doi.org/10.26398/IJAS.286

# INTERVIEWING ADMINISTRATIVE RECORDS

## A CONCEPTUAL MAP FOR THE USE OF BIG DATA FOR ECONOMIC RESEARCH

**Roberto Leombruni**
*Department of Economics and Statistics "Cognetti de Martiis", University of Torino, Italy*

*roberto.leombruni@unito.it*

*ORCID: 0000-0001-8816-2407*

INTERVIEWING ADMINISTRATIVE RECORDS

A CONCEPTUAL MAP FOR THE USE OF BIG DATA FOR ECONOMIC RESEARCH

***Abstract**. Businesses, academia and official statistics are turning more and more to novel data sources besides traditional sampling surveys, but the debate about their defining features and the challenges they pose for research is still open. In this paper I propose a conceptual map of what data are in the field of empirical economic research to clarify what are the conditions and possible strategies to fully grasp their opportunities, particularly in the case of big data of administrative origin. The conceptual map is inserted into a recent literature addressing the clarification of the very notion of data, and exemplified using three well-know cases of failures and best practices in the use of large data samples. The conceptual map is then used to discuss the case of labour market research based on social security data.*

***Keywords**: Big data for research purposes; Notion of data; Administrative data on the labour market.*


Data Availability Statement: All relevant data are within the paper and its Supporting Information files.
Competing Interests: The Author(s) declare(s) no conflict of interest.

*Ask not what you can do to the data,*
*but what the data can do for you*
Zvi Griliches

## 1. INTRODUCTION

The use of novel data sources besides traditional sampling surveys is gaining a primary role in official statistics worldwide. In Europe, an important step in this direction was taken in 2014 with the European Statistical System vision 2020, which identified in a wider and better use of new sources such as administrative and geospatial data a key strategy to answer to the challenges that official statistics was facing. The rationale of this tendency has indeed to do with the cost reduction implied by the re-use of data already collected. From a statistical point of view, however, the main interest rests on the benefits side, and in particular in the possibility of exploiting the wealth of information held by public institutions, e.g. in their fiscal or social security registers, or by private organisations such as telecommunication companies or financial institutions. The unprecedented size and timely amount of information characterising these data sources has the potential to allow the publishing of population-based real-time statistics, on themes and/or with an accuracy that in many cases are out of reach for traditional surveys.

In a somehow parallel way to what is happening in official statistics, the exploitation of administrative data is more and more characterising also academic research. Already ten years ago Raj Chetty was noting that in a leading economic journal such as the American Economic Review the share of articles using micro-data based on surveys went down from about 60% in 1980 to 30% in 2010, mirrored by an increase in articles based on administrative data from 30% to 60% of all publications. A swap in relative importance even larger in the case of the Quarterly Journal of Economics, with survey based articles going down from over 90% to 10% and administrative data based ones from about 10% to 70% (Chetty, 2012). In Italy, first examples of this are the studies carried out already starting in the mid-Eighties exploiting the National Social Security Administration archives to produce novel evidence on previously unexplored matters such as firm demography, job creation and destruction and earnings differentials (Contini and Revelli, 1986, 1987). A back of the envelope estimate says that currently about 70% of all empirical studies published about

Italian labour market are based on INPS administrative data, versus 30% on ISTAT's labour force survey[1].

The trend in using "alien" data sources with respect to statistical surveys further accelerated in the last decade with the advent of the so called big data. Two of the three "V" that are usually quoted to define them (volume and velocity) are indeed a key feature also of administrative archives of public and private organizations. To them, the "V" pointing to the variety of new and different kind of information, such as image- and textual data, took the hotspot in businesses and also in economic research, with applications ranging from the use of satellite imagery to obtain local area estimates of poverty and industrial development (Engstrom *et al*, 2021; Soren and Fisker, 2018), to the use of Twitter posts to predict labour market flows (Antenucci *et al*, 2020) or regional unemployment rates (Llorente *et al*, 2015), to the large and growing literature using on-line platforms for job advertising to estimate the dynamics and skill composition of labour demand (see e.g. CEDEFOP, 2019, and Khaouja *et al.*, 2021, for a review).

It is fair to say, however, that the surge in the use of big data – including administrative ones – has not been accompanied by a matching methodological literature investigating not only the opportunities they offer but also the challenges they pose. Still in 2007, in one of the first systematic studies on the use of administrative registers within official statistics, the authors noted that no well-established theory in the field existed (Wallgren and Wallgren, 2007). Their point was that while statistical surveys are a well-known object in research, thanks to established methodologies grounded on probability theory and inference, no comparable terms or principles were available to provide grounds to a systematic theory on statistical systems based on registers. From this point of view, the current integration of new sources of data within the production process of official statistics is providing the best setting for their study, and indeed we are witnessing many advances in the literature,

[1] Searching in EconLit all papers published in the last 10 years about Italy containing any of a large set of labour-related keywords (labour market, work, employment, unemployment, job, workforce, wages, retirement, pension system, welfare system), combined with either "INPS or Administrative Data" or with "LFS or Labour Force Survey".

particularly in the field of data integration, data quality and total error estimation (see e.g. the essays in Hill *et al*, 2020).

Outside official statistics, however, the situation is more blurred. The possibility of linking administrative data to statistical samples of a well-defined target population would offer also to researchers a solid base for statistical inference. But, also in view of current regulations about data protection and integration, this kind of anchorage is seldom available to the researcher. If one had to review in more detail the kind of data used in academic research, most of the studies would be probably classified as stand-alone applications of administrative data, i.e., as a direct use of an exerpt of administrative archives without their integration into a statistical survey. In Italy, two exceptions to this are the studies on socio-economic inequalities and health based on the linkage of administrative data on health to ISTAT's surveys (see e.g. Ardito *et al*, 2020; Petrelli *et al*, 2022) and the T-DYMM microsimulation models of Italian social security system based on the linkage of the Italian section of EU-SILC to INPS administrative data (see e.g. Conti *et al*, 2023).

From the other side, the recents advancements notwithstanding, the literature about the challenges posed by big data is still scant – on average only about 5 in one thousand articles on big data deals with epistemological aspects (Balazka and Dario Rodighiero, 2020) – and far from having reached maturity even about the very definition of its subject matter. Recent reviews of the literature actually reported the existence of different and sometimes contrasting views about what big data are (see e.g. Connelly *et al*, 2016; Fosso Wamba *et al*, 2015; Al Sai *et al*, 2019). But even besides the issue of identifying where the boundary lays between different typologies of data sources, what is somewhat surprising is the lack of an accepted consensus even about the very notion of "data" also without the "big" part.

Actually, the lack of an agreement about what the term "data" means was the point of departure of a highly quoted article in the information science literature already thirty years ago, which proposed a definition of data as a triple <entity, attribute, value> which is still today a common one in database theory (Fox et al, 1994). Point is, more than ten years later a panel composed of information science scholars reported a list of 42 different definitions of information, data, and their mutual relationship (Zins, 2007). The definitional issue is still today the focus of a large and interesting debate within the larger

Literary and Information Sciences (LIS) literature, part of which will be discussed in a section below as a background for the current contribution.

The aim of this paper is more limited in scope with respect to a fully discussion of different definitions of data and big data. It is to propose a conceptual map of what data are in the field of empirical economic research, as a basis to discuss what are the challenges for the use of big data in this specific field and what are the conditions and possible strategies to fully grasp their opportunities. To exemplify my argument, in next section I will present three examples of "going big" in applied research. I will then present the conceptual map, locating it in the current literature about the epistemological nature of data (Sections 3 and 4). In Section 5 I will provide some examples of the proposed strategy in the case of labour market research based on social security data. The final section will propose some concluding remarks.

## 2. OLD AND NEW BIG DATA STORIES

Technological aspects usually play a big role in the narrative about big data, as their use is strictly linked to several innovations we witnessed to in last decades in computing power, storage capabilities and cloud technologies (Al-Sai *et al.*, 2019). Besides the current hype on the technological innovations, however, at least the "V" referring to the volume of large datasets is out there from pretty some time. At the very beginning of modern statistical enquiry, we may say that it is more the US Census that triggered improvements in computing technologies – such as the use of punched cards for the storage and processing of data – than the reverse.

