4. The Information Problem: Big Data as a Solution for Labour Market Analysis

Jeisson Cárdenas Rubio
4. The Information Problem: Big Data as a Solution for Labour Market Analysis
4.1. Introduction

“More and better data” is a common claim of researchers and policymakers as a prerequisite to design public policies such as tackling skill mismatch issues (Cedefop 2010; OECD 2017b; Williams 2004). Information collection about labour demand through surveys involves statisticians, interviewers, and a sample of companies or individuals available and willing to respond. The cost of this kind of projects is relatively high, in terms of resources and time, and can discourage countries (especially with low budgets) from collecting and analysing vacancy data. Additionally, even if a survey is carried out, the information obtained might not be detailed enough to analyse which skills or occupations are in demand among different industries or regions (Handel 2012; OECD 2016d).

Currently, with the proliferation of the internet and higher-capacity electronic devices, large amounts of information about the behaviour of different agents are being stored daily. The storage of all this information has unlocked new territories for research in different areas of knowledge. For instance, Edelman (2012) and Askitas and Zimmermann (2015) detail several research examples using Big Data that have provided different applications for research in micro- and macroeconomics, labour and demographic economics, public economics, health, education, and welfare, among others.

Big Data may be a way to overcome the limitations of existing skills analysis. More specifically, online job portals are a promising source of valuable information about labour demand. Thus, the second section of this chapter defines Big Data, followed by an analysis of how Big Data might fill information gaps in labour demand and supply to address labour market policies and research. The fourth section discusses the potential uses of information from job portals to tackle skill mismatches (skill shortages). Big Data in specific job portals has certain limitations and, for this reason, the fifth section discusses these limitations and indicates some caveats when using this kind of data.
for analysing the labour market. Finally, the chapter describes how Big Data sources might facilitate the analysis of the labour market in a context such as the Colombian economy.

4.2. A definition of Big Data

Higher internet speed, a broader use of smartphones, tablets, cameras, computers, etc., and technologies with increasing capacities to store information have favoured the creation and storage of computerised or digital information on a large scale. Cisco (an important multinational technology conglomerate) estimates that 96 exabytes (1 EB = $10^{18}$ bytes) were the average monthly data traffic in 2016 and it is expected to increase three times by 2021 (278 EB per month) (see Figure 4.1). This era of massive information has unlocked opportunities for private and public institutions to compile, link, and analyse relatively large flows of data produced by different sources in order to better orientate important decisions and strategies. This set of massive information, including the techniques used to process and analyse the available information, is commonly labelled as “Big Data.”

Figure 4.1. IP traffic by source, 2016-2021

[Source: Cisco (2017, p. 6).]
However, there is still an extensive debate about what can or cannot be considered as Big Data. Perhaps one of the most common conventions defines this term according to three properties: volume, variety, and velocity (Laney 2001). Each of these properties will be discussed in turn. The first one refers to the most obvious property to be considered as Big Data: the size (or volume) of data. In a simple way, data with a large volume of information might be a candidate to be called Big Data. However, individuals might consider different volumes of data differently because there are different computer capacities available in the market (with more or less data storage capacity, processing, etc.), which allow people to handle a certain number of bits per second. Consequently, it is necessary to determine a standard threshold to classify data according to their size. One way to do this is by classifying data whose size represents a challenge to be processed and analysed within the average range of computer technologies available as “big.”

Nevertheless, it is important to keep in mind that the threshold to determine whether data have a high volume of information might change over time. Average computer capabilities increase over time; as technology improves, so does its capacity to process a high volume of information. Hence, what was considered as Big Data when this research started might get altered by the time this book is finished. Despite the changing nature of data, this criterion is still useful because volume allows researchers to distinguish between data sources in a technological environment that is constantly changing.

Variety refers to data structure. Unlike the information that comes from surveys, information from Big Data might not possess a well-defined structure to organise different variables in specific spaces (columns) within a database. Instead, the information might come from a range of unstructured or semi-structured sources and in different formats, such as social media, sensors, websites, mobiles, videos, etc. (Aguilar 2016). This characteristic makes data processing a challenge. Algorithms need to be developed to identify patterns (such as tags, keywords, among others) to obtain meaningful information. Thus, it is essential to consider that the concept of Big Data is not just related to volume; this concept also includes complex data qualities that make it necessary to have access to a higher capacity to store, process, and analyse the gathered information.
Finally, velocity refers to the speed at which data are generated. Nowadays, information is generated in seconds; people can share an opinion to thousands through platforms such as Twitter or Facebook and generate different reactions in an instant. Likewise, companies can post their current vacancies in real-time on various websites to quickly attract potential workers. This speed presents a challenge and an advantage for data processing and data analysis.37

For the purpose of this book, “Big Data” is considered as a relatively high volume of information, which is produced in a relatively fast way by different unstructured or semi-structured sources, and might be available in diverse formats, where the above mentioned three characteristics of volume, variety, and velocity make information processing and analysis processes a challenge per se with average technologies available in this given moment (in 2017).38

Despite many challenges, Big Data is expanding or opening a new frontier of knowledge (Askitas and Zimmermann 2015; Edelman 2012). Indeed, Big Data might fill information gaps that exist in different fields and regions where information to carry out well-oriented public policies was frequently scarce in the past (Azzone 2018). In the particular case of the labour market in Colombia, this information might give insights about the characteristics of the labour supply; and, more importantly, due to the general scarcity of labour demand data (especially in countries like Colombia), Big Data offers the possibility of having, for the first time, a detailed picture of employer requirements in real-time. The following section discusses in more detail how Big Data has provided new valuable information to analyse the labour market in different areas.

37 There are cases where information is not quickly generated (e.g. on a daily basis); nevertheless, they (e.g. medical records) might be considered as Big Data given the size of the database, which overpasses the current average computer capabilities.

38 The debate about what constitutes Big Data is still open. Özköse, Ari and Gencer (2015) or BBVA (2018) add other characteristics such as “veracity” and “value” to the Big Data concept. The former refers to the trustworthiness or credibility of data, although this is an implicit characteristic that any data should have; while the latter term establishes that information needs to provide some profit (usually measured in terms of money) to a certain institution. Nonetheless, not every institution or person seeks monetary profit from information. For instance, non-profit institutions might benefit from Big Data information in order to provide goods and/or services for free or at prices that are not economically competitive. Moreover, the value of information depends on the observer: data that might not produce any value for a company might hold some value for a researcher or a different institution.
4.3. Big Data on the labour market

“Good” data are a requisite to develop well-oriented policies and academic research, where “good” refers to data that involve analysing the representativeness of the population and, thus, a minimum standard of quality during the collection process. Supply and demand information from surveys has limitations that Big Data might alleviate in order to form a better picture of the labour market. Thus, different efforts have started to be developed from both the supply and demand sides in some countries and areas, which involve the usage of Big Data.

