ECONOMY AND SOCIETY

Volume 44 Number 3 August 2015

- Louise Amoore and Volha Piotukh/Life beyond big data: governing with little analytics
- Benjamin Braun/Governing the future: the European Central Bank's expectation management during the Great Moderation
- Dean Curran/Risk illusion and organized irresponsibility in contemporary finance: rethinking class and risk society
- Nicholas Copeland/Regarding development: governing Indian advancement in revolutionary Guatemala
- Bregje F. van Eekelen/Accounting for ideas: Bringing a knowledge economy into the picture
- Grahame Thompson/Review Essay: Socializing the Constitution?

Routledge Taylor & Francis Group

ISSN: 0308-5147

Economy and Society

Routledge Taylor & Francis Group

ISSN: 0308-5147 (Print) 1469-5766 (Online) Journal homepage: http://www.tandfonline.com/loi/reso20

Life beyond big data: governing with little analytics

Louise Amoore & Volha Piotukh

To cite this article: Louise Amoore & Volha Piotukh (2015) Life beyond big data: governing with little analytics, Economy and Society, 44:3, 341-366, DOI: 10.1080/03085147.2015.1043793

To link to this article: http://dx.doi.org/10.1080/03085147.2015.1043793

Published online: 04 Sep 2015.

Economy and Society Volume 44 Number 3 August 2015: 341–366
http://dx.doi.org/10.1080/03085147.2015.1043793

Life beyond big data: governing with little analytics

Louise Amoore and Volha Piotukh

Abstract

The twenty-first-century rise of *big data* marks a significant break with statistical notions of what is *of interest* or concern. The vast expansion of digital data has been closely intertwined with the development of advanced analytical algorithms with which to make sense of the data. The advent of techniques of knowledge discovery affords some capacity for the analytics to derive the object or subject of interest from clusters and patterns in large volumes of data, otherwise imperceptible to human reading. Thus, the scale of the *big* in big data is of less significance to contemporary forms of knowing and governing than what we will call the *little analytics*. Following Henri Bergson's analysis of forms of perception which 'cut out' a series of figures detached from the whole, we propose that analytical algorithms are instruments of perception without which the extensity of big data would not be comprehensible. The technologies of analytics focus human attention and decision on particular persons and things of interest, whilst annulling or discarding much of the material context from which they are extracted. Following the algorithmic processes of ingestion, partitioning and memory, we illuminate how the use of analytics engines has transformed the nature of analysis and knowledge and, thus, the nature of the governing of economic, social and political life.

Keywords: analytics; algorithm; big data; knowledge discovery; Bergson; technology.

Louise Amoore, Department of Geography, Durham University, South Road, Durham DH1 3LE, United Kingdom. E-mail: louise.amoore@dur.ac.uk
Volha Piotukh, Department of Geography, Durham University, South Road, Durham DH1 3LE, United Kingdom. E-mail: volha.piotukh@durham.ac.uk

Routledge
Taylor & Francis Group

To sew the pieces together

> In an amorphous space he carves out moving figures, or else, he imagines relations of magnitude which adjust themselves one to another [...] But it is not enough to cut out, it is necessary to sew the pieces together. You must now explain how those qualities which you have detached from their material support can be joined to it again (Bergson, 1912, p. 32).

> The advantage is we can throw all the data at algorithms and the algorithm sorts it, picks out the strongest relationships (SAS Analytics, 2013).

In a crowded analytics workshop in London in 2013, a data analyst demonstrates the techniques that may 'allow a user, without having to code, to segment data, to rejoin data together and to get insight into that data'. In a world of big data, he tells the assembled crowd, what matters is the capacity for businesses and governments to make sense of the data, to 'throw it at algorithms' such that the strongest relationships between elements can be identified. Understood in these terms, the work of the analytics is concerned with cutting out pieces from across a vast array of data sources and types, before stitching them together in a composite with other data elements.

There can be little doubt that the very idea of *big data* is having significant consequences for economy and society, and for human knowledge – whether in the petabytes of scientific data generated by the Large Hadron Collider at CERN, or in the ongoing debates on the social sciences' use of transactional data for the understanding of human behaviour and social transformation.[1]

The widely held view that we are living in a world dominated by the *4 Vs* of big data – increased volume, variety, velocity, veracity – has led to a focus on the significance of the *scale* and *scope* of the digital traces left in the wake of people, things, money and ideas on the move (Boyd & Crawford, 2012). As data-generating devices proliferate, and as data storage and processing power have become more scalable, the rise of twenty-first-century big data has been described as a 'goldmine' of 'magical material', a 'new oil' fuelling innovative forms of economic transaction and circulation whose 'core assets' are data (Kroes, 2013; OECD, 2013).

And yet, what precisely is meant by the concept *big data*? What distinguishes, for example, our contemporary period of vast quantities of digitized data from what Ian Hacking (1982) more specifically observes as the 'avalanche' of statistical numbers of the nineteenth century?[2] Hacking's detailed analyses of the emergence of statistical knowledge of the trends, rates and patterns of populations emphasize the importance not only of the novel availability of data on nineteenth-century population, but also the scientific and calculative techniques that rendered the data available for the 'making up of people' by the state (1986). Hacking is attentive to the intimate relationship between rationalities and technologies of governing, and mechanisms for calculating, intervening and acting upon the world (see also Miller & Rose, 1990).

It is precisely such reflection on the situated calculative techniques and processes used in the gathering, analysis and deployment of big data that we find to be absent from the contemporary ubiquitous use of the term.

As Bruno Latour notes, the 'giant in the story' is not necessarily a larger character, or a figure with greater agency, 'than the dwarf' (1988, p. 30). The spatial scale of the object, we might say with Latour, tells us little about the capacities or agency that it assembles around it. Notwithstanding such problems of ontological scale in the designation of big/small, for our purposes there are two significant aspects to the *big* in big data. First, big data pushes at the limits of traditional relational databases as tables of rows and columns, and requires new ways of querying and leveraging data for analysis, in addition to the structured query language (SQL) built for relational databases in the 1970s. As larger volumes of data of more diverse types become available for analysis, algorithmic tools have developed in parallel – for the bulk processing and analysis of stored 'data at rest', and for so-called 'real-time' analysis of streaming 'data in motion' (Gupta *et al.*, 2012, p. 43). Second, big data is big to the extent that it exceeds and changes human capacities to read and make sense of it. Specifically, the contemporary pluralization of data forms exceeds the linear techniques of punch card indexes, population sampling and probabilistic calculation characteristic of Ian Hacking's variant of large-scale printed number. The forms of sense-making that grew up around the collection of demographic data, such as census, and epidemiological data, such as statistical rates of mortality, and that persisted in early forms of computing, where the computer was a human who calculates, are transformed with new forms of human and machine reading of data (Hayles, 2005, 2012). The study of the form of sense-making that dominates twenty-first-century *big* data analysis must also address, then, the dwarf in the story – the little analytical devices without which the giant of big data would not be perceptible at all.

Perception and the little analytics

Writing on perception, Henri Bergson describes a 'transformative scene from fairyland', in which 'as by a magician's wand' perception is conjured 'so that it may have nothing in common with the matter from which it started' (1912, p. 32). For Bergson, the inescapable problem of the 'insufficiency of our faculties of perception' to capture infinite variation is one shared, albeit differently, by physics and metaphysics, by science and philosophy (1965, p. 132).[3] The 'difficulty of the problem', as described by Bergson, is that 'we imagine perception to be a kind of photographic view of things, taken from a fixed point' by an apparatus, a device or an 'organ of perception' (1912, p. 31). The knowledge of a material world given to us by perception 'works a dividing up of matter that is always too sharply defined, always subordinated to practical needs', whilst 'our science, aspiring to the mathematical form, over-accentuates the spatiality of matter' (Bergson, 1965, p. 211). In short, for Bergson perception is attuned to action, to the carving out of a series of still images, a bringing to attention that allows action to take place. To be clear, it is not the case that data analytics can be considered analogous to Bergson's many and varying accounts of the faculties of perception, but that all forms of perception, human and non-human, physical and metaphysical, are intractably bound up with attention and action, with how an image of something of interest is brought to attention for action (Crary, 1999). The practices of data mining and analysis institute curious forms of perception, drawing as they do so extensively on the spatial methods of the mathematical and physical sciences (Mackenzie, 2011; Parisi, 2013), and yet claiming to produce an enhanced version of perception, to capture a world of flux, a 'totality of images', data streams and becoming.[4]