Also the first story I'm proposing here is an old one, recounted in many statistics handbooks: The large scale election poll by the *Literary Digest* (LD) during the US presidential campaign in 1936. The LD was a popular magazine which had actually been succesfull in predicting all previous presidential elections. In 1936 they decided to sample an astounding 10 million US citizens, collecting 2.4 million answers, which is an amazing sample size also for today's standards. Their prediction was a huge victory for Alfred Landon, the republican candidate, over the incumbent president Franklin D. Roosevelt, but the result was simply the opposite: a landslide in favour of Roosevelt, who gathered 98.5% of the electoral votes – the largest victory ever in US history.

The number-one suspect for this epic fail was the sampling frame, which was based on lists of telephone and automobile owners, which resulted in the sampling of wealthier than average, pro-republicans voters. Also nonresponse bias was identified as an important factor: a recent reassessment of the matter found that pro-Landon voters were more keen to participate to the LD poll with respect to pro-Roosevelt ones (see Lusinchi, 2012, and Lohr and Brick, 2017). With a bit of a joke, an article titled *Digest Digested* appeared in the *Times* magazine in 1938 reporting the epilogue of this story: the *Literary Digest* ended its publishing history being absorbed by the *Time* magazine itself.

The next two examples are not relative to social inquiry or economics, rather to the rising field of epidemic intelligence, but they too are illustrative of the promises and perils of "going big".

The first is another well-known example, probably the first time a modern big data analytics approach gained the highest stand in scientific research, thanks to a paper published in *Nature* predicting influenza epidemics using search engine query data, the so-called *Google Flu Trends* (GFT, Ginsberg *et al.*, 2009). Building on Polgren and co-authors (2008), who already detected a correlation between virological surveillance data and search queries containing the words "flu" or "influenza", they aggregated historical logs of web searches for 50 million of the most common queries in the US, to build a system which consistently predicted epidemic outbreaks 1-2 weeks ahead of the official surveillance reports. Just four years later, however, the GFT was closed, since it was predicting almost double the number of doctor visits subsequently realised. There have been several explanations for this, including changes in user behaviours and in search engine functioning (see Shin et al., 2016, for a review). Lazer and co-authors proposed also an overfitting issue – the 50 millions records were used to fit as few as 1,152 real observations – plus two considerations. The first is a general point which is sometimes forgotten in what they called the "big data hubris": It's simply not just about the size of data. The second anticipates some aspects I'll be dealing with in the next section: "All empirical research stands on a foundation of measurement. Is the instrumentation actually capturing the theoretical construct of interest? Is measurement stable and comparable across cases and over time?" (Lazer *et al.*, 2014, p. 1204).

The last story – a successful one – is about the Artemis project of the University of Ontario, aimed at preventing disease spread in neonatal intensive

care units (NICU, see Blount *et al.*, 2010; McGregor *et al.*, 2011). The project was based on the real-time analysis of patients' data streams to identify conditions preceding the onset of medical complications. The data gathering included measures such as ECG readings, respiratory rate, blood-oxygen saturation and blood pressure metrics, for about three million data points per hour for each infant in the NICU. The interesting point here is not only the quantity and velocity of information gathered. The "on-line analysis" of data, through the automatic application of clinical rules pointing to possible medical complications, provided clinicians with a decision support based on a stream of data too large to be assessed with a traditional, "off-line" scrutiny of the same information. The deployment of the entire data gathering and analysis pipeline resulted in early warnings of infection spreading 24h before with respect to the traditional approach.

## 3. THE NOTION OF DATA

Tim Harford, in a lecture given in 2014 at the Royal Statistical Society, discussed the LD and GFT examples proposing two readings of the challenges posed by big data (Harford, 2014). The first is the "theory-free" risk somehow embedded in going big. As a representative example of this attitude he proposed a quote of the provocative *Wired* essay by Anderson (2008), *The End of Theory*: "With enough data, the numbers speak for themselves". This is actually a rather general point and a recurrent discussion in economics and statistics, that we could view as a modern reprise of the Koopmans *versus* Vining "Measurement without Theory" debate in the 1940s and 1950s (Koopmans *et al.*, 1995). The second is more specific to our theme, and is the emphasis he put on one of the defining features of big data, which he labels as the "digital exhaust" of web searches, mobile trackings and credit card payments, proposing "found data" as a (more gentle) way of identifying one of their common characteristics[2].

[2] The term "found data" was actually already common in the AI literature, where the availability of large corpora of "found", textual and audio data were key to the first achievements in speech recognition and natural language processing long before the term "big data" was even introduced (see *e.g.* Gauvain *et al*, 2000; Hirschberg and Manning, 2015). See also the quote by Griliches below.

The fact that the information used by a researcher has not been "made" on purpose for statistical use, but "found" somewhere as the outcome of a process with different purposes – being them the ones of a public administration or of an individual browsing the internet – is clearly relevant to the point. It does not help however to discriminate between the examples we presented: The Artemis project, a brilliant application of a big data approach, is by no means "found data"; the *Literary Digest* was based on "made" data, but it was wrong; the GFT was based on data collected by Google itself, so in this case the very difference between "made" and "found" is a bit fuzzy.

Griliches, anticipating this very theme, already used the term "found data" to note that most of the work in econometrics is based on data that have been collected and assembled by somebody else, often for quite different purposes, including statistical ones (see e.g. Griliches 1984, 1986; see also Triplett, 2007). He noted that this is not a bad *per se* but, rather, a defining feature of much of economic research. When data were perfect the very discipline of econometrics would possibly not exist: the "existential problem" of econometrics is "life with imperfect data and inadequate theories" (Griliches, 1984). His main point – as in the Lazer and co-authors comment quoted above – is again a stress on the issue of measurement. It is because data are imperfect that it is important to consider at least two different data generation processes: the economic model describing the behaviours of economic actors and the measurement model describing how this behaviour is recorded. "While it is usual to focus our attention on the former, a complete analysis must consider them both" (Griliches 1985, p. 198).

If from the one side Griliches is noting that in many cases data is "found", we can rephrase his point about the importance of the measurement process by saying that actually all data is "made". As we now will see, the prominence of the "made" aspect is one of the points which is actually emerging in the current literature trying to clarify the very notion of data, particularly in these years where more and more social and economic events are leaving a "digital exhaust" so that possibly anything is becoming data.

Fox and co-authors, in their early assessment of the matter, pointed to three different approaches to define data (Fox *et al*, 1994).

i. Data as "raw facts"
ii. Data as a triple <entity, attribute, value>
iii. Data as a result of measurement or observation.

A recent literature has taken up the issue of clarifying the notion of data adding details to this classification, particularly in view of the complexities brought about by the diffusion of big data (see e.g. Floridi, 2008; Borgman, 2015; Frické, 2015; Leonelli, 2015 and 2019; Hjørland, 2018; Gellert, 2022). A full review of this literature is out of scope for the present paper: for the sake of my argument I will stick to the i-iii views, which are still today the most common ones, recalling how the recent debate has contributed to clarify their differences.

The view of data as "raw facts" is the closest one to the Latin etymology of the term, *datum*. Data is "what is given", is a set of raw facts about a phenomenon which are the base of our arguments or elaborations. This is a very general definition of the term valid in common speech but also a pretty common one, for instance, in the studies within information science using the "DIKW Pyramid" conceptual map (Data, Information, Knowledge, Wisdom) in which data is the raw material on which information is built[3] (Zins, 2007). The most recent literature discussing what data are from an epistemological point of view however tends to critique this view, stressing that the process of scientific discovery does not produce "objective knowledge" about phenomena, since the scientific process itself is "theory-laden", a concept originally put forth by Pierre Duhem and recently applied to the issue of defining data by Leonelli (2015). Looking at actual scientific praxis, it has also been noted that data are always variously "cooked" within the circumstances of their collection, storage, and transmission, so that the "raw" label is actually an oxymoron (see the essays collected in Gitelman, 2013, particularly the one by Bryne and Poovey recounting Fisher's "data scrubbing" in its pioneering contributions to financial modelling). Besides the "raw" label, also the very idea of a factual piece of information "trades one difficult concept (data) for an equally difficult one (facts)" (Floridi, 2008).

