16 September 2015
Forthcoming in *The Oxford Handbook on the Law and Regulation of Technology*
edited by Roger Brownsword, Eloise Scotford and Karen Yeung (Oxford University Press, 2016)

Data Mining as Global Governance

Fleur Johns*

Abstract

Data mining technologies are increasingly prominent in development and aid initiatives in which context they may be understood to be doing work of global governance. This chapter explains how data mining may be so characterized and explores how this work may be compared to more conventional governance techniques and institutions. The chapter proceeds through three stages. First, it provides an overview of some exemplary initiatives among international institutions in which data mining plays a crucial role. Second, a playful, mundane analogy for a governance challenge is presented – the sorting of a sock drawer – and a familiar law and policy approach and a data mining approach to this challenge are compared. Third, the chapter highlights what may be at stake in the practice of data mining on the global plane and associated shifts in regulatory technique, arguing for this practice to be regarded as a matter of broad-ranging public concern.

Keywords

Data mining; global governance; international organizations; development; humanitarian aid; law; policy; regulatory technique

Introduction

Putting the terms ‘data mining’ and ‘governance’ together in a chapter heading may

* Thanks are due to the editors of this Handbook and the following people who generously read and commented on prior versions of this chapter: Lyria Bennett Moses, Janet Chan, Roger Clarke, Alana Maurushat.

evoke a number of expectations of the text to follow. Perhaps one might expect to read about data mining as an instrument in the governance toolbox; something which lawyers and others are using *for* governance, with positive and negative effects (eg Nissan 2013; Zarsky 2011). Alternatively, one might anticipate a story of the governance *of* data mining; an overview of how laws of various jurisdictions guide and restrict the practice of data mining, or should do so (eg Cate 2008; Schwartz 2001; Solove 2008). One might foreshadow, instead, a tale of data mining *about* governance; recounting ways in which the practice of governance has become something that people aspire to measure globally: through the use of indicators, for instance (Fukuyama 2013: Davis, Kingsbury & Merry 2012).

Data mining *as* governance suggests something else. It suggests that datasets, databases and data mining technologies and infrastructure are not just instruments for governance to be conducted otherwise on the global plane, nor practices opposable to law that await further or better governance, nor constraints that operate alongside but remain neatly distinguishable from law (contra Lessig 1998). Rather, these technologies and related infrastructure constitute a field and a style, or a number of related styles, of governance. To carry out data mining amid the kind of projects outlined below is to perform work that we may associate with that nebulous, ascendant term 'governance', or with 'the 'law' and 'regulation' of this book's title (Black 2002; Lobel 2004). That is, data mining operations are directive and standardizing; they constitute offices and subjectivities; they assemble information and seek to modify behaviour; they shape understanding of what is imaginable or achievable and who or what is 'right' and 'wrong' (or some proxy for those terms: efficient and inefficient; reasonable and unreasonable; just and unjust; countable and

uncountable; and so forth) according to certain norms, as well as how and why those norms might change over time. They do so, moreover, in ways that purport to address a wide range of governance dilemmas on the global plane: from disaster relief to food security; pandemic control to refugee registration; anti-corruption to environmental impact assessment and beyond (see, respectively, Meier 2011; Wang and others 2012; French and Mykhalovskiy 2013; Jacobsen 2015; Su and Dan 2014; Goetz and others 2009).

It is the argument of this chapter that in light of the operations in which it has become crucial, data mining should be understood as a practice of global governance – both as a *technique* (or set of techniques, not internally consistent: Law and Ruppert 2013: 232) and as a *site* for the assemblage and distribution of value and authority in which the public (variously configured) has significant stakes. This argument will be developed, first, by explaining data mining in general and surveying some indicative practices of global governance in which data mining plays a crucial role. In other words, this chapter begins by considering something of *what* is being accomplished and attempted on the global plane by recourse to data mining. Second, I present a deliberately 'low-tech', mundane analogy of the sock drawer (mundane, at least for, those privileged with an array of such possessions), militating against the sense of alchemy and awe by which discussions of contemporary data mining are often characterized. By this means, I will show something of *how* data mining governance proceeds in comparison to more conventional regulatory practices. Third, and finally, I will focus on *why* data mining is something with which global publics should be concerned and engaged.

1. Mining with Models

Much has been written about the collection, mining and sharing of data by national governments and corporations, for law enforcement, welfare surveillance and intelligence purposes especially, and privacy and related normative concerns provoked thereby (eg Rubinstein, Lee and Schwartz 2008; Chan and Bennett Moses 2014; Pasquale 2015). Far less scholarly or public attention has, however, been dedicated to data mining by international organizations and its potential ramifications for global law and policy (especially ramifications beyond considerations of privacy). Growing emphasis on data mining in global governance has, nonetheless, been heralded by the publication of several major reports by international organizations, both intergovernmental and non-governmental, highlighting the current and projected importance of data's automated analysis in their work (ICRC 2013; UN OCHA 2013; UN Global Pulse 2013; see generally Taylor and Schroeder 2014).

A sense of the expanding role of data mining in international organizations' work may be gleaned from a brief overview of three illustrative initiatives, described below: first (the 'UNHRC program'), a United Nations High Commissioner for Refugees (UNHRC) program for biometric registration and de-duplication of Afghan refugees living in camps in Pakistan and applying for humanitarian assistance for repatriation following the fall of the Taliban (Jacobsen 2015); second (the 'UN Global Pulse study'), a collaborative, United Nations (UN)-led initiative (involving the UN Global Pulse, the World Food Program, the Université Catholique de Louvain and a Belgian data analytics company, Real Impact Analytics) to use digital records of mobile phone transactions as a proxy for assessing and mapping non-monetary poverty (Decuyper

and others 2015); and third (the 'AIDR platform'), the Artificial Intelligence for Disaster Relief platform: a free and open source prototype designed to perform automatic classification of crisis-related messages posted to social media during humanitarian crises. The AIDR platform was developed by researchers at the Qatar Computing Research Institute and has been deployed in collaboration with the UN Office for the Coordination of Humanitarian Affairs ('UN OCHA') (Imran and others 2014; Meier 2015). To understand exactly how data mining features in each of these initiatives, some lay explanation of that term is required.

A. What is Data Mining?

Data mining entails the computerized production of knowledge through the discernment of patterns and drawing of relationships within large databases or stores of digital information – typically patterns and relationships not otherwise apparent. In contrast to 'knowledge discovery in databases' or 'KDD', data mining does not necessarily include control over data collection. It often deals with byproducts of other processes; data assembled for data mining purposes may 'not correspond to any sampling plan or experimental design' (Colonna 2013: 315-316; Azzalini and Scarpa 2012: 8).

Data with which data mining deals may be structured or unstructured, or in some combination of these two states. Structured data is organized into fixed dimensions or fields, each representing a specific yet generalizable characteristic or response to a generic query, such as name or date of birth. Unstructured data, in contrast, has no pre-defined organization and often combines many different data forms; data

constituted by a video stream is an example. The concern of data mining is 'the extraction of interesting (nontrivial, implicit, previously unknown and potentially useful) information or patterns from data in *large* databases' however that data may be assembled or structured (Han and Kamber 2001: 5). The scope of what might be 'potentially useful' in this context need not be determined a priori; that is, data mining itself may generate a sense of what merits interest, as described below (Azzalini and Scarpa 2012: 5). The 'database' so mined need not, moreover, be centralized. Much contemporary data mining concerns data that are decentralized or 'distributed' – that is, gleaned from many different, uncoordinated sites and sources (Kargupta and Sivakumar 2004; Leskovec, Rajaraman and Ullman 2014).