The question for Bergson's philosophical method, then, is not 'how perception arises', but 'how is it limited', to know 'how and why this image *is chosen* to form part of my perception, while an infinite number of other images remain excluded from it' (1912, p. 34). The task at hand, as Bergson understands it, is to 'give up your magician's wand' and 'follow the process to the end', to understand how a perception that 'should be the image of the whole' becomes limited and 'reduced to the image of that which interests you' (1912, pp. 35–36). The work of data analytics processes appears to institute a specific form of carving out and extraction from a broader extensity of big data – audio files, social networking site text, binary image files, GPS data, RFID tag reader data, digital public records and so on – and the stitching together with other data elements. Like Bergson's fairyland scene, where an image of interest is extracted from a whole, data analytics are *instruments of perception*: they carve out images; reduce heterogeneous objects to a homogeneous space; and stitch together qualitatively different things such that attributes can be rendered quantifiable. If the metaphor of big data is to continue to dominate the governing of digital life, then it cannot be understood without the little analytics that make data perceptible. As Katherine Hayles suggests, data that are 'unimaginable in their totality' and 'too vast to comprehend' are rendered 'more or less tractable' by the algorithms that make them searchable (2012, p. 230). In effect, one of the many problems with a pervasive focus on *big* and *data* is that the finite and granular minutiae of the analytics are overlooked. From the financial subject's online access to mobile and adaptive credit scoring (Marron, 2007), to the business and marketing analysis of personal data generated via mobile devices about their users (OECD, 2013), to the use of algorithmic and automated border controls (Amoore, 2011), it is the work of the little analytics that makes ever more finite interventions in the governing of life itself (Rose, 2006). In the discussion that follows we focus on how the work of analytics and algorithms not only transforms the meaning and value of data, but also inscribes the very perception of the world in which we live, govern and are governed. Animated by Bergson's injunction to 'give up the magician's wand' and 'follow the process', we discuss three processes through which little analytics reshape the landscape of what can be perceived, known and acted upon: ingestion, partitioning and memory.

Ingestion: *n*=all

Reporting on the rise of 'analytics-based decision-making', the consultants Accenture urge their business clients to 'move beyond traditional sources of data' and 'seize the opportunities for new insights' created by new sources, such as 'text analytics from social media and digital interactions' (Accenture, 2013, p. 5). What is captured here is a double transformation in the landscape of big data: a radical expansion in the forms of social interaction and transaction that can be *rendered as data*, or what Victor Mayer-Schönberger and Kenneth Cukier (2013) call 'datafication', coupled with a novel capacity to *analyse across* a variety of types of data. In short, the rise of big data witnesses a transformation in *what* can be collected or sampled as data, and *how* it can be rendered analysable. In the vocabulary of the computer scientists and data analysts, data are no longer strictly *collected*, but rather *are ingested*, such that everything becomes available to analysis, the sample being represented as infinite, or *n=all*.[5]

In the past, conventional forms of *structured data*, characterized by 'numbers, tables, rows, and columns' (Inmon & Nasaevich, 2007, p. 1), were the only forms of data to inhabit the world of databases, spreadsheets and statistical tables, and thus were the only data that could be leveraged for analysis. In many ways, the distinction between structured and unstructured data that dominates data science discourse and social science accounts is profoundly misleading. Of course, we might say that all data declared to be unstructured is always already structured, and certainly remains structured in important ways within data architectures and digital devices (Berry, 2014; Kitchin, 2014). Yet, while structured data is territorially indexable, in the sense that it can be queried on the horizontal and vertical axes of spreadsheets within databases, so-called *unstructured data* demands new forms of indexing that allow for analysis to be deterritorialized (conducted across jurisdictions, or via distributed or cloud computing, for example) and to be conducted across diverse data forms – images, video, text in chat rooms, audio files and so on.[6] In the main this has implied making unstructured data analysable by the establishment of links with already indexed structured data and the creation of new indexes.

So, for example, IBM's 'predictive policing' software uses content analytics that promise to: 'search and analyse across multiple information sources, extracting key pieces of information like new addresses, credit cards or passports that can help resolve identities, build relationship networks and trace patterns of behaviour' (IBM, 2012, p. 2). The linking of the data elements is performed through *joins* across data from different data sets, either on the basis of direct intersections with already indexed data (e.g. via a phone, credit card or social security number ingested from a database), or probabilistically, through correlations among data-points from different sources (e.g. text *scraped* from a Twitter account correlated with facial biometrically tagged images drawn from Facebook). Though in many ways the use of the join is not novel and is commonly used for querying relational databases, today the analysis operates with a much more diverse pool of data. The allure of unstructured data is that it is thought to contain patterns heretofore unseen and, therefore, a wealth of previously hidden insight. The growing use of analytics capable of reading and making sense of data, of unlocking its potential, is tightly interwoven with a 'world of promise and opportunity' thought to be buried in a text in need of an index (Inmon & Nasaevich, 2007). We can sense here 'a desire for wholeness, an embrace of the total and comprehensive' that ceaselessly 'generates a politics of mash-ups, compilation and assemblage' (Ruppert *et al.*, 2013, p. 38).

Following the process of ingestion: text and sentiment analysis

Consider a global pharmaceutical company conducting *web listening* for sentiment analysis of open-source Twitter and online epilepsy support social network data.[7] Of the 60,000 epilepsy sufferers in the United Kingdom, the majority take one of two major commercial drugs, one of which is produced by a global pharmaceutical company, we will call them Alpharm, who wish to understand the *churn* in patients moving to the alternative drug. As a problem addressed by algorithm, this kind of churn is not different ontologically from what early adopters of sentiment analysis, such as credit card companies or mobile phone companies, were analysing with respect to customers switching to new providers, for in essence they all wish to understand what kinds of sentiments or tendencies signal specific human behaviours. In the case of Alpharm, text analytics based on *R* programming language were run in *open-source* data to build what was described as a 'whole picture' of the relationships between people, their medication, their family life, their moods and feelings, their perceptions of disease and the things they do to cope with side-effects of the epilepsy medication. The text analytics used parsing algorithms to split the sentences into *tokens*, which could be words, numbers or symbols, and stemming algorithms to reduce the words to their base or root (so, for example, the words *thinks*, *thinkers* and *thinking* were stemmed to a common root *think*). A network analysis was then conducted to reveal significant nodes and links such as, for example, the association between a drug brand-name and a particular side-effect, or between an affective quality such as fear or panic and a significant life event. Questions that were asked of the data via the analytics included: 'how do people live with their disease?'; 'what are the notable patterns in the therapeutic use of music, sport or alcohol?'; 'who are the key opinion leaders in this social network?' Significantly, these were not thought to be queries that Alpharm could make of the more conventional, linear and structured data otherwise available via patient or medical practitioner surveys. The subtleties of affects such as anger, rage, depression, melancholy or anxiety were thought to be retrievable via the social media data in a way that accesses and reads life in its very emergence, in the unfolding of life.

What one can see in the use of text analytics and sentiment analysis is not merely a world of more freely available big data, or *n=all*, but more specifically a distinct mode of gathering and reading that data. It is the gathering and reading that form part of the work of the little analytics. At first glance, text analytics do not appear dissimilar from reading as such, and, indeed, the genesis of text mining has its roots in natural language processing and semantic structure. However, as Katherine Hayles has argued persuasively, machine reading is a specific kind of reading that not only allows algorithms to read text, but also alters irrevocably the way humans read and, consequently, the way humans think and perceive (2012, pp. 28–29). What matters is thus not strictly whether machines may somehow read *like humans*, but rather how the possibilities of digital forms, such as text analytics, change the practice of reading for humans and machines alike.[8] The 'hyper reading' that Hayles identifies among multiple forms of human and machinic reading consists of 'skimming, scanning, fragmenting, and juxtaposing texts', being a mode of reading attuned to 'an information intensive environment' (2012, p. 12). The reading involved in text analytics, engaged on the part of the algorithms and the humans who action a query, is just such 'hyper reading' of multiple forms and sources of data as though they were a single text.