The view of data as "raw material" is actually a reasonable one from the point of view of information systems design. As an example, when designing a datawarehouse, there is a clear point of entry for the data the system is based

[3] Quering "Data" in InfoScipedia, a large database of terms and definitions in Information Science and Technologies, 15 in 68 entries explicitely stress the "raw" aspect of data. At https://www.igi-global.com/dictionary/, retrieved October 4th, 2023.

on, and all the procedures for their elaboration and visualisation take them as "what is given". Also in this context, however, definition i) misses the necessary details to discuss in a rigorous way, among others, the quality dimension of data, which was one of the starting point of the alternative approach proposed by Fox and co-authors in their seminal contribution. The same authors provided a more recent outline of this view, widely accepted in the database community, which is actually a collection of definitions of several related items. Within this view, a *datum* or *data item* is an ordered triple <*e*, *a*, *v*>, asserting that the entity *e* has the value *v* for its attribute *a* (Redman *et al*, 2017).

The examples provided in previous section all fit well in this definition. In the *Literary Digest* poll, entities are individuals, the attribute is the intention to vote, the values are the actual intentions to vote of each individual. The case of *Google Flu Trends* is pretty similar, as long as we see web searches as individuals putting queries into a search engine "ballot". The only difference is that the possible values of the attribute (the web search) is not the close list of candidates of an electoral campaign, such as in a multiple choice question, but free text such as in an open-ended one. The Artemis project is coherent with this definition, too: The entities are the infants and the automatic medical readings populate a set of several <*attribute*, *value*> couples.

This formal definition of data is also pretty similar to the one provided by the Royal Society, together with definitions of information and knowledge in a reduced version of the DIKW model, which define them as "numbers, characters or images that designate an attribute of a phenomenon" (Boulton *et al*, 2012, p 12, cit. by Leonelli, 2015). While this view of data has the double appeal of being intuitive and to avoid the use in a circular way of the notion of fact, it has actually the same limits of definition i). In concrete research situations data are never an abstract object defined in terms of their intrinsic properties, rather, they are defined in terms of their function within specific processes of inquiry. This is an argument used by Leonelli (2015) in order to argue in favour of a "relational" definition of data. The stress on the process of inquiry brings again the issue of measurement: Also assuming there is such thing as a fact, it cannot consist of just the value of an attribute, it needs information about the way it was collected. This kind of description, that in the database terminology would be classified as metadata, is an integral part of

approach iii), which defines data explicitely as a result of measurement or observation.

An early and interesting account of data as observations can be found in Yovit's seminal paper trying to define the field of Information Science (Yovit, 1969). His starting point are "observable actions", i.e., quantities which are physical in nature such as the position of an aircraft, the result of a scientific experiment, a new product developed by a firm. As such, they are neither information nor data. In order to become data they must be transformed by a function (which Yovit calls the "T function"), which is fundamentally a measuring device which transforms the observable actions to data. Fox and co-authors, although acknowledging the profound importance of the way data is obtained, object that there are common examples of data which are not obtained by observation but are "assigned" to an entity, such as someone's name or social security number. As I will argue below, this is an objection which is easily takled with, for the sake – so to speak – of not throwing out the baby (the measurement issue) with the bath water.

Hjørland (2018) provides a recent assessment of this view, summing up the many contributions which described the nature of data not as "given" but as *capta*, i.e., "taken" and constructed, including the nice seminal posing of the matter by Jensen (1950):

> It is an unfortunate accident of history that the term *datum* (Latin, past participle of *dare*, 'to give') rather than *captum* (Latin, past participle of *capere*, 'to take') should have come to symbolize the unit-phenomenon in science. For science deals, not with 'that which has been given' by nature to the scientist, but with 'that which has been taken' or selected from nature by the scientist in accordance with his purpose.

This is why the metadata describing the process by which data has been collected, including the purposes and perspectives specific to the research activity which originated them, are not an accessory but rather an indispensable element for the use and re-use of databases, particularly in big data research (see Leonelli, 2014, and the discussion in Hjørland, 2018).

What is then "taking data" in statistics? It is a process by far more elaborate that the operation of a single, however complex measuring device to record physical or biological measures. Adrian Smith, in his presidential

address to the Royal Statistical Society in 1996, discussed what is the possible contribution of statistics for the development of an evidence-based society, in which "informed quantitative reasoning" is the base of public debate and of decision-making in government and business. He proposed a view of statistics as "the science of doing science", "whose role is to provide theory and protocols to guide and discipline all forms of quantitative investigatory procedure" (Smith, 1996). He went on proposing sort-of a check list of the tasks needed to produce reliable quantitative evidence, including:

1. The framing of questions
2. Design of experiments or surveys
3. Drawing up protocols for data collection
4. Collection of data
5. Monitoring compliance with protocols
6. Monitoring data quality
7. Data storage, summarization, presentation
8. Stochastic modelling
9. Statistical analysis
10. Model criticism and assumptions assessment
11. Inference reporting
12. Use of results for prediction, decision-making or hypothesis generation

What is missing in a view of data as "given", such as in definition i), and what is hidden in a formal view of them such as in definition ii), are all the activities from 1 to 6, characterising how statistical data are gathered, which is a mix of statistical theory and of technical competencies and process management. In principle, if the big data revolution had just to do with the size of data but all the phases of data procurement followed an adequate standard, no harm is out there. One could even forget, so to speak, statistical inference: recalling *The end of theory* in previous paragraph, "With enough data, the numbers speak for themselves". Strictly speaking, however, the size of data has to do mainly with point 2 in Smith's list, about the sampling of the population, but all other points are equally important to generate reliable evidence.

In fact, while the three examples in previous section were not easily distinguishable from the point of view of i) and ii) notions of data, the accent on the data production process cast more clear differences among them. Weak theory was indeed one of the piftalls in the *Literaty Digest* case, due to a poor choice of the population frame (point 2), together with a complete absence of monitoring of unit non-response (point 6). In the case of Google Flu Trends, in

the absence of a full documentation of the process, we can say few about points 2 to 6 – which is already a crucial "missing-metadata" issue for their use to inform decision-making. About point 1 on the framing of questions, what is sure is that it was entirely unspecific. Technically, the "question" was "What are you serching for on the web?", with completely unstructured, open-text "answers". Since the question was not directly addressing the theoretical construct of interest, no guarantees about the stability of its relation with actual disease spreading was granted.

In the case of the Artemis example, reading the methodological articles describing the project gives a sense of a huge and specific work on the design and implementation of all phases of data-procurement. From the point of view of Smith's account of how statistical evidence is produced, the Artemis big data project is way more a text-book example of "traditional" statistical enquiry with respect to the *Literary Digest* sampling survey.

## 4. A CONCEPTUAL MAP FOR THE USE OF FOUND DATA

I here present a simple conceptual map of data as *capta* that serves as a basis to discuss the use for scientific purposes of data not gathered under our direct control.

