I wrote this section with the expectation that an introduction to the practice would be included in the introduction to the article; something like this:

National statisticians began to engage in experiments with Big Data as an alternative data source for the generation of official statistics especially at the beginning of 2013. This involved investigating the potential of data generated by internet platforms (e.g., microblogging, search engines) and devices (e.g., mobile phones) as alternatives to traditional sources of data such as surveys and questionnaires for knowing population phenomena such as movements, opinions and relations. Examples of these initial experiments included analysing location data from mobile phones for generating ‘real-time’ population and tourism statistics; search query trends for generating data on migration; social media messages for generating indicators on issues such as consumer confidence; and price data scraped off the Internet for producing inflation statistics.

# Official statistics: From fundamental principles to an ethic of care

Bearing in mind also that the essential trust of the public in the integrity of official statistical systems and confidence in statistics depend to a large extent on respect for the fundamental values and principles that are the basis of any society seeking to understand itself and respect the rights of its members, and in this context that professional independence and accountability of statistical agencies are crucial (UN Fundamental Principles of Official Statistics, Adopted 29 January 2014).

Amongst national statisticians, there is much discussion of and reference to the principles and codes of practice that guide the production of official statistics. These references draw on the UN Fundamental Principles of Official Statistics, which addresses the values of utility, professional standards and ethics, scientific principles, transparency, quality, timeliness, costs, respondent burden, and confidentiality (UN 2014).[[1]](#footnote-1) These principles are echoed in the European Statistics Code of Practice, which include professional independence, quality, confidentiality, impartiality, objectivity, non-excessive burden on respondents, cost effectiveness, relevance, accuracy, reliability, timeliness, comparability, accessibility and clarity (ESSC 2011).[[2]](#footnote-2) Effectively the principles have been promulgated to secure the validity of and trust in official statistics as well as the ethics of their production.

The relevance of these principles was evident at the collaboratory when national statisticians appealed to them in relation to their assessments of the possibilities of Big Data. What the discussion made apparent is that rather than ‘applying’ the principles they require rethinking to address the different provenance and qualities of Big Data. Yet, at the same time, evaluations and interpretations of Big Data repeatedly fell back on the principles as criteria for assessing their voracity and value. In other words, the principles – not just traditional forms of data and methods – persist in their ways of thinking about data, its value and meaning. In this respect, what the framework of social intelligence and ethic of care offer is a way to recast the principles through the recognition of new kinds of togetherness that make up Big Data and the responsibilities that this entails: to how Big Data is composed and to the subjects who participate in their composition. In brief the responsibility to the relations of composition constitutes a move from *capturing* ‘snapshots’ of populations through already and pre-established categories and definitions to *monitoring* the becoming of relations as they form and reform. An ethic of care constitutes a move from an ethical responsibility to relations confined to questions of privacy and confidentiality to one that attends to relations as situated and relational achievements.

## What does a social intelligence mean in relation to official statistics?

One meaning of social intelligence concerns the relative authority of sources of data for knowing societies. On the one hand, sources of Big Data are heralded as new or better knowledge but on the other, they are typically evaluated as not meeting principles such as quality, stability, reliability, and transparency. Yet, beyond these criteria for assessing the relative authority or knowledge of data sources, there are other ‘structuring’ conditions that configure what Big Data gets taken up. As many statisticians noted what comes to be evaluated and potentially count as a source is increasingly a product of what data is in circulation and its availability and stability over time. From restrictions imposed by data owners to those built into platforms, what might become a source of official statistics is ordered by practices such as the commodification of data and the decisions of others. At the same time, the ordering – or perhaps prioritizing of sources – also involves statisticians’ assessments of a potential use and especially policy and practical application. They may juggle how they prioritise principles or criteria of evaluation (timeliness vs quality?), what data is analysed (Twitter or search engine?), or what methods are deployed (sentiment analysis or word frequency?), that is, attend to the things they can control. So, while principles are significant means of evaluating sources other practical and political orderings are also at work. Collectively, sets of social and technical relations of which statisticians are a part configure what can potentially form a social intelligence of societies. Statisticians are thus only one of many actors. Together with others they make ‘cuts’ through practices such as their deployment of principles, methods, decisions, circulation choices, and so on that configure what comes to form a social intelligence.[[3]](#footnote-3)

This points to a second meaning of social intelligence that arises from another observation of statisticians: that Big Data is inherently different, and not of the same phenomena or kind of knowledge as traditional forms of data. That is, it is incomparable and good for something else that calls for a different way of thinking and what some statisticians called a paradigm shift. A common description of this shift goes as follows: it constitutes a change in the production of statistics from ‘design then collect’ to ‘collect then design.’ Traditionally statistics are defined in advance by the questions statisticians ask and who then design methods to collect data that can answer those questions. Instead, Big Data is understood as the driver of questions, as in the now popular claim that knowledge is data-driven. Instead of imposing classification schemes a priori, Big Data involves distinguishing meaningful patterns and asking new questions of data that were previously not possible. In other words, it calls for a search paradigm, not for metrics defined in advance but for patterns and trends that emerge and serve not as measures but as signals, approximations, and possibles. As one statistician noted ‘In reality most statistical offices neither have the resources, nor the interest to trace specific people. Instead, they are interested in discovering trends and aggregates.’ It is thus a form of intelligence that is not the same as the traditional classificatory practices they usually work with and as such pose sense makingproblems that call for new principles and forms of reasoning.