As in our Alpharm example, in order for the particular object of interest to be perceptible, a certain damage is done to words and syntax, and to context. Consider the processes necessary for text analytics to read: the removal of *stop words*, including *and*, prepositions, gender suffixes in some languages, and the definite and indefinite articles *the* and *a*; *stemming*, whereby words are reduced to their stems; and the removal of punctuation marks and case sensitivity. In effect, as one sees in the pharmaceutical company's scraping of the web for a *complete life story* of a person, in order for a life to be read with data analytics, any trace of a context, movement or a story that has a recognizable narrative must first be pruned out. As Hayles points out, there remain important differences between narrative-based stories of literature and data-based *story-telling*:

> The indeterminacy that databases find difficult to tolerate marks another way in which narrative differs from database. Narratives gesture toward the inexplicable, the unspeakable, the ineffable, whereas databases rely on enumeration, requiring explicit articulation of attributes and data values. (Hayles, 2012, p. 179)

The parsing and stemming of text, then, is intrinsic and necessary to the capacity of analytics to *read* at all. The stories about the lives of epilepsy sufferers, or the purchases of retail loyalty-card holders, or the transactions of online banking customers that can be read by algorithms are not the indeterminate narratives of life stories. They are lives that are flattened and reduced to their common stems, connected with others only through correlations, links and associations. On the basis of these analytics-derived life stories, decisions are made about people, policies are implemented, resources are allocated and interventions are targeted.

Because text analytics and sentiment analysis conduct their reading by a process of reduction to bases and stems, their work exposes something of the fiction of a clear distinction between structured and unstructured data. Through processes of parsing and stemming, everything can be recognized and read as though it were structured text. The significance, then, lies not merely with some newly abundant unstructured data stream, but with the process of ingestion itself. From the Latin *in-gerere*, to carry into, to ingest suggests the drawing in of quantities of matter into an engine or body, such that the contents can be filtered, some of them absorbed and others expelled or discarded. For instance, in his first interview with the film-maker Laura Poitras, Edward Snowden refers to the 'ingestion by default' of 'bulk' communications data by the National Security Agency (NSA), wherein the software 'absorbs' that which has value or interest.[9] When one hears government statements that the 'haystack' is required in order to target the 'needle', or that analytics are akin to 'using a magnet to draw the needle out of the haystack instead of combing through the straw yourself', it is the process of ingestion that is key (Intelligence and Security Committee, 2014, 2015).

The analytics solutions that inhabit this world of ingestion – such as TIBCO® Spotfire® 5.5, capable of analysing 30 different types of data simultaneously, or IBM® Content Analytics, working with 30 sources and 150 formats – can read the data only because they are indifferent to the qualitative differences that dwell within heterogeneous data. What is most significant about the process of ingestion for us is not merely the volume of the data that can be drawn into an analytics engine, but how an object or person of interest emerges *via ingestion*, how the target is identified from the mass. Returning to Bergson, who writes about a form of ingestion, albeit one where plants and animals absorb nutrients in ways that 'care little for individual differences' (1912, p. 206):

> Hydrochloric acid always acts in the same way upon carbonate of lime […] Now there is no essential difference between the process by which this acid picks out from the salt its base, and the act of the plant which invariably extracts from the most diverse soils those elements that serve to nourish it. In short, we can follow from the mineral to the plant, from the plant to the simplest conscious beings, from the animal to the man, the progress of the operation by which things and beings seize from out of their surroundings that which attracts them, that which interests them practically […] simply because the rest of their surroundings takes no hold upon them. (Bergson, 1912, pp. 207–208)

The analytics promise to leverage all types of data stored across multiple architectures in order to *unveil* things that could not otherwise be seen, the previously unseen, hidden patterns that dwell in the folds and joins between data forms. Yet, if we understand the work of the analytics in 'seizing from the surroundings' that which interests or sustains, then we begin to see how qualitative differences between data forms become obscured by the pursuit of the object of interest.[10] The analytics extract from diverse elements that which is of interest, indifferent to the heterogeneity that surges beneath that data. Viewed in this way, the contemporary big data question of how to approach *n=all* is posed rather differently. In contrast with a word of big data that seeks out 'complete data sets never available before' (interview 1 October 2013) and where 'big data wants n, nothing else' (Hildebrandt, 2013, p. 6), *n=all* appears instead as an impossible claim. The process of ingestion draws in the data rather as Bergson's hydrochloric acid acts upon chalk, or a plant acts on diverse nutrients in the soil, that is to say indifferent to the *all* with which it communes. In this specific sense *n* will never be equal to all. In the so-called *flat files* of analytics algorithms, which quite literally flatten the multiple distinctions among data forms in order to make the data readable and analysable, the complex temporalities of the life that generated the data are entirely lost.

Partitioning: 'transform, select and filter the variables'

As IBM describe their Intelligent Miner® software, the task of analytics algorithms is 'to extract facts, entities, concepts and objects from vast repositories' (2012, p. 2). Understood thus, the work of the analytics can be conceived as one specific from of sense-making – one means by which subjects and objects of interest are partitioned from a remainder and singled out for attention. How are qualitatively different entities in a heterogeneous body of data transformed into something quantitative, something that can be enumerated? In his early work Henri Bergson differentiates between two ideas of time, the time of lived experience, or *durée réelle*, and the mechanistic time of science in which time is a succession of images or spatial frames, as in film (Ansell Pearson & Mullarkey, 2002, p. 17). In this spatial representation of time as a 'series of halts', we begin from a fixed point 'in the immobile to watch for the moving reality as it passes instead of putting ourselves back into the moving reality to traverse with it' (Bergson, 1965; see also Connolly, 2011). Understood thus, the fixed instrument of perception partitions, according to what is of interest to it, a series of immobile stills from which to derive some picture of a changing world.

As Gilles Deleuze notes, Bergson 'calls into question the order of needs, of action, and of society that predisposes us to retain only what interests us in things' and that 'tends to obscure differences in kind' (Deleuze, 1991, p. 33). Following Bergson, Deleuze understands the qualitative multiplicity of 'duration' to bear all of 'the differences in kind', while 'space' is unable to 'present anything but differences of degree (since it is quantitative homogeneity)' (1991, p. 31). The patterns of life, so readily claimed as the world captured by analytics, might be properly thought of as durational, multiple, continuous and qualitative. Like the modern physics Bergson and Deleuze describe, the analytics extract and detach data from the whole, drawing a series of discontinuous spatial images as vantage points on a mobile world. While analytics claim to afford a vantage point on emergent life patterns and tendencies, in practice they spatialize time and substitute differences in kind for differences in degree, collapsing qualitative difference into enumeration and action.

Following the process of partitioning: MapReduce

Let us consider more closely the form of partitioning at work in one widely deployed programming model that uses distributed or parallel computing, MapReduce. Originally designed by Google in order to transform the indexing of Google webpages, MapReduce is a framework for parallel processing across vast data sets (Dean & Ghemawat, 2004). In the context of data proliferating in different forms and across different databases and servers, the 'challenge', as understood by data scientists, is said to be the 'feasibility of reasoning over such large volumes of data' (Tachmazidis *et al.*, 2012). Commonly used in open-source software such as Apache Hadoop, the MapReduce architecture makes it feasible to analyse data precisely via its distributed form, by dividing computation into two distinct phases.

The first *map* step 'breaks down the data into manageable pieces', subdividing the aggregate problem into multiple discrete elements and 'automatically spreads them to different servers' (Ohlhorst, 2013, p. 8). The input files are *sharded*, that is to say they are not divided according to the existing structure of the files, but arbitrarily – so, for example, text files are split according to byte boundaries (Tachmazidis *et al.*, 2012). The second *reduce* step draws on input from the scattered map nodes, and joins the map results back into a final 'master calculation' (Ohlhorst, 2013, p. 8). Thus, for example, in the application of MapReduce to human genome analysis, the first map step would conduct analysis at the level of the genome, such as genotyping, with the output fed to the reduce step, where calculations are made across the aggregate data on multiple points of the genome (McKenna, 2010). In domains such as human genetics, meteorological data, security intelligence and business intelligence, MapReduce is thought to supply 'processing of a vast amount of data in parallel on large clusters of machines in a fault-tolerant manner' (Gupta *et al.*, 2012, p. 49).

In simple terms, MapReduce is significant because it changes the nature of what can be analysed across multiple data formats and databases, across a distributed data landscape:

> Every call, tweet, e-mail, download, or purchase generates valuable data. Companies and governments are increasingly relying on Hadoop to unlock the hidden value of this rapidly expanding data [...] Sensor output, videos, log files, location data, genomics, behavioural data are just a few of the data sources driving Hadoop use. (MapR for Apache Hadoop®, 2011)

MapReduce is thought to unlock hidden value because it makes it possible to analyse exponentially increasing volumes of diverse data for patterns and clusters that are not necessarily determined in advance. The partitioning and analysis of data using Hadoop software deploys algorithms in a process described as *knowledge discovery*. In contrast to the deductive production of knowledge from *a priori* queries or hypotheses, in this case the data analytics use inductive and abductive steps to identify previously unknown patterns in a large volume of data (Dunham, 2002). The significance here is that knowledge discovery increasingly does not begin with a set of search queries against which the data will be run. Instead, the process inductively generates queries, such that the analytics are said to 'let the data speak' (Ohlhorst, 2013).