Crucially for governance purposes, data mining may take supervised or unsupervised forms (or semi-supervised hybrids of the two). Supervised data mining proceeds from a training set of data known to have certain features: a record of past successes and failures, or pre-identified instances of the type of norm-deviating event of interest to the human (or non-human) supervisor(s). The goal is for data mining software to learn the signature, or generate a number of possible signatures, of points of interest in the training data and classify other unlabeled data employing that or those signature(s). Unsupervised data mining, on the other hand, commences without an initial model, hypothesis, or norm from which deviation must be sought. The aim is to generate and explore regularities and anomalies; to infer the properties of some function capable of predicting phenomena in the data; to create a model on that basis; and to continuously refine those inferences and the ensuing model (see generally Leskovec, Rajaraman and Ullman 2014: 415-417). Supervised mining offers a clear measure of success and failure (or degree of error) and a basis for redressing the latter; learning takes place

through the detection and correction of errors. Unsupervised mining offers no ready way of evaluating the validity or usefulness of inferences generated; part of the process is continually revisiting and discarding hypotheses which the data mining practice itself will have generated (Hastie, Tibshirani, and Friedman 2009). Even when unsupervised, however, data mining comprises part of a complex 'socio-technical system' in which humans and non-humans interact in a myriad of ways, as is apparent in the accounts of data mining endeavors set out below (Colonna 2013: 335; Nissenbaum 2010: 4-5; see generally Suchman 2007).

One might expect the design and deployment of data mining tools in relation to existential matters – disaster relief and the like – to be reflective of the human stakes at play in that work. However, because of the way data mining code and tools often get bolted together in a piecemeal fashion, customized, reused, and repurposed away from the settings in which they were originally developed, this will not necessarily be the case (Clements and Northrop 2002). Google's famous PageRank algorithm, for example, was developed as the core product of a commercial enterprise, but has been retooled for a wide range of data mining purposes outside that setting, including poverty mapping (Leber 2014; Pokhriyal and others 2015). Each of the initiatives outlined below exhibits precisely this kind of software and hardware retooling.

B. Three Illustrations of Data Mining as Global Governance

The UNHCR program was initiated in the context of the mass-repatriation of Afghan nationals from refugee camps in Pakistan back to Afghanistan, after the fall of the Taliban in 2001. Between 2001 and 2005, the UNHCR facilitated the return of over three million refugees to Afghanistan (Kronenfeld 2008). As part of this process, the UNHCR provided for every returnee to receive 'transport assistance ranging from $5 to $30 – depending on his [or her] final destination – a UNHCR family kit with plastic tarpaulin, soap and hygiene items, as well as wheat flour from the World Food Programme' (UNHCR 2002). In distributing these resources, the UNHCR used traditional identification methods to try to distinguish 'genuine' first-time claimants from 'recyclers' claiming multiple assistance packages, but found these methods wanting (UNHCR 2002; UNHCR 2003a, 2003b). At the UNHCR's request, commercial technology vendor BioID Technologies ('BioID'), in cooperation with Iridian Technologies, developed a biometric registration facility and mobile registration units for the organization's deployment of preexisting iris recognition technology, the operation of which was described as follows:

> All centers have a network of Iris Recognition cameras (ranging from 2 - 9 depending on the required capacity). The individual is asked to sit down in front of one the cameras and is briefed by the operator. A series of enrollment images are taken and sent to the server in the network. This system converts the appropriate image into an Iriscode (a digital representation of the information that the iris pattern constitutes) and checks the entire database whether that IrisCode matches with one already stored. If that is not the case,

> the individual is enrolled, the IrisCode stored in the database and a Customer Information Number (CIN) is returned to the particular workstation confirming that the enrollment has been successful…If the individual is found in the database, the system returns an alarm to the workstation with the message that a recycler has been found and also returns the CIN number that individual was originally enrolled with. The whole process from the moment the person sits down, is briefed, up to completion of enrollment takes less than 20 seconds (BioID no date).

The techniques used to extract (demodulate), analyse and classify phase information (a numeric expression – in the form of a 'bit stream' – of a pattern extracted from a set of iris images) have not been described publicly by either BioID or UNHCR (see generally Daugman 2004). Nonetheless, published descriptions of iris recognition techniques suggest that this may involve a type of data mining model known as a neural network, employing machine learning (Lye and others 2002; Cao and others 2005; Sibai and others 2011; Bowyer and others 2008; Burge and Bowyer 2013: 79-80). While neural networks vary widely, they are all predicated on the processing of numeric input through a series of interconnected nodes (some layers of which are hidden) and the attribution to connections among those nodes of associated weightings, with each layer of these nodes being comprised of the weighted sum of values in the first or preceding layer. In many instances, the weighting attributed to nodal connections is 'learned' through the processing of, and verification of performance against, a training set of input data (Roiger and Geatz 2003: 45-47; 245-264). Alternatively, it may be that this iris recognition is carried out using a decision tree: another type of predictive data mining model used for classification, again

employing machine learning (Burge and Bowyer 2013: 275; Kalka 2006). Decision trees are '[t]ree-shaped structures' that represent sets of binary tests on the basis of which data is divided and classified at each 'branch'; after training and validation of outputs, the tree can be used to 'generate rules for the classification of a dataset' without supervision (Sumathi and Suvanandam 2006: 402; Roiger and Geatz 2003: 9-11).

After a year of this system's operation, and the processing of just over 200,000 refugees, UNHCR reported that approximately 1,000 people trying to claim multiple assistance packages had been detected 'in addition to more than 70,000 families …rejected [during the same period]…under other screening methods' (UNHCR 2003b, 2003c). Those other screening methods – maintained alongside iris recognition – included 'interviewing potential returnees and examining their family photos' (UNHCR 2002). The relationship between these various screenings tactics is not explained in UNHCR literature, but that literature does suggest that the biometric screening was treated as dispositive. Indicatively, the iris recognition system was said to have performed 'flawlessly' despite the risk of data corruption posed by 'the heat and dust of Pakistan's border territories with Afghanistan', without reference to error rates associated with factors such as image compression; contact lens use; pupil dilation; corneal bleaching, scarring, inflammation and other pathologies (UNCHR 2003b; on error rates, see Al-Raisi and Al-Khouri 2008; Vatsa, Singh and Noore 2008; Bowyer and others 2013). Similarly, according to the UNHCR, concerns that use of the technology might intimidate, raise traditional objections to women being photographed, or compromise privacy proved unfounded: 'only the eye is seen onscreen'; '[t]ests on women and children are done by female refugee agency

workers'; and 'the code describing the iris has no link to the name, age, destination or anything else about the refugee' (UNHCR 2003c). Commentators have, however, been critical of the organization's failure to disclose the risk of false matches likely to arise in large-scale applications of biometric technology, or to put in place measures 'to detect and correct for such false matches', especially in view of the fact that data anonymization might hinder their detection (Jacobsen 2015: 151-2). Even if the prospect of undetected error could be adequately and publicly addressed (not the focus of this chapter), the UNHCR program still raises issues of changing regulatory style and shifting distributions of authority to which we will return below.

