nature sustainability

REVIEW ARTICLE

https://doi.org/10.1038/s41893-018-0153-6

Social-media data for urban sustainability

Rositsa T. Ilieva[1,2]* and Timon McPhearson[1,3,4]

A voluminous and complex amount of information — 'big data' — from social media such as Twitter and Flickr is now ubiquitous and of increasing interest to researchers studying human behaviour in cities. Yet the value of social-media data (SMD) for urban-sustainability research is still poorly understood. Here, we discuss key opportunities and challenges for the use of SMD by sustainability scholars in the natural and social sciences as well as by practitioners making daily decisions about urban systems. Evidence suggests that the vast scale and near-real-time observation are unique advantages of SMD and that solutions to most SMD challenges already exist.

We live on an urban planet[1]. Human behaviour and values in cities are affecting, and may even drive, the future of global sustainability[2]. As societies have become more globalized, dynamic and transient, planning for the sustainable city of tomorrow has become an elusive, and some argue even futile[3] endeavour. Global urban science remains fragmented and disconnected from global and local policy[4,5] and planning, highlighting the need for new tools and data to advance understanding of complex urban dynamics, and to support decision-making for sustainability transformations. A key question then is what new heuristics and data sources can help us capture the growing complexity of social–ecological–technological interactions in cities[6] to help build a new global urban science[4,7] and provide new knowledge for improving decision-making towards more-sustainable urban futures[6,8].

In the era of information and communications technology (ICT), the Internet of Things (IoT), many types of big data, and ubiquitous technology at our fingertips, urban geolocated data from social media promises to expand our understanding not only of where people are and what they do, but also what they value (Fig. 1). Human behaviour and values are critical to align sustainability planning and policy with the needs and interests of residents[9]. Near-real-time dynamic observations on unprecedented scales — city, regional and global — are now within reach (Box 1). While research using data from social media such as Twitter, Instagram and Flickr has steadily grown in different scientific domains — such as digital humanities, urban ecology, epidemiology, tourism, disaster management and marketing — comprehensive accounts on the use of geolocated big data from social media for sustainable city planning are still rare. To our knowledge, to date no paper has reviewed the literature to synthesize the opportunities and challenges that SMD pose to sustainability research and provide examples of how scientists and practitioners are using SMD to advance sustainability goals. This is a consequential omission since more-liveable, healthy and prosperous human settlements in the twenty-first century cannot be attained with the planning tools of the twentieth-century city.

Here, we review evidence of SMD-based research carried out over the past decade, and discuss SMD as a critical component of big data (Fig. 2) and its applications for different spheres of sustainable urban development. We consider sustainable development to be grounded in the three-pronged approach suggested in the Brundtland Report (http://www.un-documents.net/wced-ocf.htm), but take into account more recent iterations put forward in the 2030 Agenda for Sustainable Development[10]. We identify distinct advantages of SMD over traditional research methods and explicitly consider how SMD can aid the pursuit of environmental, public health, social equity, infrastructural and economic goals in cities. Next to untapped opportunities, the analysis reveals crosscutting challenges to the usability and reliability of SMD in scientific research (Box 2). We demonstrate that although constraints such as missing demographics, exiguous locative data, volatility of social-media behaviour, and privacy concerns are pervasive, they are not unsurmountable (Table 1, see also Supplementary Information, section C). As SMD are becoming increasingly integrated in human life and urban governance, an overt appreciation of their promise and limitations has implications for sustainability research over the next decade and the implementation and monitoring of the global 2030 Agenda for Sustainable Development.

Emerging opportunities

Our review of over one hundred SMD-focused publications (Supplementary Information, section A) reveals that SMD hold promise for urban-sustainability research and planning across five interconnected spheres of action: environmental sustainability, public health, social equity, mobility and economic development. Spatial and semantic information from SMD have, so far, effectively allowed scientists to study park visitation and its determinants in New York City (United States)[11], predict floods in the city of York in the United Kingdom[12], detect depression across metropolitan areas in the United States[13], uncover potential unintended social outcomes of urban renewal in Johannesburg, South Africa[14], estimate travel demand analysis in Chicago, United States[15], and relate housing prices to specific points of interest in Shenzhen, China[16]. While SMD-based research is not inherently superior to traditional social-science research approaches, which propelled much of the twentieth-century city, it offers new heuristics and data sources for understanding human behaviour and values in cities. The sheer volume, limited cost, rapid collection times, granularity and range of participants that SMD afford are already revolutionizing sustainability research both in terms of research methods and the sort of questions that scientists can ask (Fig. 1). Here, we present recent examples of specific applications of SMD to the study of cities but also acknowledge that this is the first wave of urban SMD research: much work has still to be done to consolidate it into a reputable subfield of sustainability science (Box 3).