How do analytics techniques like Hadoop MapReduce transform the nature of what can be rendered perceptible and analysed? In contrast with the more strictly statistical structured query language (SQL), designed in the 1970s for use with relational databases, advanced analytics work with the uncertainties of possible links and connections. As the author of the world's most highly cited computer science paper on data mining software, Rakesh Agrawal, explains, past forms of data analytics 'used a statistical notion of what was interesting', such that the 'prevailing mode of decision making was that somebody would make a hypothesis, test if it was correct, and repeat the process' (Agrawal & Winslett, 2005, p. 5). With the advent of large databases, distributed computing and extensive unstructured data sources, however, 'the decision making process changed', and a series of algorithms would 'generate all rules, and then debate which of them was valuable' (2005, p. 8). While in statistical forms of large-volume data the object of interest emerges from the testing of probabilistic assumptions or queries, with the output being a subset of a database, in knowledge discovery the matter of interest is iterative and emergent, with the output consisting of previously unknown patterns and relationships.

In this way, the work of the analytics is to discover, aggregate and interpret rules for items within a subset of data. It is *the rules generated by the rules* within the analytics that will determine what is of possible value. Let us imagine, for the purposes of illustration, a rudimentary knowledge discovery process in which it is inferred that transactions in a database which contain item *x* also contain item *y*. So, it may be that 24 per cent of customers who purchase a novel *x* from an online retailer will also purchase music download *y*, and that 5 per cent of the total transactions in a given period contain both purchases. In this example, 24 per cent expresses the *confidence score* of the rule *xy*, and 5 per cent expresses the *support* for the rule. The problem of knowledge discovery across a large volume of transactions, then, is to run the analytics in order to identify all rules that satisfy some predetermined level of support and confidence (Zhang & Wu, 2001). What matters is thus not the intrinsic value or content of *x* or *y*, but how these data items can be associated together and what can be known about their relations with a wider set of bulk data (Agrawal *et al.*, 1993). Importantly, the use of association rules for large transaction data sets is not unproblematic, as their analysis can reveal 'hundreds of thousands of rules at reasonable levels of support or confidence', with many of them being 'redundant or obvious' and therefore 'not interesting' (Klemettinen *et al.*, 1994).

A process of 'pruning' is therefore required in order to reduce the things of interest to the 'strongest relationships in the data' (Raeder & Chawla, 2011, p. 100).

If the advent of what has come to be known as *big data analytics* is changing the nature of what questions or queries can be asked, or of what can be calculated, or the nature of analysis itself, then what is the significance of this? Does it matter to political and social life, to how we govern and are governed? The interception and analysis of terabytes of unstructured data for security purposes has attracted a great deal of political attention in the wake of Edward Snowden's revelations about the NSA's and GCHQ's PRISM and TEMPORA programmes (Greenwald, 2014; Harding, 2014).[11] The locus of the debate, though, has resided primarily with the question of *mass surveillance* and the collection and storage of personal communications and transactions data in bulk (LIBE, 2013). There has been scant attention paid to the processes of data partitioning and reassembly, and what these might mean for the relationship between a mass or bulk volume of data and an object or person of interest. Somewhat hidden and unremarked upon in one of the leaked PRISM slides on 'collection and dataflow' is the 'scissors' process, which 'sorts data types' (http://www.washingtonpost.com/wp-srv/special/politics/prism-collection-documents/m/). Though we can observe very little of the classified processes at work in the partitioning and sorting of the data in PRISM, there are materials in the public domain that make it possible to understand the processes of cutting and stitching involved in analytics processes such as *scissors*.

In June 2013, the US Government Accountability Office (GAO) published a decision on the technology corporation IBM's challenge to the CIA's award of a contract for data analytics services to Amazon Web Services. IBM had challenged the award of the contract to Amazon on the grounds that the CIA's evaluation team did not fairly evaluate the technical and financial aspects of the bids. At issue was the two companies' 'materially different interpretations of the scenario requirements' (GAO, 2013, p. 4). While Amazon priced their bid on the basis of continual 24/7 analysis of a volume of 100 terabytes of data, IBM had envisaged analysis of a series of 'batches of 100 terabytes' of data. In preparing their bids, the software design teams were required to address, and to price, the CIA's scenario for data analysis:

> This scenario centers around providing and hosting an environment for applications which process vast amounts of information *in parallel on large clusters (1000s of nodes) in a fault tolerant manner using MapReduce*. The solution to this scenario should automatically compute for the segmentation and parallel processing of datasets via the MapReduce framework [...] Assume a cluster large enough to process 100 TB of raw input data. Assume 6 reads/second and 2 writes/second. Assume 100% duty cycle on all machines. (GAO, 2013, p. 5, emphasis added)

When IBM queried the temporality of the 100 terabytes and 100 per cent duty cycle, they specifically asked how many data analysts would make simultaneous queries of the data in the scenario. The CIA responded to the query by appealing to the existing practices of data analytics in the commercial sphere, inviting the bidders to bring to the state the techniques already thought to be best practice in economy and commerce: 'The contractor should propose commercial best practices derived from their commercially available solutions to provide data analytics via the MapReduce software framework to concurrent users from multiple organizations' (GAO, 2013, p. 7). Here the divergent responses of IBM and Amazon to the scenario reveal rather more than two competing interpretations of the requirements. They afford a glimpse of how the data analytics in processes such as *scissors* sort large volumes of data, and the proximity of security applications such as PRISM and TEMPORA to the commercial data analytics used every day to tell us which book we might like to buy next. Amazon's established commercial practice of analysing *clickstream* data on its customers in close to real time and on a continuous cycle, it seems, better met the CIA's requirement for analytics to deal with large volumes of unstructured Internet data to be queried by multiple concurrent users, from border and immigration control to counter-terrorism officers.

The capacity to integrate data analytics across multiple analysts, and to map and reduce across multiple nodes, exhibited here by Amazon, contrasts with IBM's extraction of *batches* of data for analysis. One response to the CIA's scenario appears to sustain a somewhat conventional social science approach to sampling, and a particular relation between the subset and the *whole* of big data. In the other response one can see the ceaseless stream of ingestion, partitioning and reassembly that affords novel iterative approaches to *sample* and *whole*, where, in effect, people and objects continually cross back and forth across the sample and the whole. The distributed analysis of data streams, as David Berry writes, sustains 'some form of relationship with the flow of data that doesn't halt the flow, but rather allows the user to step into and out of a number of different streams in an intuitive way' (2011, p. 143). In Amazon's MapReduce framework for the CIA, it is the identification of patterns of note across different data streams that gives rise to a threshold at which a target or person of interest is identified.

The mobile thresholds of support and confidence for an association rule – the very key to setting the gauge for the analytics – have become highly significant political boundaries for our times. The threshold is the moment when the *strongest relationships* are identified, the moment when someone or something of interest becomes perceptible. The moving of the threshold changes who or what is surfaced from the data and brought to attention. In the historical origins of data analytics this threshold was commonly defined in terms of a *frequent set*, where the co-occurrence of retail consumer items in patterns of purchases met a predetermined level of support and confidence. Co-occurrence in itself is not always a matter of interest, for example, milk co-occurring with bread in basket data would have high levels of support and confidence, but would not constitute an object of interest. Where similar MapReduce processes are used to set the threshold of risk for border controls, or the threshold for a 'nexus to terrorism' (de Goede, 2012), the threshold of support and confidence becomes a border in itself, where the co-occurrence of particular data elements will give rise to a person or object of interest. Understood thus, the little analytics are instrumental in what is called *target discovery*, the defining of a political threshold of perceptibility where a person of interest comes into view.

The partitioning and assembly processes not only structure something of the threshold of political visibility, but also redefine the lines of sovereign authority. In the design of software for sovereign information sharing (SIS) for the US Transportation Security Administration (TSA), for example, the analytics are said to enable 'computation across autonomous data sources in such a way that no information other than the intersection results is revealed' (Agrawal *et al.*, 2005, para 2.1). Such insights into the work of the analytics are critical to the contemporary form of governing life. Though the classified nature of programmes such as PRISM and TEMPORA makes it impossible definitively to identify whether sovereign information sharing is among the techniques used, the computer scientists reveal clearly how their analytics make it possible to share the subset data on persons who cross a threshold *of interest*, while annexing the big data sample from which it was drawn. In the TSA example available in the public domain, the airlines encrypt their PNR data, and the security authorities encrypt their watch-lists, with the analytics running the 'intersection results' for patterns, associations and matches. 'The TSA agrees that the use of the intersection results will be limited to the purpose of identifying suspects', write the computer scientists, 'but it will store all the metadata' (Agrawal, *et al.*, 2005, para 6.5). In the TSA's sovereign information sharing system, the analytics become part of the condition of possibility for sovereign power. Not only does SIS appear to make possible sovereign decisions on who or what poses a risk to US transportation security, but it also establishes the threshold at which data on persons of interest are pulled to the perceptible surface in the *map* process and extradited to the *reduce* step, where a calculation is made about them.

