms. Instead, a human with domain and language expertise must check suggestions for plausibility. As few people are experts on so many levels, this necessitates an interdisciplinary research approach, where people with the technical and quantitative skills to design AI need to work with thematic experts. How best to use ‘human-in-the-loop’ systems is a topic of ongoing research (Li, 2017).

A further fundamental challenge is the **reliability of big data**. Administrative data, much less online data, are not collected for research purposes, but rather result as by-products of other political, societal, or economical processes. Therefore, the formation and quality such data is typically less known and controlled than for more traditional sources of data used for research.

As one practical example, Elliot (2015) performed data analysis on unemployment rates in Cambridge, England, using a non-probability sample[12] from novel administrative data. As part of her research, she unexpectedly found that a very affluent area had very high unemployment rates. Upon speaking to data collectors, she realised that many homeless individuals had been recorded at the government employment agency in this area, resulting in very misleading figures. This shows how being critical of the data collection process is essential to avoid erroneous findings, and leads to major methodological challenges to be considered in context of big data.

5.2 Methodological

In general, application of statistical or machine learning methods on big data sets can yield results that do not at all represent 'true' values or relationships in the population. While classical sources of research data, such as experiments and probability samples, provide protection against systematic errors and allow to quantify errors that arise by chance, this is not possible for most types of big data.

[12] See discussion below.

Therefore, big data often require a very different methodological approach than traditional data sources. There are a number of reasons for this, but in terms of statistical methodology, the main issue with big data sources is that they typically result from non-probability sampling. Throughout the last century, the gold standard of statistics has been probability sampling, in which case a) every unit in the population has a chance of being selected in the sample, and b) this probability can be accurately determined. Non-probability sampling violates these assumptions (Vehovar et al., 2016), and is therefore prone to selection bias and limited generalisability from a sample to the population.

Especially in the context of data retrieved from the internet, including web surveys and big data, non-probability sampling is becoming more common. The reasons for the increased use of non-probability samples includes the need for higher survey response rates, as non-response rates continue to increase in all modes of survey administration,[13] and preference for cheaper survey modes (e.g. internet versus phone).[14] While more readily available, non-probability samples from surveys create a degree of self-selection bias. Self-selection bias can be partially addressed in online surveys in several ways. For example, respondents can be recruited from a random sample of a population register or address list, or randomly selected from a panel. While preferable to online surveys with self-selection, these strategies are slower, more cumbersome, and more expensive.[15]

It is often difficult to determine how reliable big data are, especially when derived from online sources. For example, some studies have used LinkedIn data to estimate international migration of workers (Bogdan et al., 2014; Barslund & Busse, 2016). However, users of LinkedIn skew higher-skilled, making this approach more appropriate for higher-skilled workers and IT-related positions, rather than the general workforce. If the intention is to know more only about the population sampled, rather than the general population, then there is little issue. However, most often researchers are interested in generalising results. Similarly, big data from many sources tends to over-represent people who have the financial means to afford smart phones, computers, and the internet, as well as those who are younger and more digitally literate. This can be a particular challenge when trying to perform analysis for people with labour market vulnerabilities.

This highlights the importance of critically reflecting the data collection process to avoid erroneous findings. Despite covering a large number of observations, big data typically do not allow for well-established approaches to make valid generalisations to whole populations. Therefore, understanding potential errors inherent to these data is crucial to judge the reliability and quality of information obtained from them. In general, these issues are not compensated by the sheer size of big data, because systematic errors do not decrease with sample size. 'Compensating for quality with quantity is a doomed game' (Meng, 2018). Therefore, big data increases the need for researchers to be aware of the limitations and pitfalls of classical methods for estimation and inference in case of non-probability sampling.

A key question is if, and under which conditions, non-probability samples can be used to obtain unbiased results.[16] Unlike probability sampling, where valid estimation and inference is possible by design of the sampling procedure, non-probability sampling requires assumptions and available auxiliary information that can be used to model the selection process and/or relevant characteristics of interest, such as the migration of workers outlined above.[17] Therefore, the usability of non-probability samples heavily depends on the validity of assumptions and the availability of suitable auxiliary information. To obtain such auxiliary information, data harmonisation between different big data and traditional data sources may be required. Data linkage may be a significant issue generally, but especially in many countries that lack consistent identification numbers across administrative systems (Connelly et al., 2016).