4.3.1. Labour supply

4.3.1.1. Household surveys for the analysis of the labour supply

Traditionally, on the labour supply side, information has been collected from household surveys (e.g. employment rates by age, region, gender, etc.). Generally, these household surveys are characterised by a sampling frame (based on a census) that is representative of a specific population, a set of questions, and flows that customise the sections participants complete. Such surveys collect the main characteristics of the labour supply over a certain period. In most cases, the surveys are carried out by the Office for National Statistics (ONS) of each country, which follow certain quality standards provided by an international institution such as the ILO. These standard procedures make household surveys one of the main sources of information to calculate indicators of the labour market, such as participation and unemployment rates, wages, etc.

Despite its indisputable advantages, household survey information has some limitations that might be overcome with Big Data. First, information collection through surveys requires time for design, validation, collection, and consolidation, among other processes, that might delay the publication of the resulting database for analysis. When data are available, the researcher needs time to process the information, toanalyse possible alternatives, or to address specific issues. However, an important disadvantage of such methods is that some time will elapse from the moment when the survey is designed until the
completion of the final database, and during this time the analysis of data might become outdated and invalidate the research findings due to changes in the socio-economic environment. Indeed, Reimsbach-Kounatze (2015) highlights that many OECD countries only have access to labour supply information several weeks (at best) after the data were collected.

Second, another limitation of household surveys is their fixed structure as a pre-designed questionnaire, which collects information on a variety of topics from people for various monitoring, planning, and policy purposes. Surveys also have budget and time constraints. For instance, the UK Labour Force Survey (LFS) aims to measure “economic activity and inactivity, all aspects of people’s work, job-search for the unemployed, education and training, income from work and benefits” (ONS 2015). Clearly, variables that are beyond this scope are not measured. Moreover, adding one single question increases the survey’s cost and might also affect its structure, flux, response rate, and results. This makes it difficult for survey designers to include other relevant labour supply questions. Thus, household surveys are a rigid tool that attempt to measure social issues, whose dimensions might change over time.

Third, due to sample constraints, household surveys have a statistical limitation. The more the data are disaggregated (e.g. region, sector, age, education, etc.), the more imprecise are the estimates. For instance, the GEIH survey has available labour market results, such as employment or unemployment shares, disaggregated by city and SIC (Standard Industrial Classification, revision 3). This information is useful to analyse unemployment rates by region, major occupational groups, etc; nevertheless, the level of detailed information (granularity) obtained from household surveys might not be sufficient to cover topics, which might be particularly useful for institutions and individuals (e.g. sector employment composition, the skills possessed by individuals, and occupations).

Additionally, household surveys (and, in general, other kinds of surveys) are not exempt from limitations, such as measurement errors (the difference between a measured quantity and its true value), issues when collecting the information (e.g. interviewees might provide imprecise or false information or interviewers can make mistakes when recording the data, etc.). Thus, as stated above, household surveys have important limitations: 1) a time lag between designing the survey, data collecting and processing, and analysing the results; 2) a fixed structure that makes it difficult to include or modify questions and
update categories; 3) a design to be representative for a certain population at a disaggregated level; and 4) other potential limitations such as measurement errors and issues in data collection.

Consequently, several variables of interest for policymakers and researchers are not provided by household surveys. Such is the case of job networking, among other job seeker behaviours. Therefore, although household surveys are one of the main sources of labour supply information, there exists relevant uncovered information, which might be provided by using Big Data information.

4.3.1.2. Big Data and labour supply

So far, the contribution of Big Data information to knowledge about labour supply has come from two sources. The first source uses search engines such as Google Insights, and the second source uses social media and networking sites to monitor (over a relatively short period) the behaviour of job seekers. Regarding the former, search engines track millions of searches in real time concerning different topics, such as weather, news, products, and, importantly for this document, job searches. Consequently, these word searches can be used to identify trends in people’s behaviour. For instance, Askitas and Zimmermann (2009, p. 6) found that the usage of certain keywords—for example, “unemployment office or agency,” “unemployment rate,” among others—by German people on Google has a strong correlation with, and therefore is a useful predictor of, the unemployment rate in Germany. The underpinning idea is that people will use certain words related to job searches in Google when they are (or are likely to be) fired, or when it is difficult to find a job. Thus, access to people’s searches on these kinds of search engines can provide information before the results of official surveys are available.

Regarding the latter, social media and networking sites might be a source of labour supply information. Specialised social media platforms and websites, such as LinkedIn and BranchOut, have arisen in the last decade. For instance, LinkedIn is one of the most well-known professional networks as it is present in more than 200 countries and has more than 552 million users (with around 250 million users active every month), who make their curriculum vitae public in order to be contacted or contact potential employers (LinkedIn, n.d.). The information available through these social media platforms might provide
insights about the skills and other characteristics of the labour supply (see, for instance, State et al. 2014).

Interestingly, information from social media platforms has helped researchers to build or further refine their employment indicators. Such is the case for Antenucci et al. (2014), who created indexes of job loss, job searches, and job postings in real-time by tracking keywords such as “lost job,” “laid off,” and “unemployment”, among others. Thus, regarding labour supply, social media and networking sites as well as search engines have created the opportunity for researchers to deepen their understanding in certain topics.

4.3.2. Labour demand

Perhaps the use of Big Data for labour demand analyses has raised higher expectations among researchers, policymakers, etc., than its use for labour supply. These expectations might be motivated by the fact that, traditionally, labour demand information has been scarcer than labour supply information (Kureková, Beblavy, and Thum, 2014). As explained in more detail in this section, labour demand information shares many of the same limitations as labour supply information, such as sample constraints and granularity. However, unlike labour supply information, labour demand surveys and the analysis of employer requirements tend to be less frequent (especially in countries like Colombia; see Chapter 3). Paradoxically, as Hamermesh (1996) emphasises, one reason that explains why studies about labour demand have been relatively ignored or scarce is that the “creation of large sets of microeconomic data based on household surveys has spurred and been spurred by development of new theoretical and econometric techniques for studying labour supply” (p. 6).