The UN Global Pulse study represents another example of predictive data mining being used to address a perceived paucity of reliable data in developing countries. In this instance, however, a traditional, 'verification-driven' approach was used, employing statistical analysis, rather than a 'discovery-driven' or machine learning approach (Colonna 2013: 337-340). The starting point of the study was the thesis – drawn from a series of prior studies – that 'phone usage data represent a clear barometer of a user's socio-economic conditions in the absence or difficulty of collecting official statistics' (Decuyper and others 2015: 1). On this basis, the study sought to test the further hypothesis that 'metrics derived from mobile phone data', specifically CDRs (or call detail records, including caller and callee identification data, cell tower identification data, dates and times of calls) 'and airtime credit purchases' (data comprised of the relevant user's identifier, the top-up amount, dates and times of top-ups) might serve as a 'real-time proxy' for 'food security and poverty indicators in a low-income country context' (Decuyper and others 2015: 1).

The method used to test this hypothesis entailed calculation of mathematical relationships across two data sets, both aggregated by geographical areas home to between 10,000 and 50,000 inhabitants in 'a country in central Africa' (Decuyper and others 2015: 2-3). The first data set – drawn from mobile phone company records maintained for billing purposes – was comprised of caller home location data, measures of caller 'top-up' (or airtime credit purchase) behavior, and measures of caller 'social diversity' (how equally a caller's communication time is shared among that caller's contacts): the latter having been shown otherwise to be a 'good proxy' for variation in poverty levels (Decuyper and others 2015: 2-3; Eagle and others 2010). The second data set – drawn from a 2012 survey of 7500 households across the country in question, made up of 486 questions, including questions related to food access and consumption – was comprised of a 'set of numerical metrics related to food security', some of these were question-specific measures and some were composite measures related to several questions (Decuyper and others 2015: 3-4). The second data set was designed to provide 'ground truth' data by which to validate the first (Decuyper and others 2015: 2).

Correlations (numeric representations of the interdependence of variables) were computed among thirteen mobile phone variables and 232 food consumption and poverty indicators. Relationships among those variables were then modeled using regression analysis (that is, modeling around a dependent variable of interest to explore its predicted or possible relationship to one or more independent variables and the contribution that the latter may make to variation in the former) (Decuyper and others 2015: 4). The results of these analyses were taken to support 'a new

hypothesis' that 'expenditure in mobile phone top up is proportional to the expenditure [on] food in the markets' (Decuyper and others 2015: 5).

These results from the UN Global Pulse study encouraged the authors to envision that governments and other 'partners' running 'programs and interventions' concerned with food security and poverty could collaborate with mobile carriers to generate 'an early warning system' of 'sudden changes in food access' and have their policy 'guide[d]' accordingly, including using this 'early warning' as a prompt to gather further information, through in-depth surveys for example (Decuyper and others 2015: 6-7). Although the authors of the UN Global Pulse study did not address how such targeted, follow-up surveys might be conducted, it is conceivable that any such survey methodology might employ a further set of data mining techniques. Because the use of mobile phones as platforms for survey data collection in developing countries has risen, so research on data mining techniques designed to automate data quality control during mobile survey data collection is also growing (Chen and others 2011; Birnbaum and others 2012). Using training sets known to contain both fabricated and 'relatively accurate' survey responses, machine learning data mining for this purpose seeks to 'find anomalous patterns in data' on the basis of which one might detect 'fake data' (such as data relating to home visits that malingering data-gatherers never conducted) or 'bad data' (emanating from 'fieldworker[s] acting in good faith' but subject to some 'misunderstanding or miscommunication') (Birnbaum and others 2012). Thus, the sort of in-depth inquiry that the UN Global Pulse study anticipates following from its 'early warning' mechanism may itself take the form of data mining, at least in part, aimed at purging flawed data.

Misinformation and superfluous data are also targeted by the AIDR platform (which, as indicated above, refers to the Artificial Intelligence for Disaster Relief platform) which 'tackles the overwhelming amount of information generated by mobile phones, satellites, and social media' in the midst and aftermath of a humanitarian disaster to 'help aid workers locate victims, identify relief needs, and… navigate dangerous terrain' (Meier 2015). To do so, the AIDR platform 'collects crisis-related messages from Twitter ("tweets"), asks a crowd to label a subset of those messages, and trains an automatic classifier based on the labels' as well as 'improv[ing] the classifier as more labels become available'. This approach, combining human and automated classification, is designed to train 'new classifiers using fresh training data every time a disaster strikes' ensuring 'higher accuracy than labels from past disasters' and meeting the changing informational needs of disaster victims and responders (Imran and others 2014: 159-160; Vieweg and Hodges 2014).

Data collection in this context – that is, the secondary data collection associated with organizing messages on the AIDR, not the primary data collection associated with Twitter users determining about what, when and how to write a message – is initiated by an individual or collective AIDR user entering a series of keywords and/or a geographical region for purposes of filtering the Twitter stream. On this basis, a 'crowd' of annotators provides training examples – each example comprised of a system-generated message with a human-assigned label – to train for classification of incoming items. Training examples may be obtained from the collection initiator or 'owner' using AIDR's 'internal web-based interface' or by calling on an external crowdsourcing platform: AIDR makes use of the open source platform PyBossa. This interactive training process generates output – made available, through output

adapters, as application programming interfaces ('APIs') – in the form of messages sorted into categories that may be collected and used to create crisis maps and other types of reports and visualizations, using the APIs (Imran and others 2014: 160-1).

The mining of data collected through AIDR is effected by the 'AIDR Tagger' and the 'AIDR Trainer'. The AIDR Tagger, responsible for each Tweet's classification, is comprised of 'three modules': a feature extractor (which extracts certain features from a Tweet); a machine learning module; and a classifier (which assigns one of the user-defined labels to the Tweet). The AIDR Trainer feeds the learning module of the AIDR Tagger, using 'trusted' training examples sourced from the collection owner or crowd-sourced examples processed via PyBossa (Imran and others 2014: 161-2). The learning module adopts a 'random forest' data classification methodology: an aggregation of several, successively split decision trees (Imran and others 2015; Boulesteix and others 2014: 341). Once trained to compute proximities between pairs of cases, random forest classification may be extended to unlabeled data to enable unsupervised clustering of data 'into different piles, each of which can be assigned some meaning' (Breiman and Cutler no date). Clustering entails automated gathering of data into groups of records or 'objects' that exhibit similarities among them, and dissimilarities to objects assembled in other groups. In unsupervised clustering, the likenesses or associations that comprise a particular group's relatedness are not known or predicted in advance; rather, these emerge through machine learning (Berkhin 2006). The AIDR platform has been tested in relation to Typhoon Yolanda in the Philippines and the earthquake in Pakistan in 2013, as well as in the Nepal earthquake in 2015 and elsewhere (Vieweg and others 2014; Imran and others 2014; Meier 2015).