I take from Yovit (1969) the idea that data are produced by means of some "*T* function" applied to "observable actions". About the latter, the qualification "observable" is a bit of a tautology, while "actions" can be restrictive, so I will say in a more generic way that the research activity has to do with some phenomenon of interest, which I call $\varphi$. To denote in a more generic way also the process going from $\varphi$ to a numerical representation of it I will use the term "map". Indeed, "function" and "map" are terms often used interchangeably, but the latter delivers more directly the idea that we are representing some aspects of the phenomenon of interest, as in geographical maps. Besides, sometimes the term function is used defining at the same time a specific codomain – such as a function mapping into $\Re$ – while in times of big data the codomain is often unstructured, such as in sound, textual or image repositories. For the same reason, I will not adopt at this level the representational approach stating that the result of the map is a triple $< e, a, v >$. Rather, the mapping of the phenomenon under study produces a couple that pairs $\varphi$ to some digital representation of it. I hence define data as a labelled set of digits resulting from

the application of a map on a phenomenon under study, as in figure 1 below. When the set of digits does actually not possess any structure induced by the map, we may define this kind of data as a digital copy of φ, as in the case of textual or image data, whose repository is usually termed as a datalake instead then a database. Such unstructured data can enter *as are* into the stochastic and statistical modelling phases in Smith's check list, as in the case of GFT, or in the case of the application of sentiment analysis techniques to financial news or for public opinion mining (see e.g. Saberi and Saad, 2017, and Man *et al.*, 2019).

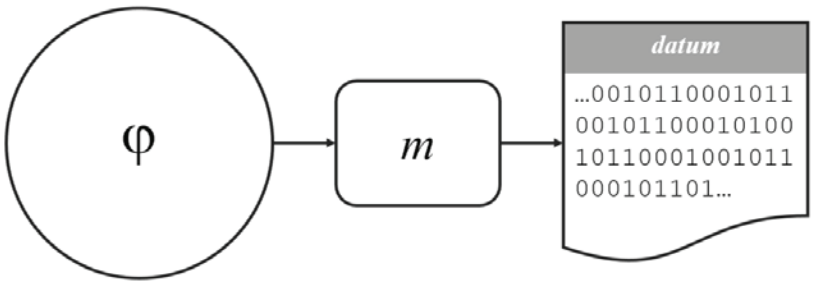


**Figure 1: Data as a digital mapping of a phenomenon of interest.**

Alternatively, an explicit activity of feature extraction may be implemented before statistical modelling to annotate the digital copies and obtain structured information. In this case, a second map is applied to the digital copy instead than on φ, in order to extract an information set in the form <*attribute*, *value*>, as in Figure 2. In this case, as is common in database jargon, we may talk of a "structure-on-read" schema (Cackett *et al.*, 2013).

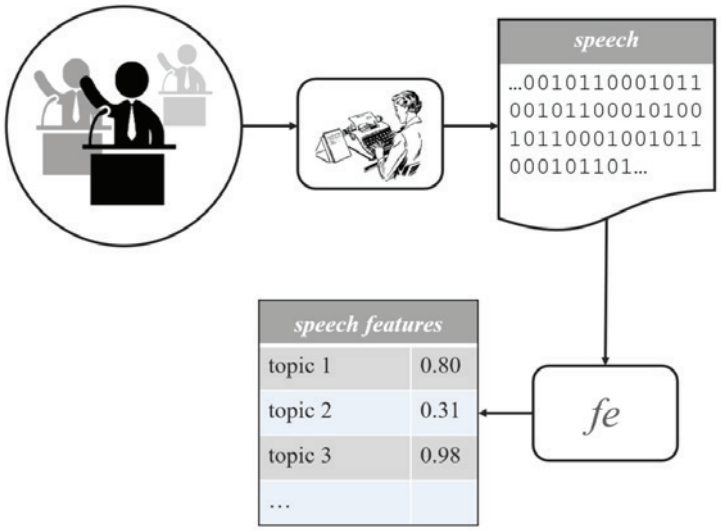


**Figure 2: Structure *on read* mapping.**

The map may also represent (or measure) the phenomenon of interest directly along a grid of attributes. This is the case of traditional statistical surveys, where the values of the attributes are the answers provided by individuals to an interview, and it is also the case of most administrative data, where e.g. social security contributions paid in favour of workers are registered into transactional databases (figure 3). Internet of Things (IoT) streams of data and the clinical readings in the Artemis project falls in this category, and also in this case it is useful to think to sensor readings as answers to specific questions: The very characteristic of the so-called "structure-on-write" schema is that the data production process tracks exactly the construals needed by the data producer[4].

[4] Note that considering the map *m* as a set of questions avoids the point made by Fox and co-authors that some attributes are "given" instead than "measured", such as social security numbers or names. Strictly speaking we are not measuring them with some device, we just ask.

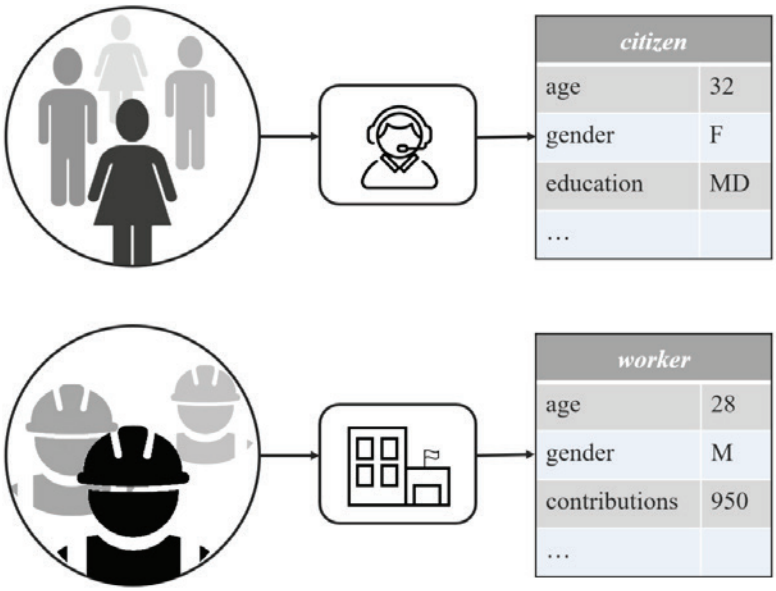


**Figure 3: Structure *on write* mapping**

Let us use the conceptual map to consider the case in which the purposes of a researcher are not aligned to the ones of the data producer. In general terms, following again Smith's check-list, the main critical aspects involve the survey design, the data quality monitoring and the framing of questions (see e.g. Johnson and Moore, 2005; Wallgren and Wallgren, 2007).

The issues about survey design have to do with what is the scope of the phenomenon of interest mapped by *m*. The representativity of the "found" data with respect to the interests of the researcher is presumably one of the most difficult issues in using, e.g., social media data to study public perceptions, and a key issue also with economic data of administrative source, e.g. in dealing with informality. From this point of view, however, provided that the map *m* is coherent with the purposes of a researcher, the proposed conceptualization does not solicit further considerations.

The issues about data quality have to do with the functioning of the map. The main point here is that while in statistical surveys data quality is mainstreamed in a consistent way, in the case of data collected for different purposes not all attributes may bear the same interest to the data producer, and

this usually has a large impact on quality. As an example, incomes are typically recorded with high accuracy in Tax and Social Security records, since it is a key information for the administrative purpose of the agencies collecting data; while the information about education – when recorded at all – has usually lower levels of quality (see e.g. Stüber *et al.*, 2023, and Adriaans *et al.*, 2020, for a study on German data and a review). Also from this point of view, the proposed conceptualization does not solicit further considerations with respect to the general literature about data quality.

Both representativeness and quality, then, may be dealt with during the statistical analysis of the data, e.g. discussing in a proper manner the external validity of the results and modelling errors in variables (as in Griliches' argument).

The third potential issue is about the framing of questions. One of the main disadvantages in using administrative data is that it is the data producer who chooses which questions to ask and typically its interests are very different with respect to the researchers' ones (Wallgren and Wallgren, 2007). A more subtle issue is that also the statistical units which are mapped may be different, and/or the "questions" used in the map may not measure exactly the concepts the researcher is interested in (see again Johnson and Moore, 2005, and Wallgren and Wallgren, 2007; and the thourough study in Kapteyn and Ypma, 2007).