The alternative form of reasoning that a search paradigm suggests is perhaps captured in Herbert Simon’s (Simon, 1947) conception of ‘satisficing’ rather than ‘optimizing’ or ‘maximizing’ in decision-making. In opposition to abstractions such as utility theory he advanced an understanding of how people reason in practice. That is, decisions involve juggling numerous criteria and arriving at a ‘good-enough’ solution rather than engaging in an infinite search for all the possible ones, evaluating them and then arriving at the best one. But rather than a solutionist answer, Jennifer Gabrys and Helen Pritchard (Gabrys and Pritchard, 2015) offer that the adequacy of an answer depends on the question. They instead define ‘just good enough data’ to counter the reliance on measurement accuracy as the only objective and criterion for evaluating environmental data gathered through citizen sensing practices. Instead, they argue that measurements meet different objectives, which are often not known in advance such as just identifying that a pollution event is happening or the emergence of unexpected relations, connections and points of concern. What Gabrys and Pritchard draw attention to is that the potential uses or value of data is often not known in advance and data collection and interpretation is also a search for its potential rather than already known value. It constitutes a move from settling things in advance such as categories and definitions to instead searching for them in data.

This is a social intelligence or knowing that is attentive and responsive to and keeps open the possibilities of new social formations, relations ways of living and being rather than foreclosing them. It also keeps open the kinds of insights and forms of knowledge that might emerge. But it further relates to the observation noted above that Big Data is not the same as other data; the search paradigm is not a search for what is already known through traditional sources of data. It is for knowledge that emerges from social processes and relations; not socially developed collective knowledge (as understood as collective intelligence) but an intelligence composed out of relations between people and technologies.

Of course, all data is composed of relations and interdependencies but what the observation suggests is that something is different with Big Data. In addition to the redistribution and reassembly of arrangements, actors, practices and processes statisticians made the further observation that it is closer to what they refer to as the ‘ground truth’, to the actual behaviour of people. Rather, we would suggest a third meaning of social intelligence here, that Big Data is an intelligence about everyday digital relations and sociality; it is an intelligence that comes out of the relations between people and technologies as they conduct their lives. The difference can be illustrated through a traditional practice of statisticians, which is to classify people. Traditional practices involve defining possible categories a priori and then identifying and fitting people into them. But rather than stable these change dynamically and subjects persistently tend to escape or overflow categories (Ruppert, 2012). In contrast, with Big Data identifying and classifying are done through different procedures of affinity and attraction that emerge from everyday digital relations. Similarities between people are defined by their attraction to or liking (purchasing, viewing, etc.) of similar things or by algorithms that have affinities to particular people as they detect, sort and classify them. While both have feedback loops – traditional categories are adjusted as ‘new’ ones appear and algorithmic sorting rules are modified by the data they generate - the former is usually a ‘correction’ whereas the latter the logic and basis of category formation.

Each of these understandings of social intelligence focuses on the myriad relations that compose Big Data. The cuts, patterns and relations that make up social intelligence carry with them or inherit ‘a web of relationalities that contributed to make them possible’ (Puig de la Bellacasa, 2012) and it is this inheritance that forms the basis of an ethic of care.

## What is an ethic of care for official statistics?

Statisticians express care for official statistics in two key ways. One concerns the principles, which indicate a form of caring for data, such as its quality, accessibility and clarity. But the principles also involve caring for their relations and accountabilities to subjects through data protection, confidentiality, consent, and trust. While the origins of these principles are a mix of legal, governmental, political and professional rationales and requirements, they tend to operate as values and commitments. This is evident in claims made by statisticians such as ‘just because you can, doesn’t mean you should’ use Big Data sources.

Accountability for official statistics, as expressed in the principles, is to the conditions of their generation – as stated in the opening quote – and not to how they are used and what is done with them. It is based on the assumption that those conditions can be more-or-less determined and controlled by a statistical office. While debatable, the point we wish to make here is that the consequences or what is done with official statistics is not part of the principles/ethics. Instead, they attend to the processes and practices of their generation. This is a distinction similar to that which has been drawn between procedural ethics and ethics in practice within the social sciences (Guillemin and Gillam, 2004). Procedural ethics are understood as an estimation of the ethical issues that might be involved in research whereas ethics in practice involve the recurrent, iterative, and uncertain ethical moments that happen in the field and which may be odds with that covered procedural ethics review.

What then does this mean in relation to Big Data especially when knowing and worlds are emergent and potential effects are not possible to know in advance? If anything, Big Data makes it even less possible to attend to consequences; instead, it and the search paradigm demand remaining ‘attentive to the unknown knocking at our door’ (Deleuze, 1992), which requires a way of caring for relating to the new (Puig de la Bellacasa, 2012).