Each of these initiatives brings slightly different data mining techniques to bear upon a perceived dilemma, for intergovernmental and non-governmental organizations, of a lack, overload or chronic unreliability of data likely to be useful for governance and other institutional actors, primarily in developing countries. Critics have raised concerns that are salient for these types of initiatives, including worries with regard to the technological circumscription of choice, the overestimation of technologies' reliability, and their propensity for non-transparency and 'function creep': that is, using data collected for unanticipated and unannounced purposes (Brownsword 2005; Mordini and Massari 2008; Jacobsen 2015). Also circulating in scholarly literature are worries about the economic and political logics said to be 'underlying' these measures (Sarkar 2014; Pero and Smith 2014). The aim of this chapter is not to reproduce or appease these concerns. Rather, this chapter focuses on shifts in global regulatory style or governance practice that these examples may signify, not as a matter of underlying logic, but rather on their surface (on the critical richness of the surface, see Hacking 1979: 43). In order to track some of these surface shifts, let us turn away from the technical language of data mining towards a mundane analogy, to compare some conventional approaches to knowledge-production and ordering in law and regulation with a data mining approach to the same.

2. The Sock Drawer

Let us imagine a banal 'regulatory' challenge: the need to order a messy sock drawer in a way that renders it usable and acceptable to a number of people likely to access it. There are various ways one might approach this task, and a range of considerations

that might come up throughout, as described below. Each of the tactics or considerations detailed in the first part of this section is roughly analogous to a strategy or possibility that might emerge in the course of some traditional global governance practice: perhaps, say, in the course of multilateral treaty drafting and negotiation, or treaty modification after adoption (whether through later amendment; subsequent, more specialized agreement; or some parties' entry of reservations – opt-outs or qualifications to derogable treaty provisions) or in the process of treaty ratification and implementation by parties. The second part of this section seeks to represent, by admittedly obtuse analogy, how the same regulatory challenge might be approached through one particular type of data mining practice.

A. *Conventional Governance of the Sock Drawer*

If one were to set about trying to 'govern' a messy sock drawer using conventional legal and regulatory techniques prevalent globally, one might begin by setting out a general principle or preambular aim: say, in order to promote timely, comfortable and aesthetically pleasing dressing, socks in the drawer shall be sorted into pairs and single socks discarded. Already, this principle contains a condition: that of availability in the particular drawer (and household) in question. It also includes a clear, question-begging omission: *whose* timeliness, comfort and pleasure should be at issue; a particular individual's, those of the members of a certain household or group, or a population's at large? In other words, what is the scale and scope of the pair-or-discard imperative and who has a stake in it?

Alternatively, one might begin the process of governance by confronting the initial framing of the exercise. Should socks be kept in a closed drawer at all? Would an open tub, or a series of pigeonholes on a wall, work better as a storage mechanism? In what ways and for whom would one or the other of these options be 'better'? Are matching socks actually more aesthetically pleasing or comfortable than unmatched socks? According to which criteria or for whom?

Having confronted these questions (and answered them – however provisionally), it may be considered timely to get stuck in to the task of human sorting. The experience of doing so would likely lead to the consideration and adoption of further rules, conditions, and exceptions. Perhaps only available socks in reasonable condition should be sorted into pairs and holey ones discarded (raising a further question: how and by whom should 'reasonable condition' be assessed)? Accordingly, one might add a rule allocating that responsibility and explaining how it should be exercised: a rule limiting the sorting imperative, for instance, to available socks judged to be in reasonable condition by their owner, taking into consideration any holes or other signs of wear and tear. The issue of sock ownership then rears its medusan head; perhaps 'possession' is a preferable alternative?

Even after ownership or possession is established to the satisfaction of the constituency at hand, other concerns may arise, either from the outset or as one encounters socks of different kinds and conditions. Should socks made of high quality, expensive material – cashmere socks, for example – or socks manufactured using an environmentally costly process – polyester socks, for instance – be recycled instead of discarded, to minimize waste and maximize sustainability? Should socks to

which the possessor has an emotional attachment – those originally hand-knitted as a gift, perhaps – be exempted from the pair-or-discard imperative, to prevent emotional harm? Should especially woolly socks be retained in a cold climate setting – even when odd – and more readily discarded in a temperate or tropical location? Should socks be sniffed in the process of their evaluation and smelly socks thrown away? If so, by whom should this smell test be conducted, and what happens if that person comes down with a cold, obstructing their nose? Considerations such as these may encourage further exceptions or more detailed directives to be adopted and responsibilities assigned.

Issues of participation, equity and compliance will also arise. Who has access to the sock drawer in question and how might they regard the sorting scheme adopted? What of those who do not have access to this drawer, or to socks at all? Are either or both of these 'constituencies' likely to take an interest in, support, and adhere to the sock-sorting arrangements developed? If not, and if their support is considered necessary or desirable, how might they be encouraged to do so? This may be partly a matter of cultivating or reflecting prevailing tastes: are the sock pairings proposed likely to strike the sock wearers in question as intuitively 'right'?

One could represent the sorting process so developed as a decision tree: a series of binary choices building on one another. Alternatively, one could understand this sorting process in terms of clustering: it will be acceptable to some to gather roughly the same category of sock – say, all black*ish* socks – and to form pairs from among that cluster. How one chooses to represent the process will likely have an effect on how its overall acceptability may be viewed. Yet the method of sorting – however

represented – is unlikely to displace the recurrent demand for dialogue around the sorts of questions raised so far.

To guard against misapplication or misinterpretation of such rules as are adopted, one might introduce a review possibility by, say, inviting some trusted third party to judge the suitability of the pairings for wearing in public and to rule some pairings in and out. One might also opt to try on a pair of socks, or a succession of pairs, in front of an audience from which opinions as to their stylishness may be 'crowd-sourced', electorally or by consensus resolution of a representative body. These are both familiar techniques in conventional governance practice on the global plane (see generally Kingsbury, Krisch and Stewart 2005; Best and Gheciu 2014). Others may conduct selective, trial outings in certain pairs of socks, to determine how comfortable or likely to fall down they may be when worn. Some might prefer to delegate the sock-sorting process altogether, asking someone to undertake the task on the basis of guidelines provided or with untrammeled discretion. Others might seek external input while retaining ultimate sock-sorting responsibility: expert advice, for example, as to the optimum number of socks required to ensure one has a clean pair available each day, given a specified number of launderings per week; a cost-benefit analysis on sock retention versus sock renewal, after an audit of the socks in stock; or scientific input as to the projected loss of body heat from the foot and ankle under different climatic conditions and its effect on the body. Again, these are analogous to regulatory techniques widely used in global law and policy.

Regardless of the process adopted or outcomes ultimately realized, 'governing' a sock drawer using some or all of these familiar regulatory strategies places the practice of

governance itself in the front and centre of deliberation. It is apparent that different methods will satisfy different constituencies and that continual reconsideration and/or review as to both method and outcome may be required to settle unforeseen questions, concerns and dilemmas as they arise. This iterative revisitation seems, moreover, likely to be multi-directional: potentially running up and down the hierarchy of rules, from overarching principle to the most nuanced of exceptions and back again, and involving horizontal comparison across different rule and sock categories and different subsets of the sock wearing constituency.