Consequently, the main sources of information used for the analysis of labour demand have come from sectoral surveys (such as industry surveys) or even from household surveys. Even though these data sources have strengths, such as national standardisation and representativeness, the collection of labour demand information through surveys is likely to be costly, both in terms of resources and time, and these surveys might not provide enough information to workers, governments, and other institutions about human resources needs.
4.3.2.1. Sectoral surveys

In the UK’s “Vacancy Survey,” carried out by the Office for National Statistics (ONS), around 6,000 trading businesses\(^{39}\) are interviewed monthly to provide “an accurate and comprehensive measure of the total number of vacancies across the economy and fills a gap in the information available regarding the demand for labour” (ONS, n.d.). The survey’s main results are published in the Labour Market Statistical Bulletin within six weeks of the reference date of the survey and reveal the monthly number of vacancies in the UK. Additionally, there is a time series available regarding the total number of vacancies (seasonally adjusted) by industry, which are aggregated (following SIC 2007 sections, 22 groups) by the size of businesses\(^{40}\) (ONS 2016a; ONS 2016b), and a time series comparison between the total number of vacancies and the total number of unemployed people (Beveridge curve) (ONS 2016c). Moreover, the UK Employer Skills Survey (carried out by the DfE) provides detailed information about job requirements; specifically, skills and occupations demanded by employers (at the four-digit SOC level if possible), and industries (22 major groups at a one-digit level according to SIC 2007). This survey is a biennial study, and its main results are published over the months following each survey (Vivian 2016).\(^{41}\)

Likewise, less developed regions like Colombia have made different efforts to collect and analyse labour demand. As mentioned in Chapter 3, Colombia has conducted sectoral surveys, such as the annual industrial survey and services survey. Despite these considerable efforts, these kinds of sectoral or cross-sectorial surveys (e.g. EFCH\(^{42}\)) present severe limitations for the analysis of labour demand and, consequently, skill mismatches. First, as the name suggests, sectoral surveys are applied for a specific sector. The EFCH survey

---

\(^{39}\) It excludes agriculture, forestry, and fishing.

\(^{40}\) 1-9 employed; 10-49 employed; 50-249 employed; 250-2,499 employed; and 2,500+ employed.

\(^{41}\) At the time this document was written, the last available report was published in 2015, and the results for the 2017 survey was set to be available in summer 2018 (ONS n.d.).

\(^{42}\) To carry out this survey took an investment of $397,349 (from the Ministry of Labour and the Interamerican Bank of Development, IDB), plus the DANE had been providing survey preparation in terms of sample designs, logistics, and advice since 2010 (CONPES 2010; DANE 2018c).
in Colombia is only applied to companies related to industrial services and sales-retail activities. Therefore, some sectors might be excluded, and their labour demand composition and dynamics will remain unknown. Second, not all types of companies are included in the sampling frame. The industrial EFCH survey interviews establishments with 10 or more employees and whose annual production is above £125,000. Moreover, the EFCH survey’s results are available according to “functional areas” such as “Production,” “Management,” and “R&D.” Thus, these sources might not be enough to provide detailed information about which skills (or occupations) are in demand among different industries or regions (Handel 2012; OECD 2016d).

Likewise, the Colombian annual industry survey, which is one of the main sources for labour demand information, interviews establishments with the same criteria than the EFCH. Indeed, the EFCH is a subsample of the annual industry survey sample. Consequently, many companies (generally small or medium-sized companies) in a sector might not be included in the sample, so even within the relevant subsample, part of the labour demand is ignored.

Perhaps more advanced regions such as the UK are less exposed to this aggregation problem. With a greater budget, these regions can design surveys with a higher disaggregation level, as is the case of the UK Employer Skills Survey mentioned above. Nevertheless, even with a larger budget, the results from industry surveys might be produced with a relatively low frequency. For instance, the main findings from the UK Employer Skills Survey are released every two years. Policymakers, educational institutions, and researchers, among others, need to wait at least two years to access the information collected by the survey about labour demand requirements. There is a long period between the time the survey is carried out and data are processed, cleaned and released; labour conditions might change during the two years it takes to prepare data reports, and consequently, some results might be outdated. Regarding this problem, less advanced countries like Colombia are in a worse situation. For instance, in 2018, at the time this section was written, the last EFCH survey had been conducted in 2012 (a period of 6 years).

In some industry surveys, companies or a group of experts are asked about the number of vacancies that opened in the last period (e.g. within the last year), the number of vacancies that each company is expected to have in the next period (e.g. within the next year), the expected volume, and some
general characteristics (e.g. experience) of people that they will need in a certain period of time (e.g. the following three months, six months, a year, etc.). By providing information about current and future labour demand dynamics, such questions address the problem of the low frequency of data results.

Based on this labour demand information, two different approaches have been developed to anticipate future labour market needs: skill forecast and skill foresight. The first term refers to forecast exercises that “use available information or gather new information with the specific aim of anticipating future skills needs, mismatches and/or shortages. Forecast results are meant to provide general indications about future trends in skill supply and/or demand in the labour market” (OECD 2016d, p. 39). The latter term, skill foresight, aims to “provide a framework for stakeholders to jointly think about future scenarios and actively shape policies to reach these scenarios” (OECD 2016d, p. 39). Both these exercises are valuable because they estimate future employer requirements and address the education and VET system according to possible future needs.

Nevertheless, once again, efforts like skill forecast and skill foresight are relatively expensive in terms of money and time, and their results are too specific to be of use to the broader labour market. For instance, in Colombia, prospective labour market studies focus on specific sectors, such as coffee production and building construction. Moreover, projections from skill foresights or skill forecasts might be biased or mistaken. For example, companies might experience sudden expansion (or contraction) periods, which can unexpectedly increase (or decrease) the creation (or destruction) of future vacancies. Thus, labour demand estimates might under- or overestimate the number of vacancies and their characteristics. Likewise, experts might not accurately predict the course of a sector over the long term. Additionally, parameters to make economic projections might be outdated. Consequently, projections based on these data would ignore economic changes that have occurred between 2005 (date of the most recent census available at the time this section was written) and the date when a new census is conducted, and economic projections are re-estimated.

Therefore, sectoral surveys and exercises derived from them have several limitations: 1) They require large logistical operations and a substantial amount of money to conduct a labour demand survey. Consequently, 2) considerable time is needed to design, collect, process, and release the information gathered. 3) Given budget constraints and survey designs, some companies or sectors
might be excluded from labour demand analysis. For the same reasons, 4) it is, frequently, unlikely to be able to disaggregate survey results at numerous levels: occupational, skills, industry, region, etc. Given these limitations, labour demand information is scarce and less frequent (e.g. monthly) than household surveys. Finally, 5) skill forecast or skill foresight methods might not properly foresee economic changes and their implications for skills (labour demand). Therefore, it is relatively common to find labour demand studies in the economic literature whose main sources of information are household surveys.