The ethic of care that we suggest arises from Puig de la Bellacasa’s call for the care of relations themselves on which Big Data would not exist and could not be maintained without. It is an ontological ethic that imposes the responsibility to act in ways that account for and are accountable to human and non-human relationality. It is an ethic she develops through a re-reading of Donna Haraway’s (Haraway, 1997) take on feminist discussions of the situatedness of knowledge. For Haraway, knowing and thinking are only possible through mulitiple and situated relations and ‘nothing comes without its world’; for Puig de la Bellacasa then, relations of thinking and knowing require care. If knowing is to be ethical then it must take care and address the situated relations through which it has emerged. This is an ontological rather than moral argument; it attends to the dependence on the relations and labours of others to ‘create, hold together and sustain’ (198) social worlds that is not a choice but a condition of knowledge. For Puig de la Bellacasa, this is not a controlling or normative practice but one that acknowledges the relations that compose Big Data and their situatedness.

One way to explore this comes from a principle that statisticians often referenced in the collaboratory: reducing ‘respondent burden’ by only requiring data that is ‘absolutely necessary’ and for which data does not already exist. The implication is that to respond to a questionnaire is a relation that demands a response and imposes a burden on the subject. By exploiting existing sources such as Big Data this burden can be reduced. However, this does not do away with relations to or actions by subjects; it only displaces them. Indeed, anew relationality is introduced; rather than a relation to a response (a question, a tick box) data is the result of a relation to an action (a tweet, a search, a purchase, a phone call). In the case of the former, responses are scrutinized for how subjects are similar to the categories they are expected to inhabit. Data is then evaluated in terms of accuracy, reliability, completeness, and so on. For the latter, actions are scrutinized for the similarities that emerge and categories they suggest. How then might data about such a relation be evaluated? While suspending disbelief for a moment that such a sharp distinction is possible,[[4]](#footnote-4) some further observations made by statisticians return us again to a social intelligence and ethic of care. One statistician noted that ‘users could modify their behaviour to avoid detection (e.g. opt out, rejecting cookies, usage of Google search), which would impact the future success of these methods.’ And another stated ‘Statisticians need to be aware that users’ behaviour can change and that this can create continuity issues. If users were to hide their digital trail as a result of privacy concerns, this would diminish the value of certain data sources.’ While problematizing the actions of subjects, a social intelligence and ethic of care calls for a mode of interpretation that enables, situates, accounts for and knows populations through these changing relations. This is perhaps also the ethic of a search paradigm discussed as part of a social intelligence. Attending to patterns and monitoring the various relations that compose data acknowledges their interdependencies and enables difference to emerge and change.

European Statistical System Committee (ESSC) (2011). European Statistics Code of Practice for the National and Community Statistical Authorities. Luxembourg, Eurostat.

Deleuze G. (1992) What Is a Dispositif? In: Armstrong T (ed) *Michel Foucault, Philosopher.* New York: Routledge.

Gabrys J and Pritchard H. (2015) Just Good Enough Data and Environmental Sensing: Moving Beyond Regulatory Benchmarks toward Citizen Action. In: Berre AJ, Schade S and Piera J (eds) *Environmental Infrastructures and Platforms 2015 - Infrastructures and Platforms for Environmental Crowd Sensing and Big Data.* Barcelona.

Guillemin M and Gillam L. (2004) Ethics, reflexivity, and “ethically important moments” in research. *Qualitative Inquiry* 10: 261-280.

Haraway DJ. (1997) *Modest\_Witness@Second\_Millennium,* New York and London: Routledge.

Puig de la Bellacasa M. (2012) ‘Nothing comes without its world’: thinking with care. *The Sociological Review* 60: 197-216.

Ruppert E. (2012) Category. In: Lury C and Wakeford N (eds) *Inventive Methods: The Happening of the Social.* London: Routledge, 36-47.

Simon H. (1947) *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*: Macmillan.

United Nations (2014) ‘Fundamental Principles of Official Statistics’. Resolution adopted by the General Assembly on 29 January 2014. A /RES/68/261

1. Six principles cover that official statistics must meet the test of practical utility; be developed according to strictly professional considerations, scientific principles and professional ethics; present information on the scientific standards of their sources, methods and procedures; may be generated from all types of sources such as surveys or administrative records and the source chosen with regard to quality, timeliness, costs and the burden on respondents; are to be strictly confidential and used exclusively for statistical purposes; and the laws, regulations and measures governing them are public. [↑](#footnote-ref-1)
2. Others include a clear legal mandate to collect data for statistical purposes, adequate resources, sound methodology and procedures. [↑](#footnote-ref-2)
3. Marres (2012) describes these changing actors and relations as involving a redistribution of methods and Ruppert et. al (2013) as a reassembling of methods. [↑](#footnote-ref-3)
4. What the discussions at the collaboratory further suggested is that such an approach relies on hunches, guesses, intuition and speculative searches, which are not independent of hypotheses, theories, assumptions, and pre-conceived notions. This is especially evident in how statisticians interpreted and differentiated between ‘signals’ and ‘noise’ or ‘babble’. [↑](#footnote-ref-4)