Even where delegation is involved, the regulatory strategies caricatured above are immersive in that they are likely to be predicated on, or referable to, some individualized and collective human experience of wearing socks (or not) and observing sock-wearing in others (or not). That is not to say that those devising the sock sorting strategy will have worn all the socks in question. Nonetheless, at least in a democratic setting, they would probably be exposed to some representation of the views, tastes and experiences of those who have worn or tested many different sorts of socks: by receiving delegations or petitions from sock manufacturers', sock wearers' or sock enthusiasts' associations, for example, or occasional submissions from different members of the relevant household. Certain accounts of the 'authentic' sock-wearing experience might surface in the course of this interaction, or a sense of the sock as a 'social construction'. Such accounts are, however unlikely to prove dispositive for those for whom sorting the sock drawer is a daily matter of concern (Latour 2005).

Each of these conventional strategies is also, quite patently, relative and vulnerable to counterclaim. Speculation about sock sorting may seem irrelevant, even indulgent, to someone who does not possess socks; does not regularly sleep inside, or in a room in which a chest of drawers or other storage vessel is available; or is guided most by religious or cultural teachings concerning the covering of the foot (which could be differentiated by gender and age). Moreover, questions of authority and interest – who bears authority for what purposes, how should this authority be exercised, and in whose interest – seem to remain alive throughout this inquiry, however trivial the subject matter.

B. *Data Mining the Sock Drawer*

Let us now try to envisage approaching the governance challenge posed by the messy sock drawer through data mining. In this section, possibilities for sock sorting will be examined through the lens of just one mode of data mining: an often unsupervised or semi-supervised descriptive data mining technique known as *k*-means cluster analysis. Data mining is termed 'descriptive' when its aim is not merely to divide and classify data according to known attributes or factors, but rather to represent data in unforeseen ways, disclosing 'hidden traits and trends' within a dataset (Zarsky 2011: 292; Colonna 2013: 345).

Recall that the AIDR platform discussed above makes use of the technique of clustering. Data mining may produce clusters in a range of ways: using statistical methods; genetic algorithms (search techniques based on ideas from evolutionary biology that seek to 'evolve' a 'population' of data states based on their 'fitness'); or neural networks, among others (Adriaans and Zantinge 1996: 8; Hand and others

2001: 266). Nonetheless, a *k*-means algorithm, first published in the 1950s, remains one of the most popular clustering tools (Jain 2010).

K-means clustering algorithms organize data around a set of data points, each known as a centroid, with the distance of data from a centroid representative of the degree of discrepancy or dissimilarity between them. Centroids are not predetermined; after a guess as to the appropriate number of clusters for the task at hand, and an initial, random positioning of centroids, centroid positions are recomputed and clusters reassembled iteratively with a view to minimizing the distance of data to centroids across all clusters (and, if agglomerative methods are also employed, between clusters). Often, multiple cluster optimization sequences will be run, using different, randomly selected starting points, to mitigate the likelihood of the clustering algorithm converging on 'local' rather than 'global' affinities, or making too much of outliers, and thereby missing potentially significant relationships and patterns (Berkhin 2006: 15-18; Hand and others 2001: 293-326).

In order to sort a sock drawer using *k*-means clustering techniques, one would begin with a decision about the *k*, or the number of clusters to identify. If the aim remains sorting into matched pairs, this might be based on an estimate of the number of pairs likely to be in the drawer. Two further parameters would also require initial, subjective specification: the process by which the initial partitioning of clusters will be effected (by one or other method of randomization) and the choice of metric, or similarity measure, to determine the distance between items in the cluster (which will in turn often depend on the choice of scale used, if variables in the data have been standardized prior to clustering: Mohamad and Usman 2013). The latter will include a decision as to what intrinsic features of the data should be considered when assessing

similarity and dissimilarity, or how the data should be represented (Jain 2010: 654-6). This could be based on some probabilistic calculation of the mixture of socks, or some other initial premise concerning how socks' 'pairness' might best be determined. In any event, the choice of algorithm(s) – both for initial partitioning and subsequent cluster optimization – will have a significant bearing on the way socks are sorted, as '[d]ifferent clustering algorithms often result in entirely different partitions even on the same data' (Jain 2010: 658).

For those, like the author, lacking information technology and statistical expertise, governing a sock draw by *k*-means clustering will involve employment, consultancy or delegation. Because of the likelihood that authority to make parameter-defining decisions will presumptively rest with those most familiar with data mining techniques, it is probable that the initial, parameter-defining decisions described above would be taken by the data mining specialists charged with their execution, without much by way of directive input from 'clients', sock-wearers or third parties. As Bendoly observed, drawing on semi-structured interviews with 'representatives from different facets of the data mining community', '[t]he [data] analyst is ultimately charged with the responsibility of transferring as much of relevant analytical knowledge…, or… [at] least the informational rules and relationships derived by the algorithm to the decision maker', a process that tends often to fall prey to 'black box internalization of consulting prowess' (Bendoly 2003: 646).

The clusters of socks that result from this data mining process might not correspond closely, if at all, to preexisting presumptions about or perceptions of 'pairness'. Depending on the data or data collection technologies to which it has access, an unsupervised clustering algorithm could find 'dense' relationships between socks

based on factors non-detectable by humans or otherwise insignificant to most wearers of socks. It could, for instance, group socks into pairs based on similarity (and dissimilarity from others) in terms of their weight; their snag or pilling resistance; the elongation or air permeability of the fibres of which they are made; their lint content; their flammability and so on. As Jain remarks, '[c]lustering algorithms tend to find clusters in the data irrespective of whether or not any clusters are ["natural[ly]"] present' (Jain 2010: 656).

Data mining outcomes departing wildly from expectations might prompt the sock sorters in question to have recourse to semi-supervised clustering. One could, for instance, introduce one or more 'must-link constraints' specifying that certain socks must be assembled within the same cluster (all blue socks, or socks of similar size, for example). Alternatively, one could 'seed' the algorithm with some 'correctly' labeled data (that is, correctly paired socks), the soundness of which has been extrinsically determined. Such constraints or seeding data might be provided by a 'domain expert' – someone who knows socks in general, or this sock drawer in particular – or derived from externally sourced information about the ontology of the data domain (that is, the ontology of socks) (Jain 2010: 660-661). These measures parallel, somewhat, the effect of exceptions, detailed directives, and review opportunities described in the narrative of 'conventional' regulation presented above. Alternatively, outcomes that initially seem unsatisfactory might come to be accepted as tolerable – and actionable for legal, policy or sock-wearing purposes – under the influence of technological determinism (Bimber 1994).

Irrespective of its outcomes in any one instance, the process of 'governing' a sock drawer through data mining, along the lines just envisaged, exhibits some crucial differences as compared to 'conventional' governance described above. First, the responsiveness of data mining techniques to concerns emanating from different constituencies tends to be 'back-ended' or postponed to the stage of outcome evaluation (at least as far as unsupervised or semi-supervised data mining techniques are concerned). Conventional governance techniques encourage attention to, and debate around, procedure and participation from their earliest stages, because of those factors' prominence in prevailing narratives about the legitimacy of legal and political institutions: rule of law narratives, for instance. Legitimacy concerns in the data mining context seem, in contrast, to revolve mostly around the validity and scalability of results (eg Berkhin 2006: 17). There seems no comparable imperative in data mining governance to ask the sorts of early stage 'who' or 'in whose interest' questions that are routinely asked in conventional governance practice, at least in democratic settings.