This latter issue cannot be dealt with during the statistical analysis of the data, but requires a pre-processing of the data in order to bring their information content closer to the researcher's purposes. We can modify our conceptual map to represent this situation as in Figure 4. The primary map $m$ is the one used by the data producer to obtain the information needed for its administrative purposes. The researcher, when having the possibility of directly survey the phenomenon of interest, would have used a different map more coherent with its purposes (called $k$ in the figure), collecting possibly different attributes on different entities. In order to use the administrative data for the researcher's purposes, then, a secondary map is needed to tranform the original data in order to obtain an information set as close as possible to the desired one. The secondary map ideally would be $m^{-1}k$, in which case the use of found data would be equivalent to a direct survey of the phenomenon under study.

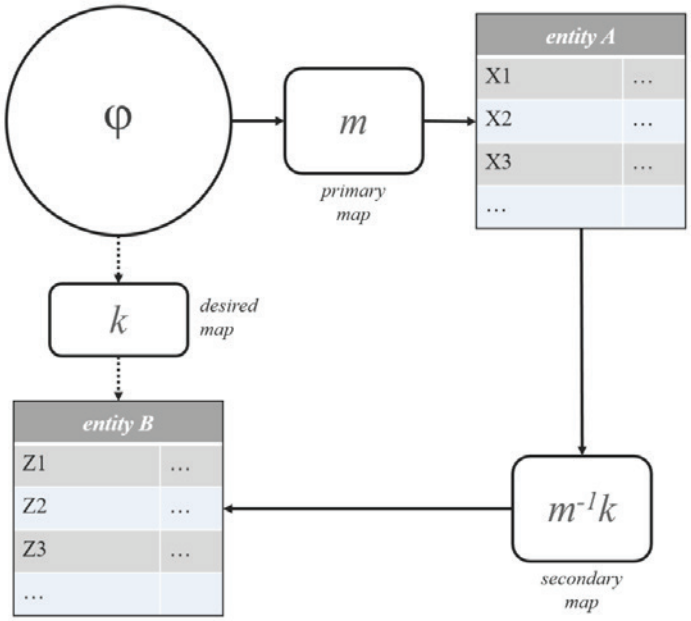


**Figure 3: Pre-processing of found data to mimick a direct survey of φ.**

In a sense, the secondary map is an "interview" posed not to the individuals in φ but to the administrative data, whose objective is to recover the kind of information that would have been obtained with a direct survey of φ applying the desired map *k*.

## 5. INTERVIEWING SOCIAL SECURITY RECORDS FOR LABOUR MARKET RESEARCH

In this section I will illustrate the need for a secondary mapping of administrative data in order to use them for research purposes, using as an example the WHIP-Health database on work and health biographies in Italy. The database is developed by the University of Turin and the Epidemiological Service of the Piedmont Region on behalf of Italian Ministry of Health, in cooperation with the many public institutions that have given access to their administrative data. It collects information of a 7% sample of the Italian population, with a longitudinal coverage of about 50 years as far as working careers are concerned, while the information on workers' health covers a more

limited period (from mid Nineties for work injuries and professional diseases and from early 2000s for hospital dismissal forms).

I will focus here in particular on work biographies and the framing of questions issue[5]. The section of the database dealing with working careers is based on the integration of more than 20 different files provided by INPS, the Italian Social Security Administration, relative to different typologies of work arrangements, different welfare provisions and employers' data; plus a database on work contracts' openings, closures and transformations provided by the Ministry of Welfare (MoW). For the sake of simplicity, to exemplify the framing of questions issue I will consider separatedly those regarding the definition of the units of observations (the entities) and those regarding the attributes construals, even though they are indeed linked with each other.

### 5.1 THE STATISTICAL UNITS

The main entities of interest for labour market studies are the two sides of the labour market itself, i.e. the individuals who offer their work services and the businesses and other institutions who may employ them.

With regard to the former, the use of administrative data entails a first hurdle. In a survey, each entity is extracted by sampling a population register in which a unique personal identifier is included. The data collection process implemented with a survey follows what we might call a top-down flow, in which one starts with the entities and then all the information of interest about them is collected with an interview. With a top-down schema, the integrity of the statistical unit definition and the association with its attribute-values is granted *by design*. In the case of administrative records, the data collection process typically follows a reversed, bottom-up flow. It is administrative transactions which are collected first – such as social security contributions collection or welfare benefits payment – and thanks to a personal identifier such as the fiscal code the data are then associated to the correct statistical unit. The integrity of the statistical unit definition hence rests entirely on the data quality of personal identifiers. Although in current information systems they possess a

[5] See e.g. Contini and Trivellato (2005) for a wider description and for applications of the work histories section of the database, and Bena *et al.* (2012) and Ardito *et al.* (2017) for the section on health biographies.

high level of certification, in longitudinal databases collecting retrospective data from legacy archives the very identification of individuals may be critical, leading to matching errors which corrupt the accuracy usually associated to administrative data (Kapteyn and Ypma, 2007). In this case, the secondary map that may be used is a probabilistic match trying to find apparently different individuals which, due to their personal characteristics and their career patterns, can be identified as the same individual. In the case of WHIP this kind of procedure was run up to the early 2000s, before INPS itself, in cooperation with the Tax Administration, started a systematic data cleansing activity of the fiscal codes used to identify persons in the individual registers.

With regard to employers, the situation is more complex. From an administrative point of view the employer is a legal entity, and no ambiguities are out there: The employer is the entity identified by the unique fiscal code which is paying the social security contributions. From a theoretical point of view, however, a researcher would like to analyse data about employers without being forced to a strictly legal definition of them. Depending on the research question, the interest could be e.g. on the local units of a firm, on firms belonging to a group, or on legal transformations of firms. In the case of Italian data all these details are not present, but some of these topics can be handled with a secondary mapping of the original data. Consider as an example relevant events such as ownership transfers**,** mergers and acquisitions, spin-offs and legal transformations. With the "legal entity" point of view of administrative data, all these events produce apparent firms' start ups and closures, even when there has not been a substantial discontinuity in the life of a firm. This generates "spurious" firm demography events, which hinders both the study of firm dynamics and a correct measurement of workers mobility and of job creation and destruction.

To cure the WHIP data about employers we implemented an algorithm which uses information about the flows of clusters of workers across different legal employers to identify longitudinal relationships in longitudinal business data (Pacelli and Revelli, 1995; see also Hethey-Maier and Schmieder, 2013). In the case of US business data, a similar correction in the statistical unit identification involved about 10-13% of all apparent worker flows (Benedetto *et al.*, 2007); in the current WHIP release this share is lower when compared to all job separations (5-8%), but it is a huge quota of apparently direct job to job transitions (between 40% and 50% in mid 2000s).

### 5.2 ATTRIBUTES CONSTRUAL

There are several attributes in the social security archives WHIP is based on which require a mapping from the original measures to construals closer to the interest of researchers. As an example, most income variables are not based on the net or gross measure of them – which are the ones typically of interest to researchers – and until recently the sector of activity and skill level were measured using old and/or non-standard classifications. These edits however, although important, do not raise particular methodological issues.

I will focus instead on a topic much investigated in current empirical literature about labour market dynamics, i.e., the long run trend towards an increase in precarious work arrangements. The issue of defining "precarity" is indeed a complex one both from a theoretical and an empirical point of view. A recent assessment of the literature found no widely accepted definition of it, while many operationalisations of the concept were actually an accommodation to the available data (Kreshpaj *et al*, 2020). Whatever the definition, however, it is fair to say that the duration of an employment relationship is a key dimension for the empirical operationalization of the concept.

To measure this important attribute within a sample survey, you can simply ask, as in ILO current recomendations: "*How long have you been employed by your current employer?*". Let us consider instead what are the "questions" available in our source data, as posed by INPS and by the MoW. The latter administration takes as a reference the legal aspect of the relationship, i.e. the labour contract between the employer and the employee. The legal basis for the data collection is the requirement for employers to communicate to the MoW all their hirings and firings, plus eventual tranformations of contracts, including their dates. We could then use this kind of information to directly answer our question about the job duration.