The partitioning work of contemporary analytics such as MapReduce, then, is political in the sense that it defines the threshold of perceptibility. This is a threshold where, as Deleuze reminds us with Bergson, 'differences in kind can no longer appear', and where science 'no longer presents anything but differences of degree, of position, of dimension, of proportion' (1991, p. 34). Analytics are technologies of degree *par excellence*. They subdivide a heterogeneous data set such that no original whole could ever meaningfully be reassembled. The partitioning has significant implications for any critical response to the pervasive use of programmes, such as Hadoop® – at our borders, in our consumer databases, inside the risk calculations of banks and so on. For the work of the analytics is not at all challenged by the demand that the *context* of data be respected in its analysis (Nissenbaum, 2010), nor that one has the right *to be forgotten* or deleted in a digital age (Mayer-Schönberger, 2009).

The processes of partitioning and analysis precisely do not require a context, nor do they need individuals who can be remembered. The critical demand for a contextual limit to the analysis of life data, or a deletion of the digital subject, gains little purchase in a world where attributes are extracted from their qualities and afforded numeric values. The analytics that partition big data make it possible to forget the person and the context, but to remember the position, the distance or proximity of association.

Memory: near real-time analytics

Presenting their analytics for stream-based event processing, analytics providers TIBCO® introduce their Spotfire® solution:

> TIBCO helps you understand the past so that you can better anticipate the future. We call it the two second advantage … What good is it to know that you've lost your customer after the customer has left your premises? What good is it to know that there is a power outage after a city is in darkness? Or that fraud has occurred after the money has left the bank? This is the power of the two second advantage. This is the importance of being able to both understand the past and anticipate the future.[12]

TIBCO's® development of analytics that promise to identify emergent trends and to enable anticipatory action – to 'uncover opportunities nobody else can see' – signals something of the temporality of contemporary analytics. TIBCO® Spotfire® promises to 'turn data into actionable insights with dashboards, apps and analytics', so that data on unfolding events can be used to enable fast and strategic 'near-real-time' decisions. The software for such stream-based analysis identifies links between events coming from multiple data sources – for example, Twitter trends, smartphone transactional data, Facebook likes – stitching together the data signals to anticipate near futures, what they call the 'two second advantage'. Such methods are becoming ubiquitous in the commercial world, for identifying people with a propensity to *churn* and transfer their custom to a new company, and in the security domain, where *radicalization* or *attack planning* is thought to be identifiable at the intersection of multiple data events. Spotfire® also signals a broader move to simplify the interface between the analytics and the user, such that she 'doesn't need to understand the R-code running in the background'; the 'software automatically chooses the most appropriate forecasting algorithm'; and the analyst can seek help from 'what does it mean pop-ups', without having to understand how the forecast was created.[13] The relationship between past data, decisions made in the present and actions taken on a future that is seconds away, then, is significantly reconfigured by advanced analytics. What form of machinic memory do the analytics access in order to anticipate the future? How do simplified *drag and drop* interfaces, which hide the complexity from the analyst, change the orientation to decision and action?

The claim of the analytics' capacities to analyse data in near real time, or with a 'two second advantage', is based upon data stream processing. In effect, the stream of unstructured data is conceived as a continuous flow whose speed exceeds conventional perceptions of an event. For example, when a 5.8 magnitude earthquake struck the US state of Virginia in August 2011, it was suggested that the first Twitter messages reached New York in advance of the first measurable shock waves (Emerson *et al.*, 2012). Similarly, on the night of the raid on Osama Bin Laden's compound in Abbottabad in 2011, IT consultant Sohabib Athor began a string of Twitter messages with 'Helicopter hovering over Abbottabad at 1am, is a rare event'; 82.68 million retweets and 21 hours later, Athor tweeted 'Uh oh, now I'm the guy who live blogged the Osama raid without knowing it' (Emerson *et al.*, 2012). The Topsy Labs analytics identified the *virality* of the message exposure within the social web, demonstrating how open-source stream data can be analysed to reveal otherwise unknown events in their unfolding.

Advanced event stream analytics, such as those deployed in high-frequency financial trading (MacKenzie, 2011), and in applications like TIBCO® Spotfire®, imply a significant transformation of the relationship between past, present and future, a transformation that is not fully captured by the idea of a *real time*. In his commentaries on the relation between past and present, Henri Bergson signals the specific temporalities of memory. 'You define the present in an arbitrary manner as *that which is*', he writes, 'whereas the present is simply *what is being made*', and 'nothing *is* less than the present moment, if you understand by that the indivisible limit which divides the past from the future' (1912, p. 193). If, as Bergson understands it, perception of the present is more precisely located in the immediate past – 'practically we perceive only the past' – then could it be that any claim to a *real-time* present can only hope to engage 'the invisible progress of the past gnawing into the future' (1912, p. 194)? Or, does the machinic memory of advanced analytics find novel ways to contain and access images of the past, such that action in the present is a possibility? 'But, if the brain cannot serve such a purpose', asks Bergson, 'in what warehouse shall we store the accumulated images?' (1912, p. 192). Can we conceive of analytics such as TIBCO® Spotfire® as supplying an infinite data warehouse of images of the past, which may be retrieved and replayed at any time?

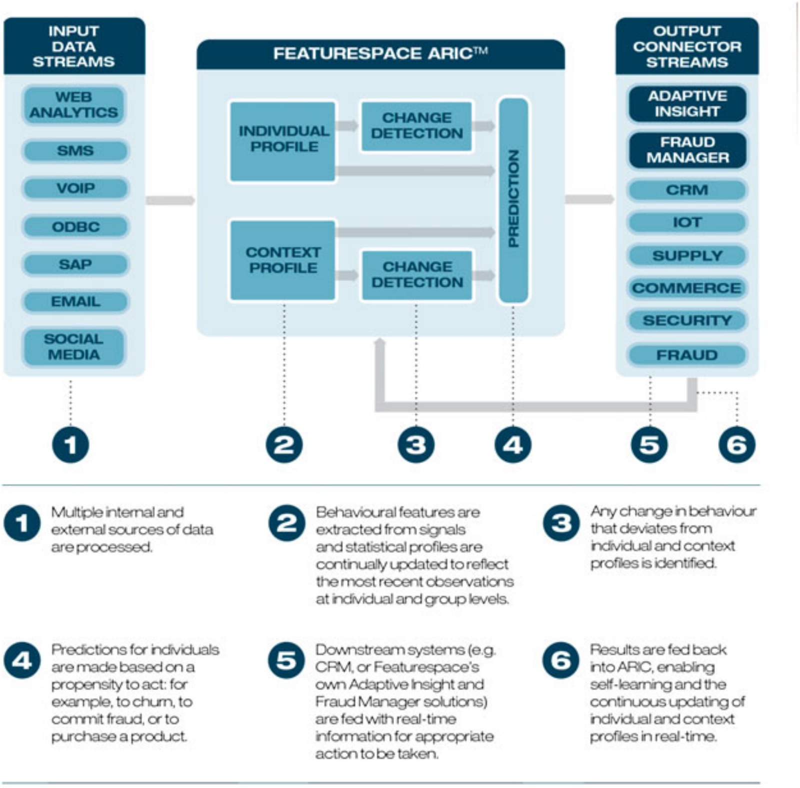
Of course, in many ways the accessing of memory depicted in Bergson's account of conscious perception is a process that is impossible for algorithm to replicate. In spite of claims to transcend the limits of human memory and expand capacity for action, big data analytics have limited capacity to incorporate the durational time of consciousness and experience. What matters in the algorithmic recalling of a series of past events as data is not strictly their temporal character, but their spatial distances data-point to data-point, their links, associations and correlations, one to another. As Katherine Hayles has expressed the different temporalities at work, the distinction 'between measured time and time as temporal process can be envisioned as the difference between exterior spatialization and interior experience' (2012, p. 112). The system of memory appropriate to describe *real-time analytics*, then, is one of measured time in which the temporalities or durations that dwell within the life that yields the data-points are entirely lost. Gilles Deleuze depicts Bergson's duration as the time in which the present endures: 'the "present" that endures divides at each instant into two directions, one oriented and dilated toward the past, the other contracted, contracting toward the future' (1991, p. 52). It is only in duration, one might say with Bergson and Deleuze, that the capacity to perceive past, present and future as qualitatively different things is possible. By contrast, the data analytics that promise a two second advantage can conceive only of a spatial point sliced through time, a point where action can be taken.