4.3.2.2. Household surveys for labour demand analysis

Traditionally, household surveys have functioned as inputs to analyse labour demand issues. These sources provide information about the intersection between labour supply and labour demand (filled labour demand) over a certain period. Household surveys provide data about labour demand in the following way: employed people can occupy one or more job vacancies, consequently, the total number of employed persons weighted by the number of jobs held by each one of them is equal to the total number of vacancies filled (satisfied demand; see Chapter 2).

This information about satisfied labour demand has been used in different studies as an approach to analyse labour demand dynamics. Moreover, the availability of a relatively long series of household data has allowed analysing relevant trends and changes of the (filled) labour demand (Acemoglu and Autor 2011; Autor and Dorn 2012; Autor, Katz, and Kearney 2006; Salvatori 2018).

However, to analyse the labour demand based on what people report on household surveys is limited. First, as explained above, survey constraints (e.g. money and time) might not allow disaggregating the results at a skill or occupational level (e.g. 4-digit level ISCO). Second, household surveys only take into account the current/past skills or characteristics of the workforce; what is unknown are employer requirements to fill their vacancies, which is an important aspect of labour demand to reduce possible mismatches (Autor 2001; Mavromaras et al. 2013); thus, the acquisition of information is based

43 For instance (as mentioned in Chapter 3), the World Bank has conducted a Skills Measurement Program to assess skills in low- and middle-income countries (Pierre et al. 2014).
on what people (labour supply) report, and does not consider an essential part of the labour market: employer requirements.

This issue is an important limitation when considering employment share as a proxy of the labour demand. Total employment is at the intersection between labour supply and demand. Nevertheless, the level of employment might significantly differ from the true level of demand because of unfilled labour demand (vacancies). For instance, employers might demand high-skilled jobs, but there is no labour supply to fill them; consequently, by only using total employment, the fact that there is an important demand for high-skilled workers would be ignored.

Therefore, household surveys are a valuable input to analyse filled labour demand and its long-term changes. Nevertheless, this information is limited in the following aspects: 1) there are constraints (e.g. time and money) that affect the level of aggregation and the frequency of data collection; 2) these surveys do not capture information about employer requirements, which is essential to address issues such as skill shortages. Consequently, all the problems mentioned above for sectorial and household surveys restrict the capacity of researchers and policymakers to tackle skill mismatches.

4.3.2.3. Big Data and labour demand

As previously mentioned, the collection of labour demand information is relatively less systematic than labour supply information. Moreover, even when labour demand information is available, different limitations make skill mismatch analysis a challenge. However, it seems that the proliferation of high-volume information (such as the internet) and techniques to analyse it have brought the opportunity to evaluate possible skill mismatches (skill shortages) through the analysis of employer requirements.

Nowadays, internet is an important source of information. This source is widely used for different purposes, and it stores relevant information regarding the behaviour of agents such as employers. As Autor (2001) highlights, the internet provides an opportunity to collect more and possibly better labour market data. Indeed, online information contains a large number of detailed observations about labour demand, and it can be accessed mostly in real time and at a relatively inexpensive cost (Barnichon 2010; Edelman 2012).
Moreover, the use of the internet by employers to advertise and find suitable applicants, and by individuals to find a job, has dramatically increased. As mentioned by Maurer and Liu (2007) and Smith (2015), both employers and job seekers have increasingly used the internet to find a vacancy or to advertise. In fact, by 2007, more than 110 million vacancies and 20 million unique resumes were stored in online US sources (Maurer and Liu 2007, p. 1). More recently, Kässi and Lehdonvirta (2018) suggest that the volume of online new vacancies has grown roughly 20% worldwide from 2016 to 2018. Likewise, the number of job seekers looking for a job using online sources has increased. For instance, in the US, the share of people who used the internet to find a job increased from 26% in 2000 to 54% in 2015 (Smith 2015).

The use of online job portals as a source of information has grown among researchers and has also attracted the attention of policymakers because they seem to provide quick and relatively inexpensive access in order to analyse information about employer requirements. Job portals are websites where companies make public their current (or future) vacancies. Companies describe, to some extent, the job position and the attributes that a potential worker should have in order to be considered as a candidate. Additionally, job seekers can screen and select vacancies, and contact potential employers. In other words, job portals help to connect employers with job seekers and vice versa.

Information from job portals, however, is not produced for the purpose of economic analysis (indeed, in most cases, it is posted online by private businesses). Yet job advertisements can potentially function as an essential input to analyse the employers’ needs. The systematic collection of information from job portals might help to diagnose the performance of an economy in real-time (e.g. at the level of available vacancies), and to understand employer requirements and how these requirements change over time. Consequently, along with the increasing usage of the internet and job portals, studies have used online job vacancy data to provide insights about the labour demand in different countries, such as the US, Slovakia, Czech Republic, and Colombia (Guataqui, Cárdenas, and Montaña 2014; Carnevale, Jayasundera, and Repnikov 2014; Marinescu and Wolthoff 2016; Štefánik 2012; Tijdens, Beblavy, and Thum-Thysen 2015).

In this sense, Kureková, Beblavy, and Thum (2014) have emphasised that job portals can be useful to generate a better understanding of company needs,
which might enrich labour market policies. In contrast with household surveys (filled demand), information from job portals (unfilled labour demand) might be useful to reveal what occupations or types of skills are currently in demand. Moreover, this kind of data might be of more relevance in contexts where employers experience difficulties to fill job vacancies, and information from job portals might be the only data available to analyse the labour demand for skills to address labour supply according to employer requirements. Consequently, in less advanced regions such as Latin America (e.g. Colombia), where the largest skill mismatches exist, there is a lack of labour demand information (see Chapter 3), and the usage of information from job portals to measure employer requirements might have a high impact on different labour demand outcomes.

4.4. Potential uses of information from job portals to tackle skill shortages

Targeted vacancy information gathered from online job portals might improve information and public policy deficiencies regarding skill shortage problems in the following ways: it allows 1) maintaining an estimation of vacancy levels, 2) identifying skills and other job requirements, 3) recognising new occupations or skills, and 4) updating occupation classifications.

4.4.1. Estimating vacancy levels

The number of job offers, together with other labour market indicators (such as unemployment levels), help to determine the business cycle and possible mismatches in an economy. High vacancy rates might mean that the economy is in a stage of economic expansion and/or there are mismatches between the supply and labour demand.44

In this sense, online job vacancy advertisements might provide real-time access to job offers in an economy, and public policymakers might react

---

44 An example of the above is the Beveridge curve, which relates unemployment and vacancy to determine how well, or not, vacancies match with unemployed workers (Blanchard and Diamond 1989) (see Chapter 9).
or re-design public policies in a shorter period aligned to current economic changes. Given the advantages of collecting online information, different countries have started to create job vacancy databases based on information from the internet. For instance, in the US, there is the Help Wanted Online Data Series created by The Conference Board (Conference Board n.d.), and in Australia, there is the Internet Vacancy Index developed by the Australian Department of Education, Employment, and Workplace Relations (DEEWR) (Australian Government 2018). Both provide measures of labour demand (advertised vacancies) at various levels, including at a national, state, regional, and occupational level (Reimsbach-Kounatze 2015).