For the years prior to 2009 however MoW data are not available, hence one has to resort to the INPS' source. The collection of social security data does not originate from hirings and firings communications, but on contribution payments' forms, which entail a completely different data transaction[6]. The

[6] Starting from 2005, the digital transmission of contributory data has been radically changed, improving some critical aspects that I'm going to present. Since the WHIP data cover dependent employment starting from 1987, however, all the

issue here is that the relation between contributory forms and labour contracts is actually of a many-to-many type: as an example, when employing a seasonal worker with two different contracts in January and then in December of a given year, a firm would compile one single contributory form. The rationale is: The administration needs just to add the contributions to the previdential account of the worker, no matter if they originated from one or more contracts[7]. In this example, to trade a contributory form for a job contract would underestimate the mobility of workers between contracts, and precarity. From the other side, if a worker stays in the job but there is a change in some aspects of the job itself – e.g. the province of work – the employer has to compile two different contributory forms, leading in this case to an overestimation of workers mobility and precarity. In this case, to derive information about job contracts one needs a complicated remapping of the data, aimed at splitting/joining contributory forms in order to identify (in a probabilistic way) the job contracts which generated the contributory spells.

The situation, however, is more complex than this: similarly to the discussion regarding the entity "firm", one should also consider whether the object of interest for the researcher is really the legal aspect of a job (*i.e.* the employment contract) or some other construal. For the sake of simplicity, let us consider the MoW data, which measure with high accuracy hirings and firings, and consider a worker who had two successive contracts of one year with the same employer. Taking at its face value the MoW datum, at month 13 we would classify the worker as having a tenure of one month, instead than 13, which, depending on the research question, may not be appropriate. A similar issue has been considered in the US' Current Population Survey. Prior to 1983, CPS supplements on tenure asked workers "*When did you start working at your present job?*". The term "job" is itself an ambiguous one: A worker employed for 10 years promoted to a managerial position 1 year prior to the survey may have been counted as having 10 years or 1 year of tenure, depending on

arguments exposed still apply for the procedures handling the older decades of the work careers.

[7] Also the start of the labour contract is actually recorded in contributory forms but it is affected by a huge missing data problem, presumably because of the low administrative relevance of it and of the many-to-many nature of the relation, which implies a non univocal association between contributory forms and work contracts

whether s/he interpreted the tenure with the current employer or in the managerial position. The Bureau of Labor Statistics then switched the wording of the question to a formulation closer to the ILO one quoted above, creating a break in the job tenures' time series (US Bureau of Labor Statistics, 2022).

This is not a minor issue, since successive labour contracts with the same employer are a very common situation in the Italian labour market, with an increasing trend. It is actually now common to hire workers even on a daily basis, with the activation of two or three labour contracts within the same week-end. In the MoW data sample available in WHIP, the share of hirings with contracts which lasted one single day was 12% in the years 2014-2019, and one out of five had a complete tenure within one week. In this case, to consider different legal arrangements as if they were independent one from the other would overestimate the mobility of workers between contracts, and precarity.

To my knowledge, the issue about what is the precise notion (or the set of notions) of employment spell that should be used in labour studies has been rarely discussed, contrary, e.g., to the notion of unemployment. A reference for the debate can be found in the ILO *Employment Relationship Recommendation*, 2006 (EER). The EER is actually focused on the identification of a subordinate condition in cases of casual work arrangements and of concealed employment relationships, but it provides also a reference for a clarification of the concept. The EER operationalise it suggesting that "the determination of the existence of such a relationship should be guided primarily by the facts relating to the performance of work and the remuneration of the worker, notwithstanding how the relationship is characterized in any contrary arrangement, contractual or otherwise, that may have been agreed between the parties". The EER goes on requiring that the relationship has a certain duration and continuity, but with a very loose interpretation of these concepts. In Europe, prevailing interpretations consider as a unique employment relationship sequences of several short-term contracts, even when comprising intermittent working and non-working periods (Risak *et al*, 2013).

Although it apparently delivers a poorer information with respect to MoW data, the recording present in INPS data was actually sufficient to approximate a definition of employment relationships close to the ILO one. To derive this kind of information – i.e. to answer to the question "*How long have you been employed by your current employer?*" – instead of sticking to the information

about the start and end of each work contract one has to "interview" the administrative records, looking for sequences of contributory spells with a certain continuity involving the same individual and the same employer, in order to ascertain how long this employment relationship has been lasting.

## 6. CONCLUDING REMARKS

Administrative data are an early instance of the family of non statistical sources collectively labelled as big data, delivering structured information with a volume and velocity which granted them a long standing role in official statistics and academic research. The advent of a wave of novel data sources such as IoT and social media data triggered a discussion which is shedding light on issues which were already present in the literature but have long been under-explored, particularly as regards their use for academic research.

In this paper I reviewed this recent debate particularly about the clarification of the very notion of data. A point which has been stressed by many authors is the importance of explicitely viewing data as *capta*, i.e. as an outcome of research activities and not only as a raw material for subsequent analysis. An immediate consequence of this is the importance of documenting not only what data are from a formal point of view – which entities and which attributes they represent. All production process has to be documented, both to make clear what are the theoretical perspectives adopted in their collection and as a basis to facilitate their re-use.

I went on proposing a conceptual map of data coherent with this view, which I used as a basis to focus on the re-use of administrative data for economic research. The main point I stressed is the fact that the different purposes of the institution who "made" the data and the researcher who "found" them reveal themselves already in the framing of the questions at the very beginning of the data production process. This implies that both the statistical units and the attributes available in the data may differ in a substantial way from the ones of interest for the researcher. Using as an example the WHIP database on work and health biographies in Italy, I discussed the case of firm data demography, which in social security files is typically based on a legal definition of employers, implying an over-estimation of firm closures and openings and consequently of job creation and destruction; and the case of

tenure estimation, which again is often based on the legal definition of work contracts, implying a mismeasurement of workers mobility and precarity.

The relevance of these issues rests in the need of studying the functioning of labour markets avoiding the perils of data driven research. From a descriptive point of view, it is indeed interesting to have statistics about the flows of contracts based on their legal definition. Also for research, they may be important e.g. for studies about collective bargaining and the evaluation of labour market legislation. From a perspective focused on employment precarity, a measurement of labour market flows based on the legal representation of them may instead be significantly misleading. An employment relationships with a restaurant or a hotel with a weekend commitment may be based on a single vertical part-time contract or on a recurrent sequence of very short ones. We may well evaluate their relative stability in different ways, but from the point of view of tenure, of human capital accumulation and firm-specific experience, they are hardly distinguishable. In two studies about the impact of tenure and experience on work safety, a definition of employment relationship closer to the ILO construal allowed to detect that one of the health costs of precarious work is mediated by short average tenures and the shift between different employers and tasks (Giraudo *et al.*, 2016, and Bena *et al.*, 2013). A measure of the same risks based on a legal definition of tenure would have implied a mis-classification of workers, hiding or attenuating this potential health spillovers of labour market flexibilization. Similarly, a measure of unemployement based on a tentative reconstruction of the statistical definition of it allowed to identify a causal impact of long unemployment spells of cardiovascolar health, which was cancelled out sticking to the apparently precise administrative measure of it based on unemployment benefits recipiency (Ardito *et al.*, 2017). As Tukey put it in his seminal work which anticipated modern data science, it is "far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise" (Tuckey, 1962).

It is a bit of a paradox that the clarification of the notion of data has received several contributions coming from biology, literary and information science and philosophy, while statistics and data science are among the few fields in which we do not ask ourselves what data are, and when we do it we stick to a definition of it close to the prevailing one in database theory. This

may be because the process of statistical evidence production was already grounded on established theories and practices which materialized themselves into specialised institutes for the delivery of statistical surveys, as summed up by Smith (1996).

The advent of big data and the diffusion of stand-alone uses of administrative data requires to put back the accent on the data production process and on the purposes and perspectives on which it is based on. As Griliches put it, the first question we should pose is not which models we can run on the data, but what are the data telling to us, since it is not us who framed the questions.