[13] Derived from the Presentation of Emilia Rocco and Alessandra Pettrucci, InGRID-2 Webinar, 27-28 February 2020.

[14] Derived from the Presentation of Danny Pfeffermann and Arie Preminger, InGRID-2 Webinar, 27-28 February 2020.

[15] Derived from the Presentation of Jelke Bethlehem, InGRID-2 Webinar, 27-28 February 2020.

[16] Derived from the Presentation of Danny Pfeffermann and Arie Preminger, InGRID-2 Webinar, 27-28 February 2020.

[17] Derived from the Presentation of Natalie Shlomo and Ton De Waal, InGRID-2 Webinar, 27-28 February 2020.

5.3 Practical

Using big data is contingent on accessing them, which can pose a significant challenge. Most big data are proprietary, namely owned by companies, and especially large multinationals. In many cases, potentially useful data will not be shared with researchers because data represents a competitive advantage for businesses. In other cases, data will only be shared on a very limited basis, or only with researchers working for the company. This can jeopardise the neutrality of findings.[18]

Just taking the examples from sections above, JPMorgan Chase's data is proprietary, Facebook ad's API is subject to change and access can be restricted, Turkcom's data is proprietary, and webscraping or API access of online job boards can be restricted. For data retrieved from online sources, the website may alter data unexpectedly, further challenging any efforts to replicate and confirm results.

In the case of administrative data from government sources, researchers will be working under a strict set of conditions determined by the data owners (Connelly et al., 2016). This means that most researchers will not be able to access potentially useful data, which not only prevents some research from taking place, but also makes it impossible to replicate findings. Beyond businesses protecting valuable data, they may not be accessible due to privacy concerns as well. Even if big data are accessible, their use may still entail significant privacy concerns (Horton & Tambe, 2015).

Furthermore, storing and processing big data can entail large computational costs, so analysing big data requires investments in technology (Harding & Hersh, 2018). While hardware and cloud services are becoming cheaper, this can still restrict access for researchers and institutes without suitable financial means.

Finally, processing big data requires significant skill and expertise, requiring investments in human resources (Harding & Hersh, 2018). In many cases, domain experts in the social sciences may not be able to perform analyses themselves, even if they have a reasonably strong background in more traditional statistics and econometrics. Social scientists may therefore require additional training required for data management, like restructuring data to achieve variable by case matrix required for most analysis techniques, and being able to run and interpret different AI techniques. Certain machine learning methods require a great deal of fine tuning for optimal performance, which places further strain on computational requirements (Harding & Hersh, 2018). Traditionally, economists use regression to fit models to data. However, big data may require different techniques such as random forest, least absolute selection and shrinkage operator (LASSO), or deep neural networks. These and other applications of machine learning are becoming more widespread as they regularly perform better than standard econometric methods (ibid.).

[18] For example, see Berg and Johnston's critique of Hall and Johnston's labour market analysis for Uber (2019).

6. Conclusions

This paper has argued that big data can be an excellent supplement to existing data sources, offering much more granular information collected on a shorter timeframe. Big data can offer better insights into vulnerable populations who may not be well captured by traditional sources.

Big data may be especially promising in certain fields where traditional data sources, like official surveys, are often insufficient. Migration and non-standard employment are two such areas. Various types of big data are being used to better understand these areas, and such efforts should continue to be honed. Additionally, online data from job boards and online advertising have great potential to generate new insights. Job board data is already being used to model supply and demand in the labour market. This includes timely findings on what jobs are in demand, where, and what skills are required. As online job boards become more ubiquitous, it is likely that job boards will only increase in utility. Online advertisement data is particularly interesting to better understand discrimination of vulnerable populations, such as how evenly opportunities are distributed. Even as digitalisation offers huge new opportunities, it also risks retrenching old inequalities, or creating new ones.

While online forms of big data are very much in vogue, administrative data should not be overlooked. In many cases, administrative data are the easiest type of big data for researchers to implement, being more systematic than big data from online sources. Government tax records, education records, neighbourhood characteristics, immigrant landing records, birth cohorts, etc., can be especially useful for research on vulnerable populations.

Despite these promising points, using big data for social science is associated with theoretical, methodological, and practical challenges. These are not insurmountable, but require very careful consideration to reach valid, useful, and generalisable conclusions.