Second, any revisitation of early-stage choices made in data mining – whether in the supervision of machine learning or otherwise – seems likely to be structured around field-specific considerations and options to a greater degree than in conventional governance practice. In the literature on *k*-means clustering, for example, 'cross-validation' tends to entail one or more of the following: comparing structures generated by the same algorithm (or the same combination of algorithms) under different parameters; comparing structures generated by different algorithms from the same data; or comparing one or more of those structures with so called 'ground truth' data, often obtained and represented through some other combination of data

collection and mining techniques (Jain 2010: 656-658). Consider, for example, the way the UN Global Pulse study compared the outputs of different data collection exercises, treating survey data as 'ground truth' data, without elaborating much on how the latter were collected or represented. Whereas conventional governance practice, since the late 19th century at least, has tended to encourage openness to quite robust and penetrating cross-disciplinary forays from a range of fields (exemplified by the 'Brandeis brief'), opportunities along these lines seem far more limited in the field of data mining (on the 'Brandeis brief', see Doro 1957). Describing data mining as 'interdisciplinary', one popular data mining textbook explained as follows the discipline's narrow sense of that term: 'Statistics, database technology, machine learning, pattern recognition, artificial intelligence, and visualization, all play a role' (Hand and others 2001: 4).

Third, the influence of taste, disposition, culture, style, faith, education, class, race, gender, sexuality and experience – and the recognition of contingencies and loyalties framed in one of these modes, or otherwise – seems more submerged, or dependent on representation-by-proxy, in the data mining context than in 'conventional' governance settings (on the difficulty of detecting mechanized reliance on proxies for race and gender, see Chan and Bennett Moses 2014: 672). Questions of identity and allegiance do not seem as likely to surface during sock sorting by data mining as they might in a 'conventional' dialogue around how socks should be sorted and which ones discarded. In the face of some data mining process judging a particular sock worthless due to its loss of elasticity, it might seem difficult – excessively emotional perhaps – to recall that one's grandmother knitted it, whereas conventional governance practice quite regularly invites human input more or less of this kind. In data mining (as in

some other modes of quantitative knowledge practice), contingencies and attachments tend to be transposed onto numeric attributes, weightings and randomization mechanisms, and worked through experimentally: by tweaking parameters and running the process again, to see what eventuates. Moreover, it is significant that neither subjects nor objects are necessary features of a data set for data mining purposes. A sock may be disaggregated and dispersed into any number of data points for purposes of relating it to another sock; no reassembly of those data points into something recognizable as a sock need ever occur for the directives yielded by that data mining to seem actionable. Similarly, it is the relationship of numbers representative of intervals in an iris image to those processed from another iris image (both deliberately anonymized) that authorizes someone to be ruled a 'recycler' – and both discredited and disentitled accordingly – in the UNHCR program. The data mining practice operating in both these settings need never engage a subject or object as such in order to yield an actionable predicate. As Louise Amoore has written, with respect to the dispensability of subjects, 'the digital alter ego becomes the de facto person' (Amoore 2009: 22; cf Clarke 1994).

In the same vein, whereas conventional governance has been accompanied by several centuries' worth of anxious reflection on 'bias' in decision making and attempts to mitigate human shortcomings in this regard, accounts of 'bias' in data mining literature seem to articulate quite poorly with this tradition. Barocas, Hood and Ziewitz have observed that '[t]here is a history of diagnosing "bias" in computer systems', but that key questions remain: 'what is it to say that an algorithm is biased—and how would we know? What counts as "unbiased"?' (Barocas and others 2013). Crawford has advocated approaching these questions agonistically (Crawford

2015). Yet data mining practice and literature seem to proceed in an entirely different, decidedly non-agonistic register when considering bias. One influential data mining textbook observed, for instance: 'different clustering algorithms will be biased toward finding different types of cluster structures (or "shapes") in the data, and it is not always easy to ascertain precisely what this bias is from the description of the clustering algorithm' (Hand and others 2001: 295). Despite this, a finding of usefulness tends to subsume and displace all other concerns: '[T]he validity of a cluster is often in the eye of the beholder… if a cluster produces an interesting scientific insight, we can judge it to be useful' and put it to work, irrespective of bias (Hand and others 2001: 295).

Fourth, and finally, questions of authority or jurisdiction seem more difficult to tackle, or even raise, in relation to data mining governance than in conventional governance practice. The UN's mining of data lawfully obtained from telephone companies in 'a country in central Africa' in the UN Global Pulse study, and the AIDR's labeling of data mined from Twitter, do not seem to call the relevant institutions' jurisdiction into question to the same degree as if those institutions had sought involvement in conventional governance practice in the countries in question. A claim that data mining entails some assumption or redistribution of power to 'give judgment' and pronounce the law (or pronounce something having law-like effects) for others seems alien and overblown when set against standard representations of the discipline (on jurisdiction understood in these terms, see Dorsett and McVeigh 2012: 4). Data mining tends, instead, to be cast diminutively as a practice of 'self-learning': 'Data mining is, in essence, a set of techniques that allows you to access data which is hidden in your database' (Adriaans and Zantinge 1996: 127).

3. Conclusion: Data Mining as a Matter of Concern

Approaching data mining as a practice of governance reveals that its operations are not altogether unlike those of some other, more familiar practices of global governance. Many conventional techniques of global law and policy are pattern creating and knowledge extracting. Think how defined terms, lists, and multi-part tests, of the sort regularly enshrined in legal instruments and policy directives, tend to create pathways for governance decision (on lists, for example, see Johns 2015).

Other techniques of governance typically expressed as institutions or entities might equally be understood as classifying, predictive or knowledge producing. States and corporations could be thought of as ordering devices: ways of drawing phenomena into and out of relation and generating maps, regularities and patterns not otherwise obvious. Yet we tend to think of these entities as much more than that; they tend to be anthropomorphized and treated as articles of faith or agents of reason. Though created by law and policy, these institutions are commonly understood to confer, distribute and embody authority, generate and dispense value, and evoke allegiance in ways that the former – names, lists and tests – are typically not.

Some legal and policy devices start off being thought of in the first, more diminutive, technical register (as lists are conceived), then migrate to the category of value-creating, power-producing, identity-defining institutions in which the public has a stake (as corporations are conceived). For example, think of how securitization practices came to be conceived after the 2007-2008 global financial crisis; no longer

are collateralized debt obligations, credit default swaps and other financial products thought of as benign instruments of concern to only a limited, savvy group (Swan 2009; Erturk and others 2013). Some institutions travel the opposite route, becoming more technique than authoritative entity over time. Consider, for instance, the way that a national stock exchange has changed from being a location considered pivotal to national and global economies – a place where people went to work, engaged in quirky rituals, and maintained, together, a significant institutional presence and identity – to being a moniker for an array of computer and telephone networks engaged in largely automated interaction around a common set of symbols and pricing metrics (Mitchie 2001: 596).

This chapter argues for a reclassification of data mining akin to that which securitization has lately undergone: from the 'merely' technical category – the concern of a highly specialized few – to the category of governance institutions and practices of global public concern. It does so, in part, because of the material and political significance of the decisions that data mining is called upon to support and guide globally, as made clear in Part 1: decisions about how to distribute limited aid resources; about how to prioritize and target anti-poverty measures and investigations; about how to locate, evaluate and address humanitarian need in emergencies. It does so, also, because of the way that data mining 'decision support' transforms experiences and possibilities of governance, as shown in Part 2. Data mining makes many governance-related tasks easier. In so doing, however, it makes some questions routinely raised in and around conventional governance practice harder to put forward, consider, or address. One need not claim privileged access to some underlying logic to recognize this (although privileged access may be necessary for

certain modes of action on this recognition); material shifts in global governance are taking place on the surface of data mining practice all around.