This is certainly not an easy task. Among the challenges to fully grasp the opportunities of big data, there are the issues linked to the data protection legislation ruling the processing of personal and sensitive data. The two conflicting aims of privacy protection and data needs for scientific research currently translates in a trade-off between the accessibility to anonymised microdata distributed as public use files, but poor in information detail, and the richness of confidential data that is however accessible – when accessible at all – only on the premises of the data holder. Safe and accessible data has to pay the price of a substantial reduction in the information content available in the source data (Trivellato, 2019), and this hinders not only their use but also an effective activity of data wrangling and understanding.

The regulations about personal data is actually a wider issue, posing a potential obstacle to the full exploitation also of census and survey data particularly in the field of economic policies evaluation (see Crato and Paruolo, 2019, for a recent assessment). A further point, more specific to the use of administrative data, is what we can define a "missing-metadata" issue. It is a common experience for practitioners in the field that data are provided "as are", without a full clarification of the concepts besides a bare schema of the database of origin and of the query which was used to fetch data. The point has not a straightforward solution, since the framing of questions, in the case of public administrations, is embedded into complex layers of different laws, decrees and regulations more than on choices from the part of actual data managers. To "interview" administrative data, trying to design a secondary map of the data answering the needs of a specific research question, can then serve a dual purpose. It is the necessary step to derive information from the data as coherent as possible with the construals of interest, and a possible basis for a

clear and comprehensible documentation of the data themselves, using as a way of documenting data the traditional and well known form of a questionnaire, listing the "questions" posed not to actual respondents but to the administrative source data.

## REFERENCES

Adriaans, J., Valet, P., and Liebig, S. (2020). Comparing administrative and survey data: Is information on education from administrative records of the German Institute for Employment Research consistent with survey self-reports? *Quality & Quantity*, *54*(1), 3–25. https://doi.org/10.1007/s11135-019-00931-4

Al-Sai, Z. A., Abdullah, R., and Husin, M. H. (2019). Big data impacts and challenges: A review. *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, 150–155. https://doi.org/10.1109/JEEIT.2019.8717484

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired*. https://www.wired.com/2008/06/pb-theory/

Antenucci, D., Cafarella, M., Levenstein, M., Ré, C., & Shapiro, M. D. (2014). *Using Social Media to Measure Labor Market Flows* (Working Paper 20010). National Bureau of Economic Research. https://doi.org/10.3386/w20010

Ardito, C., D'Errico, A., Leombruni, R., Ricceri, F., & Costa, G. (2020). Life expectancy inequalities and their evolution in Italy. How these impact on the equity of the pension system? *European Journal of Public Health*, *30*(Supplement_5), ckaa165.764. https://doi.org/10.1093/eurpub/ckaa165.764

Ardito, C., Leombruni, R., Mosca, M., Giraudo, M., & d'Errico, A. (2017). Scar on my heart: Effects of unemployment experiences on coronary heart disease. *International Journal of Manpower*, *38*(1), 62–92. https://doi.org/10.1108/IJM-02-2016-0044

Balazka, D., & Rodighiero, D. (2020). Big data and the little big bang: An epistemological (r)evolution. *Frontiers in Big Data*, *3*. https://www.frontiersin.org/articles/10.3389/fdata.2020.00031

Bena A, Giraudo M, Leombruni R., and Costa G. (2013). Job tenure and work injuries: A multivariate analysis of the relation with previous experience and differences by age. *BMC Public Health*, *13*, 1–9. https://doi.org/10.1186/1471-2458-13-869

Bena, A., Leombruni, R., Giraudo, M., & Costa, G. (2012). A new Italian surveillance system for occupational injuries: Characteristics and initial results. *American Journal of Industrial Medicine*, *55*(7), 584–592. https://doi.org/10.1002/ajim.22025

Benedetto, G., Haltiwanger, J., Lane, J., & McKinney, K. (2007). Using worker flows to measure firm dynamics. *Journal of Business & Economic Statistics*, *25*(3), 299–313. https://doi.org/10.1198/073500106000000620

Blount, M., Ebling, M. R., Eklund, J. Mikael., James, A. G., McGregor, C., Percival, N., Smith, K., & Sow, D. (2010). Real-time analysis for intensive care: development and deployment of the Artemis analytic system. *IEEE Engineering in Medicine and Biology Magazine*, *29*(2), 110–118. https://doi.org/10.1109/MEMB.2010.936454

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. The MIT Press. http://catalogue.bnf.fr/ark:/12148/cb44284618q

Cackett, D., Bond, A., & Gouk, J. (2013). *Information Management and Big Data: A Reference Architecture.* Oracle Corporation.

CEDEFOP. (2019). *Online Job Vacancies and Skills Analysis*. CEDEFOP. https://www.cedefop.europa.eu/en/publications/4172

Chetty, R. (2012). Time trends in the use of administrative data for empirical research. *34th Annual NBER Summer Institute. Cambridge, Mass.(July 9–27, 2012)*.

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

Conti, R., Michele, B., Boscolo, S., Puccioni, C., Ricchi, O., Tedeschi, S., & Fabrizi, E. (2023). The Italian Treasury Dynamic Microsimulation Model (T-DYMM): Data, structure and baseline results. *Italian Department of the Treasury Working Paper Series WORKING PAPERS*, *1*.

Contini, B., & Revelli, R. (1986). Natalità e mortalità delle imprese italiane: Risultati preliminari e nuove prospettive di ricerca. *L'industria*, *2*, 195-.

Contini, B., & Revelli, R. (1987). The process of job creation and job destruction in the Italian economy. *Labour*, *1*(3), 121–144. https://doi.org/10.1111/j.1467-9914.1987.tb00122.x

Contini, B., & Trivellato, U. (2006). *Eppur si muove. Dinamiche e persistenze nel mercato del lavoro italiano*. Il Mulino.

Crato, N., & Paruolo, P. (A c. Di). (2019). *Data-Driven Policy Impact Evaluation: How Access to Microdata is Transforming Policy Design*. Springer International Publishing. https://doi.org/10.1007/978-3-319-78461-8

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V. R., Tsou, A., Weingart, S., & Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*, *66*(8), 1523–1545. https://doi.org/10.1002/asi.23294

Engstrom, R., Hersh, J., & Newhouse, D. (2022). Poverty from space: Using high resolution satellite imagery for estimating economic well-being. *The World Bank Economic Review*, *36*(2), 382–412. https://doi.org/10.1093/wber/lhab015

Floridi, L. (2008). Data. In W. A. Darity (A c. Di), *International Encyclopedia of the Social Sciences*.

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, *165*, 234–246. https://doi.org/10.1016/j.ijpe.2014.12.031

Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. *Information Processing & Management*, *30*(1), 9–19. https://doi.org/10.1016/0306-4573(94)90020-5

Frické, M. (2015). Big data and its epistemology. *Journal of the Association for Information Science and Technology*, *66*(4), 651–661. https://doi.org/10.1002/asi.23212

Gauvain, J.-L., Lamel, L., & Adda, G. (2000). Transcribing broadcast news for audio and video indexing. *Communication of the ACM*, *43*, 64–70. https://doi.org/10.1145/328236.328148

Gellert, R. (2022). Comparing definitions of data and information in data protection law and machine learning: A useful way forward to meaningfully regulate algorithms? *Regulation & Governance*, *16*(1), 156–176. https://doi.org/10.1111/rego.12349

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), Articolo 7232. https://doi.org/10.1038/nature07634

Giraudo M, Bena A, Leombruni R., & Costa G. (2016). Occupational injuries in times of labour market flexibility: The different stories of employment-secure and precarious workers. *BMC Public Health*, *16(1)*, 1–11. https://doi.org/10.1186/s12889-016-2834-2

Gitelman, L. (2013). *Raw Data Is an Oxymoron*. MIT Press.

Griliches, Z. (1984). *Data Problems in Econometrics* (SSRN Scholarly Paper 300716). https://papers.ssrn.com/abstract=300716

Griliches, Z. (1985). Data and econometricians—The uneasy alliance. *The American Economic Review*, *75*(2), 196–200.