In particular, aspects of data quality, such as the potential selectivity and bias, must be carefully considered. Methods that are designed for probability samples will likely fall short in accounting for errors that arise from non-probability samples, and may heavily exaggerate the accuracy of results. Dedicated methods for non-probability samples can - under certain conditions - provide more adequate results for point estimation and inference. Such methods are particularly important as many big datasets, and more recent survey data, increasingly represent non-probability samples.

Finally, big data demands multidisciplinary expertise and creativity. To the extent possible, interested social science researchers should become acquainted with new techniques required for big data. To aid this process, research institutes should make appropriate investments in technology and human capital. At the same time, theoretical and substantive knowledge of social science topics is irreplaceable, and often times must be implemented alongside AI processes.

References

Atziori, L. et al. (2010). The internet of things: A survey, *Computer Networks*, *4247*.

Barslund, M., & Busse, M. (2016). How mobile is tech talent? A case study of IT professionals based on data from LinkedIn, CEPS, June (https://www.ceps.eu/ceps-publications/how-mobile-tech-talent-case-study-it-professionals-based-data-linkedin/).

Berg, J., & Johnston, H. (2019). Too Good to Be True? A Comment on Hall and Krueger's Analysis of the Labor Market for Uber's Driver-Partners, *ILR Review*, *72*(1), 39-68.

Bircan, T. et al. (2020). Review of Migration Theories and the Quality and Compatibility of Migration data on the National and International Level.

Bogdan, S. et al. (2014). Migration of professionals to the us evidence from LinkedIn data.

Böhme, M.H. et al. (2020). Searching for a better life: Predicting international migration with online search key-words, *Journal of Development Economics*, *142*, p. 102347.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon, *Information, communication & society*, *15*(5), 662–679.

Carnevale, A.P. et al. (2014). Understanding online job ads data, Georgetown University, Center on Education and the Workforce, Technical Report (April).

Chen, H. et al. (2012). Business intelligence and analytics: From big data to big impact, *MIS quarterly*, pp. 1165-1188.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends', *Economic record*, *88*, 2–9.

Connelly, R. et al. (2016). The role of administrative data in the big data revolution in social science research, *Social Science Research*, *59*, 1-12.

Coyle, K. (2006). Mass digitization of books, *The Journal of Academic Librarianship*, *32*(6), 641–645.

De Mauro, A. et al. (2016). A formal definition of Big Data based on its essential features, *Library Review*, *65*(3), 122–135.

Elias, P. (2014). Administrative data, Scivero Verlag, Berlin.

Elliott, J. (2015). Advancing the Administrative Data Research Network: Next Steps for Facilitating Excellent Research.

Ettredge et al. (2005). Using Web-based search data to predict macroeconomic statistics, *Commun. ACM*, *48*, 87.

Eurofound (2017). Aspects of non-standard employment in Europe (https://www.eurofound.europa.eu/publications/customised-report/2017/aspects-of-non-standard-employment-in-europe).

Eurofound (2021). Migration and mobility (https://www.eurofound.europa.eu/fr/topic/migration-and-mobility).

European Parliament (2017). 'Asylum and migration in the EU: facts and figures', June (https://www.europarl.europa.eu/news/en/headlines/society/20170629STO78630/asylum-and-migration-in-the-eu-facts-and-figures).

Facebook (2019). Updates To Housing, Employment and Credit Ads in Ads Manager (https://www.facebook.com/business/news/updates-to-housing-employment-and-credit-ads-in-ads-manager).

Farrell, D., & Greig, F. (2016). Paychecks, paydays, and the online platform economy.

Gendronneau, C. et al. (2019). Measuring Labour Mobility and Migration Using Big Data.

Harding, M., & Hersh, J. (2018). Big Data in economics, *IZA World of Labor* (https://wol.iza.org/articles/big-data-in-economics/long).

Horton, J.J., & Tambe, P. (2015). Labor Economists Get Their Microscope: Big Data and Labor Market Analysis', *Big Data*, 3(3), 130–137.

Houttekier, D. et al. (2011). Study of recent and future trends in place of death in Belgium using death certificate data: A shift from hospitals to care homes, *BMC public health*, *11*, 228.

International Organization for Migration (2015). Key Migration Terms (https://www.iom.int/key-migration-terms).

Jenner & Block (2019). HUD Brings Housing Discrimination Charge Against Facebook (https://jenner.com/library/posts/19022).

Kässi, O., & Lehdonvirta, V. (2018). Online labour index: Measuring the online gig economy for policy and research, *Technological Forecasting and Social Change* (https://linkinghub.elsevier.com/retrieve/pii/S0040162518301331).