Moreover, online job vacancy information is not limited to counting the number of job offers in the economy. Indeed, one of the most important advantages of online job vacancy advertisements is that they provide detailed information about employer requirements. This aspect allows researchers, policymakers, among others, to delve into topics (that previously were relatively difficult or costly to obtain updated information on) and to identify demand for skills and other job requirements.

4.4.2. Identifying skills and other job requirements

Perhaps one of the most promising uses of online vacancy information is the identification of job requirements in a relatively short time to enable public policy design. As will be seen in more detail in Chapter 5, companies post their job vacancies on job portals along with detailed candidate requirements to fill each position (skills, education, experience, etc.). This detailed information creates an opportunity to monitor job requirements at a disaggregated level (e.g. 4-digits occupation level) and, for instance, advise VET institutions regarding what skills they need to train people in to increase their employability.

In this sense, one of the most important ongoing projects, at the time this book was written, is the “Big Data analysis from online vacancies” project carried out by the European Centre for the Development of Vocational Training (Cedefop for its acronym in Spanish). The Cedefop combines its efforts with Eurostat and DG Employment, Social Affairs, and Inclusion to collect data on skills demand using online job portals. With this information, the Cedefop attempts to monitor skills and other job requirements at an occupation level.
and identify emerging skills and jobs in Europe to advise training providers to revise or design new curricula according to current labour demand requirements in Europe (Cedefop 2018).

Moreover, private companies such as Burning Glass Technologies provide and analyse labour demand information using job portals for countries like the US and the UK. For instance, this company has reported that 80% of middle-skill job advertisements demanded digital skills in 2016, which represents an increase of 4% compared with 2015 (Burning Glass Technologies 2017, p. 3).

4.4.3. Recognising new occupations or skills

As was mentioned in Chapter 3, labour market changes rapidly and new occupations or skills might emerge or disappear over time. The identification of these new patterns in labour demand is relevant because it allows curricula to be adapted by training providers and, as a consequence, it prepares people for technological change. Patterns in labour demand can be identified by recording labour demand information from job portals. For instance, Emsi, a labour market analytics company, has started to build a skill taxonomy, which has identified the growing demand for relatively new skills, such as “Cloud engineers/architects” and “Cloud computing” (Verougstraete 2018). Verougstraete (2018) mentions that this information might be useful to understand how to adapt the labour supply according to changes in labour demand—especially for the most innovative sectors such as IT and tech.

4.4.4. Updating occupation classifications

With a demand for identification of occupations and skills, and new emerging patterns for job requirements, job portal data might facilitate the updating of occupation classifications with real-time information. As mentioned in Chapter 2, occupation classifications are usually not updated as fast as labour market changes occur. A significant amount of time and financial resources are required to analyse information collected from companies and other stakeholders to update an occupation classification. However, with the relatively quick and inexpensive collection of online job advertisements it is now possible to identify job requirements (skills, educational level, tasks, etc.) of each
occupation; hence this information might become an essential contribution to update occupation classifications according to changes in labour demand.

For instance, as recognised by the ILO (2008, p. 2), “some countries may not have the capacity to develop national classifications in the short to medium term. In these circumstances it is advisable for countries initially to focus limited resources on the development of tools to support implementation of ISCO in the national context, for example a national index of occupational titles.” In these circumstances, online job advertisements might provide relevant information to adapt ISCO classifications to a regional context.

Consequently, information from job portals can be used for a range of different topics. Authors such as Turrell et al. (2018) use job vacancy information to understand the effects of labour market mismatch on UK productivity. Moreover, Rothwell (2014) employs advertisement duration as a proxy of vacancy duration in order to determine skill shortages in the US. Additionally, Marinescu and Wolthoff (2016) and Deming and Kahn (2018) use online job advertisements to determine the portion of wage variance explained by employer skill requirements (e.g. cognitive, social, writing, etc.) in the US. However, one of the most promising uses of this information is the identification of skill mismatches. The study of labour demand for skills is a key input to overcome informational barriers between labour demand and supply (Kureková, Beblavy, and Thum, 2016). Yet, as the next section will address, despite the potential of vacancy information, it is essential to take into account its possible limitations, so as to avoid potential biases when analysing information from job portals.

**4.5. Big Data limitations and caveats**

It is important to note that despite the advantages of Big Data, such as the greater volume of information it allows researchers to analyse, there exist some limitations that might affect the analysis of labour demand via information from job portals. Consequently, any study that uses online job advertisements should consider the following issues: 1) data quality; 2) job postings do not necessarily represent real jobs; 3) data representativeness; 4) internet penetration rates, and 5) data privacy.
4.5.1. Data quality

Data quality is one of the most important factors that determines the reliability of any database for statistical purposes. According to the quality framework and guidelines provided by the OECD, data quality is a multi-faced concept within which the relative importance of each dimension depends on user needs. These dimensions are relevance, accuracy, credibility, timeliness, accessibility, interpretability, and coherence (OECD 2011, pp. 7-10) (Table 4.1).

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance</td>
<td>&quot;Degree to which the data serve to address the purposes for which they are sought by users. It depends upon both the coverage of the required topics and the use of appropriate concepts.&quot;</td>
</tr>
<tr>
<td>Accuracy</td>
<td>&quot;Degree to which the data correctly estimate or describe the quantities or characteristics they are designed to measure.&quot;</td>
</tr>
<tr>
<td>Credibility</td>
<td>&quot;Refers to the confidence that users place in those products based simply on their image of the data producer...This implies that the data are perceived to be produced professionally in accordance with appropriate statistical standards, and that policies and practices are transparent. For example, data are not manipulated, nor their release timed in response to political pressure.&quot;</td>
</tr>
<tr>
<td>Timeliness</td>
<td>&quot;Reflects the length of time between their availability and the event or phenomenon they describe but considered in the context of the time period that permits the information to be of value and still acted upon.&quot;</td>
</tr>
<tr>
<td>Accessibility</td>
<td>&quot;Reflects how readily the data can be located and accessed.&quot;</td>
</tr>
<tr>
<td>Interpretability</td>
<td>&quot;Reflects the ease with which the user may understand and properly use and analyse the data. The adequacy of the definitions of concepts, target populations, variables and terminology, underlying the data, and information describing the limitations of the data, if any, largely determines the degree of interpretability.&quot;</td>
</tr>
<tr>
<td>Coherence</td>
<td>&quot;Degree to which they [data] are logically connected and mutually consistent.&quot;</td>
</tr>
</tbody>
</table>

Source: OECD 2011.