Griliches, Z. (1986). Comment on Behrman and Taubman. *Journal of Labor Economics*, *4*(3), S146–S150.

Harford, T. (2014). Big data: A big mistake? *Significance*, *11*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2014.00778.x

Hethey-Maier, T., & Schmieder, J. F. (2013). *Does the Use of Worker Flows Improve the Analysis of Establishment Turnover? Evidence from German Administrative Data* (Working Paper 19730). National Bureau of Economic Research. https://doi.org/10.3386/w19730

Hill, C. A., Biemer, P. P., Buskirk, T. D., Japec, L., Kirchner, A., Kolenikov, S., & Lyberg, L. E. (2020). *Big Data Meets Survey Science: A Collection of Innovative Methods*. John Wiley & Sons.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Hjørland, B. (2019). Data (with big data and database semantics). *KO Knowledge Organization*, *45*(8), 685–708.

Jensen, Howard E. 1950 "Editorial note." In H.P.Becker *Through Values to Social Interpretation: Essays on Social Contexts, Actions, Types, and Prospects*. Durham, NC: Duke University Press, vii–xi

Johnson, B., & Moore, K. (2005). Consider the source: Differences in estimates of income and wealth from survey and tax data. *Special Studies in Federal Tax Statistics*, 77–99.

Kapteyn, A., & Ypma, J. Y. (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics*, *25*(3), 513–551. https://doi.org/10.1086/513298

Khaouja, I., Kassou, I., & Ghogho, M. (2021). A survey on skill identification from online job ads. *IEEE Access*, *9*, 118134–118153. https://doi.org/10.1109/ACCESS.2021.3106120

Koopmans, T. C., Vining, R., & Hastay, M. (1995). 'Measurement without theory' debate (Review of Economics and Statistics, vol. 29, 1947, pp. 161–72 (cut); vol. 31, 1949, pp. 77–93 (cut); and Journal of the American Statistical Association, vol. 46, 1951, pp. 388–90). In D. F. Hendry & M. S. Morgan (A c. Di), *The Foundations of Econometric Analysis* (pp. 491–524). Cambridge University Press. https://doi.org/10.1017/CBO9781139170116.046

Kreshpaj, B., Orellana, C., Burström, B., Davis, L., Hemmingsson, T., Johansson, G., Kjellberg, K., Jonsson, J., Wegman, D. H., & Bodin, T. (2020). What is precarious employment? A systematic review of definitions and operationalizations from quantitative and qualitative studies. *Scandinavian Journal of Work, Environment & Health*, *46*(3), 235–247.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: traps in big data analysis. *Science*, *343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Leonelli, S. (2014). What difference does quantity make? On the epistemology of big data in biology. *Big Data & Society*, *1*(1), 2053951714534395. https://doi.org/10.1177/2053951714534395

Leonelli, S. (2015). What counts as scientific data? A relational framework. *Philosophy of Science*, *82*(5), 810–821. https://doi.org/10.1086/684083

Leonelli, S. (2019). What distinguishes data from models? *European Journal for Philosophy of Science*, *9*(2), 22. https://doi.org/10.1007/s13194-018-0246-0

Llorente, R., Morant, M., Macho, A., Garcia-Rodriguez, D., & Corral, J. L. (2015). Demonstration of a spatially multiplexed multicore fibre-based next-generation radio-access cellular network. *2015 17th International Conference on Transparent Optical Networks (ICTON)*, 1–4. https://doi.org/10.1109/ICTON.2015.7193681

Lohr, S. L., & Brick, J. M. (2017). Roosevelt predicted to win: Revisiting the 1936 Literary Digest poll. *Statistics, Politics and Policy*, *8*(1), 65–84. https://doi.org/10.1515/spp-2016-0006

Lusinchi, D. (2012). "President" Landon and the 1936 Literary Digest poll: Were automobile and telephone owners to blame? *Social Science History*, *36*(1), 23–54. https://doi.org/10.1017/S014555320001035X

Man, X., Luo, T., & Lin, J. (2019). *Financial Sentiment Analysis (fsa): A survey*. 617–622.

McGregor, C., Catley, C., James, A., & Padbury, J. (2011). Next generation neonatal health informatics with Artemis. *Studies in Health Technology and Informatics*, *169*, 115–119.

Pacelli, L., & Revelli, R. (1995). Trasformazioni societarie, scorpori, fusioni: Un metodo di individuazione mediante dati di fonte Inps. Biffignandi S. Martini M., Il registro statistico delle imprese.

Petrelli, A., Sebastiani, G., Di Napoli, A., Macciotta, A., Di Filippo, P., Strippoli, E., Mirisola, C., & d'Errico, A. (2022). Education inequalities in cardiovascular and coronary heart disease in Italy and the role of behavioral and biological risk factors. *Nutrition, Metabolism and Cardiovascular Diseases*, *32*(4), 918–928. https://doi.org/10.1016/j.numecd.2021.10.022

Polgreen, P. M., Chen, Y., Pennock, D. M., & Nelson, F. D. (2008). Using internet searches for influenza surveillance. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *47*(11), 1443–1448. https://doi.org/10.1086/593098

Redman, T. C., Fox, C., & Levitin, A. (2017). *Data and Data Quality*. Routledge Handbooks Online. https://doi.org/10.1081/E-ELIS4-120008897

Risak, M., Rauws, W., Sredkova, K., & Portmann, W. (2013). *Regulating the Employment Relationship in Europe: A Guide to Recommendation No. 198*. ILO/ELLN.

Saberi, B., & Saad, S. (2017). Sentiment analysis or opinion mining: A review. *Int. J. Adv. Sci. Eng. Inf. Technol*, *7*(5), 1660–1666.

Shin, S.-Y., Kim, T., Seo, D.-W., Sohn, C. H., Kim, S.-H., Ryoo, S. M., Lee, Y.-S., Lee, J. H., Kim, W. Y., & Lim, K. S. (2016). Correlation between national influenza surveillance data and search queries from mobile devices and desktops in South Korea. *PLoS ONE*, *11*(7), e0158539. https://doi.org/10.1371/journal.pone.0158539

Smith, A. F. M. (1996). Mad cows and ecstasy: Chance and choice in an evidence-based society. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *159*(3), 367–383. https://doi.org/10.2307/2983324

Soren, S., & Peter, F. (2018). *What is the State of the Manufacturing Sector in Mozambique?* (Availability Note: Information provided in collaboration with the RePEc Project: http://repec.org).

Stüber, H., Grabka, M. M., & Schnitzlein, D. D. (2023). A tale of two data sets: Comparing German administrative and survey data using wage inequality as an example. *Journal for Labour Market Research*, *57*(1), 8. https://doi.org/10.1186/s12651-023-00336-9

Triplett, J. E. (2007). Zvi Griliches' contributions to economic measurement. In *Hard-to-Measure Goods and Services: Essays in Honor of Zvi Griliches* (pp. 573–589). University of Chicago Press. https://www.nber.org/books-and-chapters/hard-measure-goods-and-services-essays-honor-zvi-griliches/zvi-griliches-contributions-economic-measurement

Trivellato, U. (2019). Microdata for social sciences and policy evaluation as a public good. In N. Crato & P. Paruolo (A c. Di), *Data-Driven Policy Impact Evaluation: How Access to Microdata is Transforming Policy Design* (pp. 27–45). Springer International Publishing. https://doi.org/10.1007/978-3-319-78461-8_3

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67.

US Bureau of Labor Statistics. (s.d.). *Employee Tenure Technical Note—2022 A01 Results*. Retrieved 8 october 2023, https://www.bls.gov/news.release/tenure.tn.htm

Wallgren, A., & Wallgren, B. (2007). *Register-based Statistics: Administrative Data for Statistical Purposes*. John Wiley & Sons.

Yovits, M. C. (1969). Information science: Toward the development of a true scientific discipline. *American Documentation*, *20*(4), 369–376. https://doi.org/10.1002/asi.4630200421

Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, *58*(4), 479–493. https://doi.org/10.1002/asi.20508