Kingsley, S. et al. (2020). Auditing Digital Platforms for Discrimination in Economic Opportunity Advertising, *arXiv:2008.09656 [cs]* (http://arxiv.org/abs/2008.09656).

Li, G. (2017). Human-in-the-loop data integration, *Proceedings of the VLDB Endowment*, *10*(12), 2006–2017.

Manyika, J. et al. (2011). Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election, *Annals of Applied Statistics*, *12*(2), 685–726.

Michel, J.-B. et al. (2011). 'Quantitative analysis of culture using millions of digitized books', *science*, *331*(6014), 176–182.

Moore, G.E. (1965). *Cramming more components onto integrated circuits*, McGraw-Hill New York, NY, USA.

Reinsel, D. et al. (2018). The Digitization of the World from Edge to Core, p. 28.

Shvachko, K. et al. (2010). The hadoop distributed file system.

Taylor, L. et al. (2014). Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same?, *Big Data & Society*, *1*(2), p. 2053951714536877.

Vehovar, V. et al. (2016). Non-probability sampling, *The Sage handbook of survey methods*, pp. 329–345.

Zagheni, E. et al. (2014). Inferring international and internal migration patterns from twitter data.

InGRID-2
Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy

Referring to the increasingly challenging EU2020-ambitions of Inclusive Growth, the objectives of the InGRID-2 project are to advance the integration and innovation of distributed social sciences research infrastructures (RI) on 'poverty, living conditions and social policies' as well as on 'working conditions, vulnerability and labour policies'. InGRID-2 will extend transnational on-site and virtual access, organise mutual learning and discussions of innovations, and improve data services and facilities of comparative research. The focus areas are (a) integrated and harmonised data, (b) links between policy and practice, and (c) indicator-building tools.

Lead users are social scientist involved in comparative research to provide new evidence for European policy innovations. Key science actors and their stakeholders are coupled in the consortium to provide expert services to users of comparative research infrastructures by investing in collaborative efforts to better integrate microdata, identify new ways of collecting data, establish and improve harmonised classification tools, extend available policy databases, optimise statistical quality, and set-up micro-simulation environments and indicator-building tools as important means of valorisation. Helping scientists to enhance their expertise from data to policy is the advanced mission of InGRID-2. A new research portal will be the gateway to this European science infrastructure.

This project is supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 730998.

More detailed information is available on the website: www.inclusivegrowth.eu

Co-ordinator
Monique Ramioul

KU LEUVEN HIVA
RESEARCH INSTITUTE FOR WORK AND SOCIETY

Partners

TÁRKI Social Research Institute Inc. (HU)
Amsterdam Institute for Advanced Labour Studies – AIAS, University of Amsterdam (NL)
Swedish Institute for Social Research - SOFI, Stockholm University (SE)
Economic and Social Statistics Department, Trier University (DE)
Centre for Demographic Studies – CED, University Autonoma of Barcelona (ES)
Luxembourg Institute of Socio-Economic Research – LISER (LU)
Herman Deleeck Centre for Social Policy – CSB, University of Antwerp (BE)
Institute for Social and Economic Research - ISER, University of Essex (UK)
German Institute for Economic Research – DIW (DE)
Centre for Employment and Work Studies – CEET, National Conservatory of Arts and Crafts (FR)
Centre for European Policy Studies – CEPS (BE)
Department of Economics and Management, University of Pisa (IT)
Department of Social Statistics and Demography – SOTON, University of Southampton (UK)
Luxembourg Income Study – LIS, asbl (LU)
School of Social Sciences, University of Manchester (UK)
Central European Labour Studies Institute – CELSI (SK)
Panteion University of Social and Political Sciences (GR)
Central Institute for Labour Protection – CIOP, National Research Institute (PL)

InGRID-2

Integrating Research Infrastructure for
European expertise on Inclusive Growth from
data to policy Contract N° 730998

For further information about the InGRID-2
project, please contact
inclusive.growth@kuleuven.be
www.inclusivegrowth.eu
p/a HIVA – Research Institute
for Work and Society
Parkstraat 47 box 5300
3000 Leuven
Belgium

The views expressed during the execution of the InGRID-2 project, in whatever form and or by whatever medium, are the sole responsibility of the authors. The European Union is not liable for any use that may be made of the information contained therein.