Following the process of memory: Featurespace

What kind of process of memory is at work in so-called real-time analytics? The past of the analytics is a preselected collection of actual pasts and, when it is used, only discrete elements are recalled. Let us illustrate this process at work in the Adaptive Real-Time Individual Change Identification (ARIC) engine produced by Featurespace. Described as 'the only adaptive behavioural analytics software in existence', ARIC is used in the detection of casino and credit card fraud, customer churn, marketing opportunities and security threats 'in over 60 countries', and in 'processing over 20 million transactions per day' (Featurespace, 2013). The ARIC engine analyses thousands of disparate data sources, including SMS data from mobile phones and e-mail metadata, identifying 'signals', or small but interesting changes in the patterns of data. The signals are described as 'symptoms' of human behaviour that can be converted, via processes of parsing, stemming and partitioning, into 'features'. The features are observed in their multiple relations to other people, things and groups with correlating features, such that 'predictions for individuals are made based on a propensity to act: for example, to churn, to commit fraud, or to purchase a product' (Featurespace, 2013). Changes in behaviour that 'deviate from individual and context profiles' are identified, and the results are 'fed back into ARIC' in a process of 'self-learning' and 'continuous updating of profiles in real time' (Figure 1). In effect, the analytics in ARIC combine conventional Bayesian conditional probabilities with a capacity for machine learning on the basis of small modifications in observed behaviours.

The form of memory at work in Featurespace's analytics engine is an iterative movement back and forth across past 'features', present recursive recalibrations of the rules of the analytics (the machine learning) and future projections of human propensities to act. This process observes no qualitative distinctions, not only in the vast and heterogeneous array of data that is ingested from different architectures, but also in the transformations that take place in the present as threshold between past and future. As one analyst described the process to us in interview, ‘we just keep iterating until the results are satisfying’. The processes of memory and iteration in Featurespace, or TIBCO® Spotfire® or indeed in the analytics used by GCHQ and the NSA to build a ‘pattern of life’ (*The New York Times*, 2013) are far removed from what Bergson termed ‘attention to life’ (1912, p. 63). Where ‘attention to life’ bears witness to the ‘adaptation of the past to the present, the utilization of the past in terms of the present’ (Deleuze, 1991, p. 70), the features or patterns of life sought by major supermarket chains and national security agencies know no limitations, no indeterminacies, nothing that is not available to action.

Where the durational time of consciousness confronts the indeterminate future by shedding some light gathered from selected past states, combining with present states, it does so in the knowledge that ‘the rest remains in the dark’ (Bergson, 1912, p. 194). Amid their claims to predict human propensities, by contrast, analytics engines, such as ARIC and Spotfire®, confront an indeterminate future in order precisely to leave nothing in the dark and nothing undetermined. Though the analytics share with consciousness the selection



1 Multiple internal and external sources of data are processed.

2 Behavioural features are extracted from signals and statistical profiles are continualy updated to reflect the most recent observations at individual and group levels.

3 Any change in behaviour that deviates from individual and context profiles is identified.

4 Predictions for individuals are made based on a propensity to act: for example, to churn, to commit fraud, or to purchase a product.

5 Downstream systems (e.g. CRM, or Featurespace's own Adaptive Insight and Fraud Manager solutions) are fed with real-time information for appropriate action to be taken.

6 Results are fed back into ARIC, enabling self-learning and the continuous updating of individual and context profiles in real-time.

Figure 1 Featurespace’s ARIC engine.
Source: http://www.featurespace.co.uk/technology/aric.

of some discrete past events, there the commonality ends. For the analytics take the light of some past states and project it forward as though there could be no dark corners remaining – all propensities will be known, all future acts anticipated. When analytics like ARIC are being used to monitor social media in the Arab Spring uprisings, or to monitor Twitter in the predictive policing of urban protest, it is of great significance that all memory of every past infraction is thought to be retrievable, all futures foreseeable. Indeed, the *event* in event stream analysis is annulled as such, along with the *real* in real-time analytics. For nothing new or eventful can emerge, such is the machine time of the iterative replay of the past state, modified for recent deviations and gnawing into the uncertain future.

Conclusions: the loss of a fallible world

Asked to explain the distinction between the advanced analytics deployed in the commercial spheres of retail and banking and those used for state-oriented security purposes, an analyst reflects on the problem. 'We license software for companies to detect fraud and we license it to governments for counter-terrorism', he says, 'but we don't know what they do with it because that would be classified'. 'I assume', he concludes, 'they are using the most advanced analytics' (interview, 1 October 2013). What is perhaps most notable here is that, at least from the standpoint of the software designers and analysts, the overriding process is one of ingestion, partitioning, retrieving and analysing large volumes of data in order to identify people, objects or patterns of interest. In terms of computer science and mathematics, it scarcely matters what those data are, or from whence they came – they could be financial transactions, supermarket purchases, or national security watch-lists – what matters is the capacity to identify something of interest amid the mass, and the capacity to action that analysis.

It is precisely because of the growing ubiquity of big data analytics across diverse spheres of life that social science should pause in the rush to *exploit the value of big data* and attend carefully to what form of analysis is engaged by advanced analytics. The apparent *big* in big data departs significantly from the 'avalanche' Ian Hacking (1990) observes in nineteenth-century social statistics. Hacking describes in detail how people of interest emerge from the volume of structured data on economy and society in the nineteenth century. From the average man or *l'homme typique* of Adolphe Quetelet to the biometric composites of criminality of Francis Galton, the statisticians identified their object of interest from probabilistic calculation (see also Daston, 1995). In short, they began their inquiries with a statistical notion of what was interesting – this crime rate, or these rates of infant mortality, or this probability of suicide.

The twenty-first-century rise of big data marks a significant break with conventional statistical notions of what is *of interest*. The vast expansion of unstructured digital data, much of it considered to be open source, has been closely intertwined with the development of advanced analytical algorithms to make some sense of that data. And so, amid the cacophony of noise around the *big* in big data, we urge careful attentiveness to the work of the *little* analytics. Rather as the growth of statistical probabilistic methods made data on murder, health, employment or war perceptible and amenable to analysis, so contemporary analytics are instruments of perception without which the extensity of big data would not be perceptible at all. As art historian Jonathan Crary explains in his compelling account of historical changes in the nature of perception, 'perception transformed alongside new technological forms of spectacle, display, projection' (1999, p. 2). As instruments of perception, the technologies of analytics not only focus human attentiveness on particular persons or things of interest, but they also annul and discard as redundant much of the material context from which these persons and things emerged. In this sense, the little analytics are one element of a 'contemporary experience that requires that we effectively cancel out or exclude from our consciousness much of our immediate environment' (Crary, 1999, p. 16). And yet, they are not merely the twenty-first-century manifestation of Benjamin's 'mechanical reproduction' or Lorraine Daston and Peter Galison's (2007) 'mechanical objectivity'. As we have proposed, the invention of analytics engines has transformed the nature of analysis and, with it, the nature of what and how life can be rendered governable.

By way of conclusion, we draw out three implications of a focus on the work of analytics, as they transform the governing of economic, social and political life. First, the advent of advanced analytics ushers in a specific and novel *epistemology* of population. At first glance, the formulation *n=all* appears to render the whole of population as the sample – all data on all of life's transactions are, at least in theory, available to analysis. But the population as the sample in contemporary analytics processes does not imagine the population as a 'curve of normality' or a Gaussian bell curve of plotted attributes (Foucault, 2007, p. 63). Once we 'give up on the magician's wand' and 'follow the process', as advised by Bergson, we can see how the object of interest becomes detached from the population as such. As the mathematicians and analysts tell us, advanced analytics work not merely with a statistical notion of what is interesting, but also via an inductive process of knowledge discovery, in which the process generates the rules. Thus, in the 'chain analysis' of air transportation security algorithms, the person of interest emerges from the links of 'activities funded'; 'member of'; 'listed'; 'acquainted with'; 'travelled to'; 'countries visited'; and 'geopolitical events' (GAO, 2007). The population in the *n=all* formulation is a curious entity – a series of possible chains of association, in which there are no standard bell curves of normality, and from which anyone or anything could become a matter of interest or concern.