References

Adriaans P and Zantinge D, *Data Mining* (Addison-Wesley 1996)

Al-Raisi A and Al-Khouri A, 'Iris Recognition and the Challenge of Homeland and Border Control Security in UAE' (2008) 25 Telematics and Informatics 117

Amoore L, 'Lines of Sight: On the Visualization of Unknown Futures' (2009) 13 Citizenship Studies 17

Azzalini A and Scarpa B, *Data Analysis and Data Mining: An Introduction* (OUP 2012)

Barocas S, Hood S, and Ziewitz M, 'Governing Algorithms: A Provocation Piece' (29 March 2013) <http://dx.doi.org/10.2139/ssrn.2245322> accessed 15 September 2015

Bendoly E, 'Theory and Support for Process Frameworks of Knowledge Discovery and Data Mining from ERP Systems' (2003) 40 Information & Management 639

Berkhin P, 'A Survey of Clustering Data Mining Techniques', in Jacob Kogan, Charles Nicholas, Marc Teboulle (eds), *Grouping Multidimensional Data: Recent Advances in Clustering* (Springer 2006)

Best J and Gheciu A, *The Return of the Public in Global Governance* (Cambridge 2014)

Bimber B, 'Three Faces of Technological Determinism' in Merritt Roe Smith and Leo Marx (eds), *Does Technology Drive History? The Dilemma of Technological Determinism* (MIT Press 1994)

BioID Technologies, 'UNHCR Refugee Identification System' (no date) <http://www.bioidtech.co.uk/BioID/UNHCR.html> accessed 15 September 2015

Birnbaum B, DeRenzi B, Flaxman A and Lesh N, 'Automated Quality Control for Mobile Data Collection' in *Proceedings of the 2nd ACM Symposium on Computing for Development*, 11-12 March, 2012, Atlanta, Georgia (2012)

Boulesteix A, Janitza S, Hapfelmeier A, Van Steen K and Strobl C, 'Letter to the Editor: On the Term 'Interaction' and Related Phrases in the Literature on Random Forests' (2014) 16 Briefings in Bioinformatics 338

Bowyer K, Hollingsworth K and Flynn P, 'Image Understanding for Iris Biometrics: A Survey' (2008) 110 Computer Vision and Image Understanding 281

Bowyer K, Hollingsworth K and Flynn P, 'A Survey of Iris Biometrics Research: 2008–2010' in Mark Burge and Kevin Bowyer (eds), *Handbook of Iris Recognition* (Springer 2013)

Breiman L and Cutler A, 'Random Forests: Original Implementation' (no date) <http://www.stat.berkeley.edu/~breiman/RandomForests/> accessed 15 September 2015

Brownsword R, 'Code, Control, and Choice: Why East is East and West is West' (2005) 25 Legal Studies 1

Burge M and Bowyer K, *Handbook of Iris Recognition* (Springer 2013)

Cao W, Hu J, Xiao G and Wang S, 'Iris Recognition Algorithm Based on Point Covering of High-Dimensional Space and Neural Network' in Petra Perner and Atsushi Imiya (eds), *Machine Learning and Data Mining in Pattern Recognition* (Springer 2005)

Cate F, 'Government Data Mining: The Need for a Legal Framework' (2008) 43 Harvard Civil Rights-Civil Liberties Law Review 435

Chan J and Bennett Moses LK, 'Using Big Data for Legal and Law Enforcement Decisions: Testing the New Tools' (2014) 37 University of New South Wales Law Journal 643

Chen K, Chen H, Conway N, Hellerstein J and Parikh T, 'Usher: Improving Data Quality with Dynamic Forms' (2011) 23 IEEE Transactions on Knowledge and Data Engineering 1138

Clarke R, 'The Digital Persona and its Application to Data Surveillance' (1994) 10 The Information Society 77

Clements P and Northrop L, *Software Product Lines: Patterns and Practices* (Addison-Wesley 2002)

Colonna L, 'A Taxonomy and Classification of Data Mining' (2013) 16 SMU Science and Technology Law Review 309

Crawford K, 'Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics' (2015) 40 Science, Technology, & Human Values < http://sth.sagepub.com/content/early/2015/06/24/0162243915589635.full.pdf+html> accessed 15 September 2015

Daugman J, 'How Iris Recognition Works' (2004) 14 Circuits and Systems for Video Technology, IEEE Transactions 21

Davis K, Kingsbury B, and Merry S, 'Indicators as a Technology of Global Governance' (2012) 46 Law & Society Review 71

Decuyper A, Rutherford A, Wadhwa A, Bauer J, Krings G, Gutierrezn T, Blondel V and Luengo-Oroz M, 'Estimating Food Consumption and Poverty Indices with Mobile Phone Data' (2015) Computers and Society <http://arxiv.org/pdf/1412.2595.pdf> accessed 15 September 2015

Doro M, 'The Brandeis Brief' (1957) 11 Vanderbilt Law Review 783

Dorsett S and McVeigh S, *Jurisdiction* (Routledge 2012)

Eagle N, Macy M and Claxton R, 'Network Diversity and Economic Development' (2010) 328 Science 1029

Erturk I, Froud J, Johal S, Leaver A & Williams K, '(How) Do Devices Matter in Finance?' (2013) 6 Journal of Cultural Economy 336

French M and Mykhalovskiy E, 'Public Health Intelligence and the Detection of Potential Pandemics' (2013) 35 Sociology of Health and Illness 174

Fukuyama F, 'What is Governance?' (2013) 26 Governance 347

Goetz S, Baccini A, Laporte N, Johns T, Walker W, Kellndorfer J, Houghton R and Sun M, 'Mapping and Monitoring Carbon Stocks with Satellite Observations: A Comparison of Methods' (2009) 4 Carbon Balance 1

Hacking I, 'Michel Foucault's Immature Science' (1979) 13 Noûs 39

Han J and Kamber M, *Data Mining: Concepts and Techniques* (Morgan Kaufmann Publishers 2001)

Hand D, Mannila H and Smyth P, *Principles of Data Mining* (MIT Press 2001)

Hastie T, Tibshirani R and Friedman J, *Unsupervised Learning* (Springer 2009)

Imran M, Castillo C, Lucas J, Meier P and Vieweg S, 'AIDR: Artificial Intelligence for Disaster Relief' (23rd International World Wide Web Conference, Seoul, 7-11 April 2014) <http://dx.doi.org/10.1145/2567948.2577034> accessed 15 September 2015

— and Castillo C, Lucas J, Meier P and Vieweg S, 'AIDR: Artificial Intelligence for Disaster Relief' (Qatar Computing Research Institute, Doha, 20 May 2015) <http://www.slideshare.net/mimran15/artificial-intelligence-for-disaster-response> accessed 15 September 2015

International Federation of Red Cross and Red Crescent Societies (IFRC), *World Disasters Report: Focus on Technology and the Future of Humanitarian Technology* (IFRC 2013)

Jacobsen K, 'Experimentation in Humanitarian Locations: UNHCR and Biometric Registration of Afghan Refugees' (2015) 46 Security Dialogue 144