With regards to these conditions, given the nature of Big Data on specific job portals as sources of information, this source has a clear advantage in terms of "timeliness" compared with other sources of information such as sectoral surveys. However, as mentioned in Subsection 4.3.4, job portals and, in general, Big Data sources (such as LinkedIn) were not initially created for policy
or academic purposes. This makes the data available through these websites seem relatively disorganised; for example, without standardisation, with duplication issues, and/or with a relatively high portion of missing values. Hence, data quality and the analysis of labour demand using Big Data sources might be affected or limited by these issues of organisation. Table 4.2 lists possible problems that might affect the quality of information provided by job portals.

Table 4.2. **Possible sources that affect the quality of information from job portals**

<table>
<thead>
<tr>
<th>Potential data quality issues</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employers do not follow a specific format when they advertise vacancies</td>
<td>For instance, where the online content indicates the presence of a “job title,” there may also be information regarding the company’s name, location, etc. This unstructured way of announcing vacancies can make statistical inference difficult. For instance, to generate a simple tabulate of a specific variable (e.g. wages), it is necessary to first identify where all (or most) of the information is located on the website and put only this information together to form the corresponding tabulate. Moreover, companies use their own “language” when providing information, such as job descriptions, titles, and the required skills; thus, employers might use different words to define a similar job position (see Chapter 5).</td>
</tr>
<tr>
<td>Companies are not required to provide a standard set of detailed information about the vacancy</td>
<td>The high occurrence of missing values might create bias in the analysis of a certain database. For instance, employers might reveal the wages offered for low-skilled jobs while they might not reveal the wages offered for high-skilled jobs. In consequence, when the mean of wages offered are estimated from any subsequent database, the results would underestimate the average of real wages due to missing information of a specific (high-skilled) occupation group (see Chapter 6).</td>
</tr>
<tr>
<td>Duplication issues</td>
<td>There are two possible types of duplication: within and between job portals. The first type (“within”) refers to the situation where companies might advertise the same job position on the same job portal more than once. The second type (“between”) occurs when employers advertise the same vacancy on more than one website. Consequently, when collecting information about labour demand using different job portals, the number of job vacancies might be overestimated, hence any statistical inference might be biased (see Chapter 6).</td>
</tr>
<tr>
<td>Mistakes in the information</td>
<td>Employers might make mistakes when typing in information, and, in some cases, the information provided might be contradictory. For example, when an employer writes in the job description that work experience is not required, but in the job title it states that some work experience is required (see Chapter 6).</td>
</tr>
</tbody>
</table>

*Source: Author’s elaboration.*
All the problems cited above show that, when working with information from job portals, important issues need to be addressed to guarantee a certain level of data quality (some of these issues are also true of survey information). Clearly, the problems mentioned above can be reduced with the use of data mining techniques, such as data cleaning, classification, and imputation, among others, but they might not be completely eliminated. This result depends on the effectiveness of the algorithms used and the information provided by the employer (Chapter 10 discusses whether the vacancy database for Colombia fulfil the quality requirements established by the OECD).

Thus, the level of these data quality problems and the techniques implemented to tackle them will determine the extent to which information from job portals can be used to analyse labour demand. However, data quality is not the only concern when this type of information is used for analysis. There are other issues: job postings might not necessarily be real jobs, data representativeness, internet penetration, among other issues, might limit the usage of Big Data for the analysis of labour demand.

4.5.2. Job postings are not necessarily real jobs

Given the nature of job portals, any company or individual can post a vacancy.\textsuperscript{45} However, job portals do not have the means to verify whether the advertisement corresponds to a real vacancy or might not be interested in doing so. As Sentz (2013) remarks, when using information from portals, there are difficulties in making a one-to-one comparison between job advertisements and a real job vacancy. For instance, companies might post more job advertisements than available positions in order to receive more applications, and then hire the candidates who best fit their requirements. Another alternative is that companies (such as recruitment agencies) might advertise vacancies to collect CVs and store them in their databases. With this technique companies have already collected the data of potential workers and have the ability to quickly start the screening process in the eventuality of a job opening.

\textsuperscript{45} Depending on the job portal, advertising a vacancy might be free or associated with a cost that generally depends on the time the advertisement is active on the website.
If job portals can post jobs that are not real, or companies can open vacancies without posting them, it is then difficult to precisely determine the number of job vacancies for an occupation, sector, etc., using job portals. These issues do not mean that information from job portals cannot be used as a source to analyse labour demand. With this information in mind to utilise the proper statistical techniques, it is possible to comprehend the structure and trends of labour demand (see Chapter 8); although it may be challenging to determine the exact number of real vacancies available in a period through information from job portals.

Moreover, as Sentz (2013) discusses, even with the above problems, job vacancy advertisements are useful to understand current skill demands, such as who is (or interested in) hiring and where the most employee rotation (turnover) is occurring. For instance, an employer might advertise ten job positions for accountants in a single job advertisement when he/she will eventually only hire five of them. Despite the possibility that job advertisements might overestimate the number of available vacancies, this information might reveal occupations and the demand for skills associated with those occupations. Therefore, information from job portals is a valuable resource to support the analysis of labour demand, even if not all advertisements correspond to a real job position.

### 4.5.3. Data representativeness

Even though information from job portals contains a considerable amount of data, this does not guarantee that this information is representative of the whole economy. On one side, some companies with a specific characteristic (sector, localisation, etc.) might not commonly use job portals to advertise vacancies. On the other side, even in the unlikely situation that every company uses job portals, some specific job positions might exist that are not advertised on websites. For instance, companies might recognise that people with low skills do not tend to use the internet to find a job, and the most effective way to recruit such candidates is through informal channels, such as one-to-one or personal references (e.g. friends). In consequence, depending on the available information on job portals, in some cases, it is not possible to make any statistical inference for a labour market segment or, in other cases, there might be some restrictions when the data are analysed.
Thus, when using information from job portals, it is relevant to understand which segments of the market are properly represented by these sources of information. This discussion of data representativeness is one of the main concerns regarding the use of this type of information for policy recommendation. The representativeness issue determines whether it is possible to analyse and make public policy recommendations for labour markets based on information from job portals. However (as will be discussed in more detail in Chapter 8), testing data representativeness is a complex task. To illustrate this point, it is important to consider how household surveys or sectoral surveys guarantee data representativeness. As mentioned in Subsection 4.3.3, household surveys are based on a population census. This census enables researchers to obtain information about the total number of individuals (“universe”) and their main characteristics over a certain period. When the population and its characteristics are known, it is possible to draw a household sample. In this way, the information from household surveys guarantees that their sample results are as close as possible to the required population parameters (age, gender, etc.).