Second, the advent of data analytics brings significant *ontological* implications for thought and practice. The processes of ingestion, partitioning and machinic memory reduce heterogeneous forms of life and data to homogenous spaces of calculation. Algorithmic technologies, such as those we have described, tend to reduce differences in kind to differences in degree, as Bergson and Deleuze might say, or to distances data-point to data-point. From these processes of reduction and flattening it is thought that different kinds of life stories – patterns of life – emerge, and that interventions and decisions can be made on their basis. The affective world of an epilepsy sufferer, a sub-prime borrower, a border crosser, a terrorist or a criminal is thought to be excavable from the seams and joins of multiple data sources. But, what kinds of stories can be told with analytics? What happens to the things that cannot be spoken, or that which is not fully accessible to us even of ourselves? In current debates on protecting people from the worst vicissitudes of data mining and algorithmic decision, the emphasis is placed overwhelmingly on the restriction of processes to specified people of interest and specific queries defined in advance. As we have argued, this critical framing entirely misunderstands the work of the analytics. The stories of the mass or bulk data are precisely the means by which the queries are generated – the analytics require the mass data in order to decide on what or who is interesting, and this can only ever be retroactive.

Finally, the rise of analytics has important consequences for the form of contemporary politics. In effect, the analytical processes of ingestion, partitioning and reassembly and memory we have described make a particular claim in the world, they say *n=all*, *this is the world*, *here it is*, all data is rendered tractable. They carve out and convert radical heterogeneity into flat difference of degree, such that it appears as though everything is calculable, everything about the uncertain future is nonetheless decidable. 'With its applications which aim only at the convenience of existence', writes Bergson, 'science gives us the promise of well-being' (1965, p. 129). Today's little analytics promise a convenience of existence via the detection of all human propensities. What need, then, for politics? If politics expresses the fallibility of our world, the impossibility of resolution of all matters economic, social, ethical, then it exists because not everything is reducible and resolvable. Politics in our times confronts a ubiquitous analytics that imagines an infallible world where even the most turbulent of situations can be rendered tractable. Perhaps the most striking and troubling matter of scale in *big* data is the recognition of an extensity that always exceeds the capacity of human knowledge to collect and apprehend it. The promise and allure of the little analytics is to see those things that would otherwise be invisible, to perceive the imperceptible and to feed the insights to those who would action them. Confronted with this claim to reduce the fallibility of governing and decision, to do so in a manner that is 'fault tolerant', a political response must be mindful of the perils of the magician's wand, and point to the material, contingent and fallible processes that make this claim possible.

Acknowledgements

An early version of this paper was presented at the *Calculative Devices in the Digital Age* conference, Durham University, November 2013. Our thanks go to all participants at that event for the stimulating discussion that took place. The three anonymous reviewers for *Economy* & *Society* and Managing Editor Fran Tonkiss have been extraordinarily helpful and generous with their comments and insights; thank you. The research field-work from which the empirical elements of the paper are drawn, conducted during 2013, involved observations of data analytics industry and governmental events, and interviews with software engineers, analytics consultants and data scientists. We are grateful to everybody who has so generously given time to the project.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the RCUK Global Uncertainties Fellowship *Securing against Future Events: Pre-emption, Protocols and Publics* [Grant number ES/K000276/1].

Notes

[1] In response to the UK government's announcement of the second phase of funding for 'Big Data centres', Chief Executive of the ESRC, Professor Paul Boyle, welcomed the 'sheer volume of data that is now being created', a 'significant resource … that can shape our knowledge of society and help us prepare and evaluate better government policies in the future' (ESRC, 2014).

[2] It is not our purpose here to map a linear history of practices of data collection and analysis. Rather, we juxtapose two moments when a specific set of claims are made regarding the scale and scope of social data and its effects on the governing of societies.

[3] Bergson's reflections on perception in science are present throughout his body of work. Of particular significance here is his insistence on the shared categories of thought and sensing across science and prosaic perception, so that 'ordinary knowledge is forced, like scientific knowledge, to take things in a time broken up into particles, pulverized so to speak, where an instant which does not endure follows another without duration' (1965, p. 120).

[4] Indeed, by 1930, Bergson himself appreciated the growing capacity of 'modern mathematics' and physics to capture something of perpetual and indivisible change, to 'follow the growth of magnitudes' and to 'seize movement' from within (1965, p. 211).

[5] The earliest use of the concept of ingestion for analysis of data in multiple formats can be found in papers from IBM's research on smart surveillance and web architecture (Chiao-Fe, 2005; Gruhl *et al.*, 2004). The use of a vocabulary of ingestion coincides

with an expansion of analysable samples of digital data, such that it is said that *n=all*, or the sample is equal to everything.

6 The concept of index is used here in the sense proposed by Deleuze and Guattari to denote the capacity to designate the state of things, territorially locatable in time and space (1987, p. 124). Understood thus, for example, extraction algorithms are required in order territorially to index unstructured objects, as in the use of biometric templates derived from Facebook. It is the extracted template that makes the object searchable in time and space.

7 The case is derived from field-work conducted in London in 2013. For further examples and detailed descriptions of text mining and sentiment analysis, see Bello *et al.* (2013); Zhao *et al.* (2013); and Anjaria and Gudetti (2014).

8 Hayles defines the concept of 'technogenesis' as the 'idea that humans and technics have coevolved together', such that our very capacity for thought and action is bound up with 'epigenetic changes catalysed by exposure to and engagement with digital media' (2012, pp. 10–12). The idea is present also in Walter Benjamin's famous essay on art in the age of mechanical reproduction, where he notes that 'the mode of human sense perception changes with humanity's entire mode of existence' (1999, p. 216).

9 Retrieved from http://www.theguardian.com/profile/laura-poitras. See also Harding (2014, pp. 110, 204).

10 Though the focus of this essay is not on the interface between data architectures and software, the flattening of differences at this interface is significant. See Galloway (2012); Berry (2011).

11 Despite substantial interest in the automated analysis of large data sets for security purposes in the wake of Edward Snowden's disclosures, the use of algorithmic techniques to analyse Passenger Name Record (PNR) and SWIFT financial data has been known and documented for some time (Amoore, 2013; de Goede, 2012).

12 Insights drawn from observations at TIBCO® Spotfire® event, London, 13 June 2013.

13 Insights drawn from observations at TIBCO® Spotfire® event, London, 13 June 2013, and SAS Analytics 'How to' workshops, 19 June 2013.

References

Accenture. (2013). *Accenture analytics in action*. Retrieved from http://www.accenture.com/sitecollectiondocuments/pdf/accenture-analytics-in-action-survey.pdf

Agrawal, R., Asonov, D., Baliga, P., Liang, L., Porst, B. & Srikat, R. (2005). *A reusable platform for building sovereign information sharing applications*. SIGMOD Proceedings. Retrieved March 25, 2015, from http://www.rsrikant.com/papers/divo04.pdf

Agrawal, R., Imielinski, T. & Swami, A. (1993). *Mining association rules between sets of items in large databases*. SIGMOD Proceedings (pp. 207–217).

Agrawal, R. & Winslett, M. (2005). *An interview with Rakesh Agrawal*. SIGMOD. Retrieved March 24, 2015, from http://www.sigmod.org/publications/interview/pdf/D15.rakesh-final-final.pdf

Amoore, L. (2011). Data derivatives: On the emergence of a risk security calculus for our times. *Theory, Culture & Society*, *28*(6), 24–43.

Amoore, L. (2013). *The politics of possibility: Risk and security beyond probability*. Durham, NC: Duke University Press.

Anjaria, M. & **Gudetti, R. M. R.** (2014). A novel sentiment analysis of social networks using supervised learning. *Social Network Analysis and Mining*, *4*(3), 181–193.

Ansell Pearson, K. & Mullarkey, J. (Eds.). (2002). *Henri Bergson: Key writings*. London: Bloomsbury.

Bello, G., *et al.* (2013). Extracting collective trends from Twitter using social-based data mining. In C. Badica, N. T. Nquyen & M. Berezovan (Eds.), ICCCI 2013. LNAI 8083 (pp. 622–630).

Benjamin, W. (1999). *Illuminations.* London: Random House.

Bergson, H. (1912). *Matter and memory*. Mineola, NY: Dover Philosophical Classics.

Bergson, H. (1965). *The creative mind*. (M. Andison, Trans.). Totowa, NJ: Littlefield, Adams & Co.

Berry, D. M. (2011). *The philosophy of software: Code and mediation in the digital age*. Basingstoke: Palgrave Macmillan.