Jain A, 'Data Clustering: 50 years beyond K-means' (2010) 31 Pattern Recognition Letters 651

Johns F, 'Global Governance through the Pairing of List and Algorithm' (2015) 33 Environment and Planning D: Society and Space < http://epd.sagepub.com/content/early/2015/08/13/0263775815599307.full.pdf+html> accessed 15 September 2015

Kalka N, Zuo J, Schmid N and Cukic B, 'Image quality assessment for iris biometric' in *SPIE 6202: Biometric Technology for Human Identification III Proceedings* 6202:D1‑D11 (2006)

Kargupta H and Sivakumar K, 'Existential Pleasures of Distributed Data Mining' in Hillol Kargupta, Anupam Joshi, Krishnamoorthy Sivakumar and Yelena Yesha (eds), *Data Mining: Next Generation Challenges and Future Directions* (AAAI Press/MIT Press 2004)

Kingsbury B, Krisch N and Stewart RB, 'The Emergence of Global Administrative Law' 68 Law and Contemporary Problems 15

Kronenfeld D, 'Afghan Refugees in Pakistan: Not All Refugees, Not Always in Pakistan, Not Necessarily Afghan?' (2008) 21 Journal of Refugee Studies 43

Latour B, *Reassembling the Social: An Introduction to Actor-Network-Theory* (Oxford 2005)

Law J and Ruppert E, 'The Social Life of Methods: Devices' (2013) 6 Journal of Cultural Economy 229

Leber J, 'How Google's PageRank Quantifies Things' (Fast Company Co.Exist, 18 August 2014) <http://www.fastcoexist.com/3034193/how-googles-pagerank-quantifyies-things-like-historys-best-tennis-player-beyond-the-web> accessed 15 September 2015

Leskovec J, Rajaraman A and Ullman JD, *Mining of Massive Datasets* (Cambridge rev. ed. 2014)

Lessig L, 'The New Chicago School' (1998) 27 Journal of Legal Studies 661

Lye W, Chekima A, Fan L and Dargham J, 'Iris Recognition using Self-Organizing Neural Network' (Student Conference on Research and Development, SCOReD 2002, Shah Alam, July 2002) <http://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=1033084> accessed 15 September 2015

Lobel O, 'The Renew Deal: The Fall of Regulation and the Rise of Governance in Contemporary Legal Thought' (2004) 89 Minnesota Law Review 7

Meier P, 'New Information Technologies and their Impact on the Humanitarian Sector' (2011) 93 International Review of the Red Cross 1239

— 'Virtual Aid to Nepal: Using Artificial Intelligence in Disaster Relief', *Foreign Affairs* (New York, 1 June 2015)

Michie R, *The London Stock Exchange: A History* (OUP 2001)

Mohamad IB and Usman D, 'Standardization and its Effects on k-means Clustering Algorithm' (2013) 6 Research Journal of Applied Sciences, Engineering and Technology 3299

Mordini E and Massari S, 'Body, Biometrics and Identity' (2008) 22 Biosocieties 488

Nissan E, 'Legal Evidence and Advanced Computing Techniques for Combatting Crime: An Overview' (2013) 22 Information & Communications Technology Law 213

Nissenbaum H, *Privacy in Context: Technology, Policy and the Integrity of Social Life* (SUP 2010)

Pasquale F, *The Black Box Society* (Harvard 2015)

Pero R and Smith H, 'In the "Service" of Migrants: The Temporary Resident Biometrics Project and the Economization of Migrant Labor in Canada' (2014) 104 Annals of the Association of American Geographers 401

Pokhriyal N, Dong W and Govindaraju V, 'Virtual Networks and Poverty Analysis in Senegal' (2015) Computers and Society <arXiv:1506.03401> accessed 15 September 2015

Roiger R and Geatz M, *Data Mining: A Tutorial-Based Primer* (Pearson Education Inc. 2003)

Rubinstein I, Lee R and Schwartz P, 'Data Mining and Internet Profiling: Emerging Regulatory and Technological Approaches' (2008) 75 University of Chicago Law Review 261

Sarkar S, 'The Unique Identity (UID) Project, Biometrics and Re-Imagining Governance in India' (2014) 42 Oxford Development Studies 516

Schwartz P, 'Regulating Governmental Data Mining in the United States and Germany: Constitutional Courts, the State, and New Technology' (2011) 53 William and Mary Law Review 351

Sibai F, Hosani H, Naqbi R, Shanhani S and Shehhi S, 'Iris Recognition using Artificial Neural Networks' (2011) 38 Expert Systems with Applications: An International Journal 5940

Solove D, 'Data Mining and the Security-Liberty Debate' (2008) 75 University of Chicago Law Review 343

Su J and Dan S, 'Application of Data Mining in Construction of Corruption Risks Prevention System' (2014) 513 Applied Mechanics and Materials 2165

Suchman L, *Human-Machine Reconfigurations: Plans and Situated Actions* (Cambridge 2007)

Sumathi S and Sivanandam S, *Introduction to Data Mining and its Applications* (Springer 2006)

Swan P, 'The Political Economy of the Subprime Crisis: Why Subprime was so Attractive to its Creators' (2009) 25 European Journal of Political Economy 124

Taylor L and Schroeder R, 'Is Bigger Better? The Emergence of Big Data as a Tool for International Development Policy' (2014) GeoJournal <10.1007/s10708-014-9603-5> accessed 15 September 2015

United Nations (UN) Global Pulse, *Big Data for Development: A Primer* (UN 2013)

United Nations High Commissioner for Refugees (UNHCR), 'Afghan "Recyclers" under Scrutiny of New Technology' *UN News* (New York, 3 October 2002) <http://www.unhcr.org/3d9c57708.html> accessed 15 September 2015

— 'UNHCR gears up for 2003 Afghan repatriation' *UN News* (New York, 24 February 2003) (2003a) <http://www.unhcr.org/3e5a38924.html> accessed 15 September 2015

— 'Iris Testing Proves Successful' *UN Briefing Notes* (New York, 10 October 2003) (2003b) <http://www.unhcr.org/3f86a3ac1.html> accessed 15 September 2015

— 'Iris Testing of Returning Refugees Passes 200,000 Mark' *UN News* (New York, 10 October 2003) (2003c) <http://www.unhcr.org/3f86b4784.html> accessed 15 September 2015

United Nations (UN) Office for the Coordination of Humanitarian Affairs (OCHA), *Humanitarianism in the Network Age* (UN OCHA 2013) <http://www.unocha.org/hina> accessed 15 September 2015

Vatsa M, Singh R and Noore A, 'Improving Iris Recognition Performance using Segmentation, Quality Enhancement, Match Score Fusion, and Indexing' (2008) 38 Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions 1021

Vieweg S and Hodges A, 'Rethinking Context: Leveraging Human and Machine Computation in Disaster Response' (2014) 47 Computer 22

— and Castillo C and Imran M, 'Integrating Social Media Communications into the Rapid Assessment of Sudden Onset Disasters' (2014) 8851 Social Informatics: Lecture Notes in Computer Science 444

Wang Y, Tang J and Cao W, 'Grey Prediction Model-Based Food Security Early Warning Prediction' (2012) 2 Grey Systems: Theory and Application 13

Zarsky T, 'Government Data Mining and Its Alternatives' (2011) 116 Penn State Law Review 285