However, usually, in the case of vacancy analysis, the “universe” is unknown: for instance, the total number of vacancies available in a period by population groups (sector, occupation, localisation, etc.). Therefore, in this case, it is more difficult to know which population is represented by job portal sources. Paradoxically, the relative absence of vacancy information motivates researchers to use information from job portals; nevertheless, this absence of representativeness might limit or put in doubt the usefulness of job portal data.

Some authors such as Štefánik (2012) and Kureková, Beblavy, and Thum (2014) have addressed this issue. However, as pointed out by Kureková, Beblavy, and Thum (2014), most of the studies that have used job advertisements (printed or online) do not discuss or test data representativeness, and their findings are generalised for occupational or sectorial groups. The absence of discussion aimed at identifying data representativeness might affect the reliability of many studies. For instance, given the nature of the internet, occupations related to internet technologies (IT) tend to be overrepresented in online job advertisements. Consequently, a study that does not account for this source bias might conclude that IT skills are one of the most relevant skills required to find a job, while considering the total number of real vacancies (those advertised and not advertised on the internet), the actual share of IT occupations might be minimal.
Therefore, discussing and testing the representativeness of job portal data for academic and public policy purposes is a key issue when considering the use of these sources of information. The validity and the generalisation of results from the analysis of online job advertisements depend on the population being represented by job portal sources. For this reason, Chapter 8 discusses and tests data representativeness for the Colombian case.

4.5.4. Limited internet penetration rates

Related to the above point, the usefulness of information from job portals and, hence, their representativeness depends on internet penetration rates (the percentage of the total population that uses the internet). Although internet usage has increased (see Section 4.2), this growth might not cover some sectors, regions, etc. For instance, in Colombia, there is a remarkable disparity between rural and urban zones in terms of internet access.46 Given this limited access, employers might tend towards the use of other job advertising channels such as asking friends or colleagues to recruit potential workers.

In regions where the growth of internet access has not occurred or has occurred at a slower pace, the inferences that can be drawn from job portal data might be more restricted than in areas where internet access is more widespread. Places where there is less internet access tend to be poorer, and information about labour demand tends to be scarcer due to the prohibited cost of doing a vacancy survey. In consequence, even where the internet is not widely used, paradoxically, it might be the only reliable source of information to analyse labour demand. Hence, the statistical inference from job portals depends on internet penetration rates; however, even when internet access is relatively low, online job advertisements might be a rich source for analysing important segments of the labour market.

Additionally, as Kureková, Beblavy, and Thum (2014) mentions, it is highly likely that the internet continues to spread across different regions and socio-economics groups, so that reliance on internet-based recruitment methods will

---

46 According to the Economic Commission for Latin America and the Caribbean (CEPAL 2016, p. 12), around 10% of households in rural areas had access to the internet in 2014, while around 50% of households in urban zones had access to the internet in 2015.
increase over time. In consequence, internet penetration rates limit the statistical inferences that can be drawn from job portal data; however, those limits are becoming less relevant due to technological advances.

4.5.5. Data privacy

Online job vacancy advertisements belong to job portals or to other platforms where employers have decided to make their vacancies public. Provided that job vacancy information is shared and is administrated by a third party, this issue might affect the statistical inferences that can be drawn from those sources. First, the availability of information might change due to modifications on the platforms. As private administrators, job portals might unexpectedly change the number of vacancies or the number and/or kind of variables displayed on their websites, which in turn affects what information is available for researchers, especially when attempting to analyse changes in the economic environment (e.g. number of vacancies, wages, etc.).

Second, job portals can restrict the usage of vacancy information. In most cases, job portals prohibit the storage and usage of job advertisements for commercial purposes; however, for statistical purposes, there does not seem to be any legal restriction. For instance, the Cedefop project “Big Data analysis from online vacancies” has started to collect information from different job portals in Europe. Cedefop has informed these portals that information is going to be collected for statistical purposes, and most of the job portals have not denied access to their data. Nevertheless, as mentioned above, the project has required new statistical legislation to delineate the use of information from job portals and other non-traditional information sources.

Table 4.3 summarises the main advantages and disadvantages of different data sources for the analysis of labour demand. Both traditional (sectoral and household surveys) and non-traditional sources of information (online job portals) have advantages and disadvantages regarding the study of labour demand. Consequently (at this point), non-traditional surveys cannot replace traditional sources of information, although non-traditional sources such as

---

47 Some websites might adjust the number of variables displayed, such as wages, because potential workers might not apply for the job given previous characteristics of the vacancy.
Big Data might complement and support sectoral or household surveys and vice-versa (see Chapter 9).

Table 4.3. Advantages and disadvantages of data sources for the analysis of labour demand

<table>
<thead>
<tr>
<th>Source</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
</table>
| Sectoral surveys | • Guarantee a certain level of data representativeness  
• Provide (usually macro) indicators of labour demand                                                                                   | • Aggregated data  
• Time consuming  
• Relatively expensive  
• Fixed structure  
• Less frequent than household surveys                                                                                                    |
| Household surveys| • Guarantee a certain level of data representativeness  
• Provide (aggregated occupational or skills) indicators about the labour force  
• Generally available as long-term time series                                                                                           | • Aggregated data  
• Time consuming  
• Relatively expensive  
• Fixed structure  
• Information from the labour supply                                                                                                        |
| Job portals      | • High volume of data  
• Information in real time  
• Inexpensive  
• Disaggregation level  
• Detailed information  
• Useful for different purposes (e.g. to estimate vacancy levels, to identify skills and other job requirements, etc.) | • Data quality issues  
• Job postings are not necessarily real jobs  
• There is no a priori guarantee of a certain level of data representativeness  
• Depends on internet penetration rates                                                                                                     |

Source: Author’s elaboration.

4.6. Big Data in the Colombian context

As mentioned before, in some contexts, Big Data sources might be the only ones available to analyse different labour market topics (Kureková, Beblavy, and Thum, 2014). Specifically, in Latin American countries like Colombia, the use of information from online job portals can provide valuable first insights about “skill-shortage vacancies.”