Berry, D. M. (2014). *Critical theory and the digital.* New York, NY: Bloomsbury.

Boyd, D. & **Crawford, K.** (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679.

Chiao-Fe, S. (2005). IBM smart surveillance system: an open framework for event based surveillance. IEEE Proceedings on advanced video and signal surveillance (pp. 318–373).

Connolly, W. (2011). *A world of becoming*. Durham, NC: Duke University Press.

Crary, J. (1999). *Suspensions of perception: Attention, spectacle, and modern culture*. Cambridge, MA: MIT Press.

Daston, L. (1995). *Classical probability in the enlightenment*. Princeton, NJ: Princeton University Press.

Daston, L. & **Galison, P.** (2007). *Objectivity*. New York, NY: Zone Books.

Dean, J. & **Ghemawat, S.** (2004). *MapReduce: Simplified data processing on large clusters*. Proceedings of the OSDI. Retrieved March 25, 2015, from http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf

Deleuze, G. (1991). *Bergsonism*. New York, NY: Zone Books.

Deleuze, G. & Guattari, F. (1987). *A thousand plateaus: Capitalism and schizophrenia*. Minneapolis, MN: University of Minnesota Press.

Dunham, M. H. (2002). *Data mining: Introductory and advanced topics*. Upper Saddle River, NJ: Prentice Hall.

Emerson, T., Ghosh, R. & Smith, E. (2012). Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications. London: Elsevier.

ESRC (Economic and Social Research Council). (2014). *ESRC big data network: Phase 2: Business and local government data research centres*. Retrieved April 27, 2014, from http://www.esrc.ac.uk/research/major-investments/Big-Data/BDN-Phase2.aspx

Featurespace. (2013). *Our technology*. Retrieved November 10, 2013, from http://www.featurespace.co.uk/technology/our-technology/

Foucault, M. (2007). *Security, territory, population: Lectures at the Collége de France, 1977–78*. Basingstoke: Palgrave Macmillan.

Galloway, A. R. (2012). *The interface effect*. Cambridge: Polity.

GAO (Government Accountability Office). (2007). Data mining: early attention to privacy in developing a key DHS program could reduce risks, GAO-07-293. Washington, DC: Government Accountability Office.

GAO (Government Accountability Office). (2013). *Decision: IBM v US Federal, B-407073*. Washington, DC: Government Accountability Office.

de Goede, M. (2012). *Speculative security: The politics of pursuing terrorist monies*. Minneapolis, MN: University of Minnesota Press.

Greenwald, G. (2014). *No place to hide: Edward Snowden, the NSA and the surveillance state*. London: Penguin.

Gruhl, D., Chavet, D. & Gibson, D. (2004). How to build a WebFountain. Utility Computing, *43*(1), 64–70.

Gupta, R., **Gupta, H.** & **Mohania, M.** (2012). Cloud computing and big data analytics. *Computer Science*, *7678*, 42–61.

Hacking, I. (1982). Biopower and the avalanche of printed numbers. *Humanities in Society, 5*(3–4), 279–295.

Hacking, I. (1986). Making up people. In T. Heller (Ed.), *Reconstructing individualism* (pp. 222–236). Stanford, CA: Stanford University Press.

Hacking, I. (1990). *The taming of chance*. Cambridge: Cambridge University Press.

Harding, L. (2014). *The Snowden files*. London: Faber and Faber.

Hayles, N. K. (2005). *My mother was a computer: Digital subjects and literary texts*. Chicago, IL: Chicago University Press.

Hayles, N. K. (2012). *How we think: Digital media and contemporary technogenesis*. Chicago, IL: Chicago University Press.

Hildebrandt, M. (2013). *Slaves to big data. Or are we?* Keynote at the 9th Annual Conference on Internet, Law & Politics (IDP, 2013, Barcelona). Retrieved January 14, 2014, from http://works.bepress.com/mireille_hildebrandt/52/

IBM. (2012). *IBM content analytics: Rapid insight for crime investigation*. Retrieved November 19, 2013, from http://public.dhe.ibm.com/common/ssi/ecm/en/zzs03073usen/ZZS03073USEN.PDF

Inmon, W. H. & **Nasaevich, A.** (2007). *Tapping into unstructured data: Integrating unstructured data and textual analytics into business intelligence*. Upper Saddle River, NJ: Prentice Hall.

Intelligence and Security Committee. (2014). *Transcript of evidence given by Theresa May MP, Home Secretary*. Retrieved October 25, 2014, from http://isc.independent.gov.uk/public-evidence

Intelligence and Security Committee. (2015). *Privacy and security: A modern and transparent legal framework*. London: HMSO.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. London: Sage.

Klemettinen, M., Mannila, H., Ronkainen, P. & **Toivonen, H.** (1994). Finding interesting rules from large data sets of discovered association rules. Proceedings of the 3rd International Conference on Information and Knowledge Management, 684–699.

Kroes, N. (2013). *Speech: The big data revolution (European Commission – Speech/13/261)*. Retrieved April 27, 2014, from http://europa.eu/rapid/press-release_SPEECH-13-261_en.htm

Latour, B. (1988). *The pasteurization of France*. Cambridge, MA: Harvard University Press.

LIBE (2013). Draft report on the US NSA surveillance programme. European Parliament Committee on Civil Liberties, Justice and Home Affairs: 2188.

MacKenzie, D. (2011). How to make money in microseconds. *London Review of Books, 33*(10), 16–18.

MapR for Apache Hadoop (2011). White Paper. Retrieved January 26, 2014, from https://www.mapr.com/sites/default/files/mapr_dist_white_paper.pdf

Marron, D. (2007). 'Lending by numbers': Credit scoring and the constitution of risk within American consumer credit. *Economy and Society, 36*(1), 103–133.

Mayer-Schönberger, V. (2009). *Delete: The virtue of forgetting in the digital age*. Princeton, NJ: Princeton University Press.

Mayer-Schönberger, V. & **Cukier, K.** (2013). *Big data: A revolution that will change how we live, work and think*. New York, NY: Houghton Mifflin Harcourt.

McKenna, A. (2010). The genome analysis toolkit: A MapReduce framework for analysing DNA sequencing data. *Genome Research*, *20*(9), 1297–1303.

Miller, P. & **Rose, N.** (1990). Governing economic life. *Economy and Society, 19*(1), 1–31.

Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Palo Alto, CA: Stanford University Press.

OECD (Organisation for Economic Cooperation and Development). (2013). *The OECD privacy framework*. Retrieved April 27, 2014, from http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf

Ohlhorst, F. (2013). *Big data analytics*. London: Blackwell Wiley.

Parisi, L. (2013). *Contagious architecture: Computation, aesthetics and space.* Cambridge, MA: MIT Press.

Raeder, T. & **Chawla, N. V.** (2011). Market basket analysis with networks. *Social Networks Analysis and Mining, 1* (2), 97–113.

Rose, N. (2006). *The politics of life itself: Biomedicine, power, and subjectivity in the twenty-first century.* Princeton, NJ: Princeton University Press.

Ruppert, E., Law, J. & **Savage, M.** (2013). Reassembling social science methods: The challenge of digital devices. *Theory, Culture & Society, 30*(4), 22–46.

SAS Analytics (2013). Field-work observations of SAS knowledge discovery processes, London.

Tachmazidis, I., Antoniou, G. & **Kotoulas, S.** (2012). *Scalable nonmonotonic reasoning over RDF data using MapReduce*. Proceedings of the Joint Workshop on Scalable Semantic Web Systems (pp. 75–90) 11th November, Boston, USA.

The New York Times. (2013, September 28). NSA gathers data on social connections of US citizens.

Zhang, S. & **Wu, X.** (2001). Large scale data mining based on data partitioning. *Applied Artificial Intelligence, 15*(2), 129–139.

Zhao, L., Yingjie, R., Wang, J., Meng, L. & **Zou, C.** (2013). Research on the opinion mining system for massive social media data. In G. Zhou *et al.* (Eds.), NLPCC 2013, CCIS 400 (pp. 424–431) November 15th–19th, Chongqing, China.

Louise Amoore is Professor of Political Geography in the Department of Geography, Durham University. She is RCUK Global Uncertainties Fellow (2012–2015) and author of *The politics of possibility: Risk and security beyond probability* (Duke University Press, 2013).

Volha Piotukh is currently Postdoctoral Research Associate at Durham University, working on the ‘Securing Against Future Events’ project. She is the author of *Biopolitics, governmentality and humanitarianism: ‘Caring’ for the population in Afghanistan and Belarus* (Routledge, 2015).