Therefore, given the potential for vacancy analysis in Colombia and the high expectations generated by this topic, in order to understand the potential scope of these data sources it is first necessary to answer the following questions:
1) How and to what extent could a web-based system for monitoring skills and skill mismatches using job portals and household surveys be developed for Colombia? Specifically, 2) how can job portal data be used to inform policy recommendations, primarily to address two of the major labour market problems in Colombia, which are its high unemployment and informality rates? And 3) to what extent can these sources be used together (information from job portals [unsatisfied demand] and national household surveys [labour supply]) to provide insights into skill mismatches (skill shortage) in a developing economy?

By answering the above questions, this study contributes to our current knowledge of the advantages and limitations of novel sources of information, which attempt to address public policy issues and/or academic research problems. It provides a methodological and analytical model for countries with scarce information regarding occupations and skills in the labour market by considering possible limitations and biases surrounding vacancy data. It also provides an analysis of the labour market in terms of occupations and skills. Importantly, this research is useful to institutions to match disadvantaged workers (especially unemployed and informal workers) to jobs that they have the potential capabilities to fill, or could be used to help employees develop certain skills, which might not be easily transferable through the formal educational system or programs such as VET (Kureková, Beblavy, and Thum, 2014).

As previously mentioned, the most important ongoing project similar to this book is the “Big Data analysis from online vacancies” project conducted by the Cedefop. So far, this project has been focused on analysing skills and job requirements in Europe from job portals. A remarkable task given the necessity to capture and analyse online sources in more than 24 official EU languages, since April 2018 (Cedefop 2019). However, as summarised in Table 4.4, this study is distinct from the Cedefop project in eight aspects.
Table 4.4. **The main differences between the Cedefop and the Colombian vacancy projects**

<table>
<thead>
<tr>
<th>Source</th>
<th>Cedefop</th>
<th>Colombian vacancies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region</td>
<td>European Union</td>
<td>Colombia</td>
</tr>
<tr>
<td>Theoretical framework regarding labour market mismatches and the potential usefulness of job portals to tackle skill mismatches</td>
<td>The project is in a stage where vacancy data have begun to be downloaded and processed (exploration stage). It has not been exhaustively discussed and tested to determine the usefulness of job portals to tackle skill mismatches.</td>
<td>It provides a theoretical framework and concepts that highlight the benefits of analysing information from job portals for tackling skill mismatches.</td>
</tr>
<tr>
<td>Extraction of information</td>
<td>Job title, skill, and sector variables are collected and processed.</td>
<td>This study considers and proposes various methods to collect and process a wider number of variables, such as job title, labour experience, educational requirements, (imputed and non-imputed) wage, and skills, among others.</td>
</tr>
<tr>
<td>Methods to classify job titles into occupations and to identify skills</td>
<td>Machine learning algorithms and the use of a European skills dictionary.</td>
<td>It proposes new mixed methods to properly classify job titles into occupations and to identify skills for a country that does not have national skills dictionaries.</td>
</tr>
<tr>
<td>Analysis of variables such as educational requirements, wages, and sector, among others</td>
<td>The project (so far) is focused on describing the most demanded skills and occupations.</td>
<td>This study uses variables, such as occupations, skills, wages, educational requirements, etc., to exhaustively validate and analyse vacancy data.</td>
</tr>
<tr>
<td>Period of analysis</td>
<td>April 2018 – ongoing</td>
<td>January 2016 – ongoing</td>
</tr>
<tr>
<td>Framework to test the validity and consistency of job portal data</td>
<td>The consistency of the results has not been tested yet.</td>
<td>The vacancy database is exhaustively tested. In fact, a framework is suggested to evaluate its representativeness for each occupation at a different level of disaggregation.</td>
</tr>
<tr>
<td>Combination of job portal and household survey data to determine skill shortages</td>
<td>The vacancy data have been used to provide a preliminary overview of demanded occupations and skills.</td>
<td>It provides a descriptive and detailed analysis of occupations, skills, educational requirements, wages, among others. Vacancy data are combined with household data to monitor skill shortages.</td>
</tr>
</tbody>
</table>

*Source: Author’s elaboration.*
4.7. Conclusion

Technological changes have facilitated the generation and storage of large amounts of information at a low cost in terms of time and money. Together with an increased volume of information, a set of different techniques has been developed to process and analyse the massive information generated and available for research and analysis. This large amount of information and the techniques to manage this kind of data have been named “Big Data.” As the name suggests, this term refers to a relatively high volume of data; nevertheless, this is not the only characteristic of “Big Data;” indeed, the most common three properties assigned to this term, as described in Section 4.2, are volume, variety, and velocity (Laney 2001).

The Big Data phenomenon has attracted the attention of private and public companies, as well as researchers (among others), because Big Data might provide relevant information for the analysis of individual behaviour, especially in contexts where previously there was a lack of data. The labour market is one of these scenarios where information was traditionally limited or relatively absent, especially for the analysis of labour demand requirements. Collecting information related to labour demand by traditional methods (e.g. surveys) is relatively costly in time and monetary terms. Moreover, even in cases where there is information about labour demand, this information might not be disaggregated (or well-designed) enough to analyse employer requirements. This absence of information and, hence, labour demand analysis is one of the main obstacles to tackle possible skill mismatches. Individuals and training providers unaware of employer requirements might offer skills that are not required by the labour demand.

Consequently, Big Data—specifically job portals—might provide valuable information in real time and at a low cost for the analysis of labour demand, contributing thus to the identification of skill shortages. Compared with traditional sources of information, such as sectoral or household surveys, job portals 1) provide labour market information in a short period of time (real time); 2) enable a relatively inexpensive collection of job portal data; 3) provide a high volume of detailed information, and, hence, 4) allow their data to be disaggregated to skills and occupational levels. Given these advantages and
the potential use of information from job portals, there has been an increasing interest from researchers and policymakers to utilise online job advertisements.

However, little attention has been paid to the possible limitations and biases of the information provided by job portals, and how these issues might affect labour demand analysis. As a source of information, job portal data have the following limitations: 1) data quality; 2) job postings are not necessarily real jobs; 3) data representativeness; 4) internet penetration rates, and 5) data privacy. This chapter has discussed the need for labour demand information that job portals might fill. However, before making any statistical inferences for these sources of information, first it is necessary to know as much as possible about the biases and limitations of their data. Consequently, since Big Data have considerable limitations, as is the case of household or sectoral surveys, it is necessary to evaluate the scope of these sources of information.

At this point, Big Data is a complement rather than a substitute for traditional data collection, such as household and employer surveys, among others. Yet, in a context where information is scarce, Big Data might be the only “reliable” source available for labour demand analysis. This is the case for Colombia (and Latin America), where there are high complaint rates about the quality of workers by companies, and there is not enough labour demand information to address worker skills according to employer requirements. Consequently, the next chapters present a methodology to collect and analyse labour demand information considering possible information biases.