[1]Urban Systems Lab, The New School, New York, NY, USA. [2]CUNY Urban Food Policy Institute, City University of New York, Graduate School of Public Health and Health Policy, New York, NY, USA. [3]Cary Institute of Ecosystem Studies, Millbrook, NY, USA. [4]Stockholm Resilience Centre, Stockholm University, Stockholm, Sweden. *e-mail: ilievar@newschool.edu

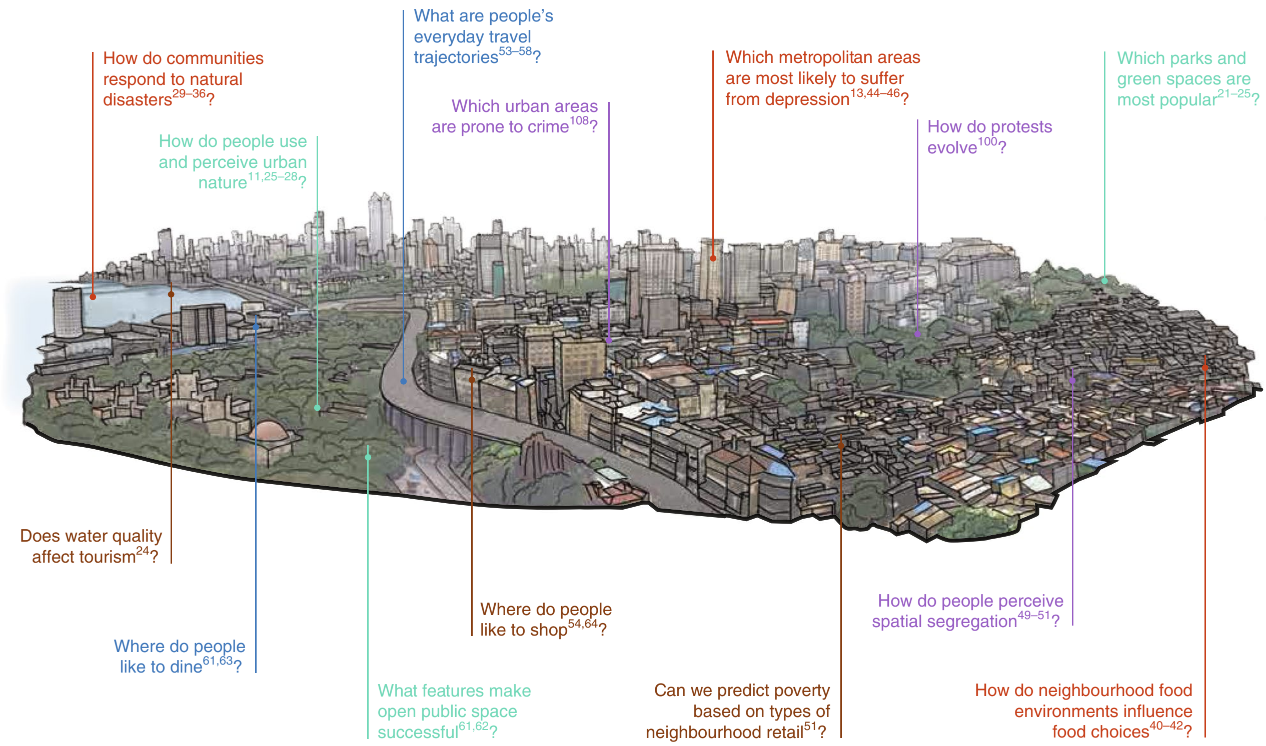

Fig. 1 | The wide range of emerging opportunities for urban-sustainability research provided by big data from social media. Evidence points to the promise of social-media data (SMD) for addressing key questions in five established domains of sustainability research: environmental sustainability (questions in green), public health (questions in orange), social equity (questions in violet), mobility (questions in blue), and economic development (questions in red). Large-scale, publicly available SMD on how people navigate, perceive, and respond to man-made and natural landscapes allows the investigation of human–environmental relationships in greater depth. SMD provide researchers and decision-makers with fresh insights into what makes open public spaces successful (refs [61,62]), what travel trajectories people pursue every day (refs [53–58]), how communities respond to natural disasters (refs [29–36]), which metropolitan areas are more prone to depression (refs [13,44–46]), and how neighbourhood environments influence food choices (refs [40–42]), among other consequential topics for sustainability planning discussed in the text. Image credit: Taylor Drake.

Environmental sustainability

Assessing the full value of regulating ecosystem services (for example, flood control or heat mitigation), provisioning ecosystem services (for example, food and water), and cultural ecosystem services (for example, aesthetic appreciation, spiritual enrichment and cognitive development) is key to environmental planning and sustainability policy[17,18]. Although progress has been made on assessing the value of regulating and provisioning services[19], the non-material benefits people derive from proximity to healthy ecosystems are yet to be adequately acknowledged in either research or policy[20]. Equally important are questions of what makes initiatives such as nature-based recreation or the development of new urban parks successful and how we can use this knowledge to inform future urban and landscape planning efforts.

Geolocated Flickr photographs have effectively been used to quantify nature-based tourism and recreation worldwide[21] and in the United States[22–24], and have been found to be a reliable proxy for visitation. Measuring park visitation and its underlying drivers in cities is even more critical given that few reliable monitoring mechanisms to gain this information exist to date. SMD have showed promise in filling this gap. Researchers have used crowdsourced data from Twitter and Flickr to develop rapid indicators for park access and visitation in New York City[11] and Minneapolis, Minnesota[25] and to study human use of and interaction with urban green spaces in Birmingham, United Kingdom[26]. SMD from Instagram have also enabled studies of how people use and perceive urban green spaces in Copenhagen, Denmark[27] as well as how they engage with different social activities in neighbourhood plazas and green spaces in New York City[28].

Compared to costly and time-consuming traditional observation methods, in which researchers need to be physically present and observe who visits urban parks and how and when visitors use different park spaces and facilities, or distribute lengthy surveys, SMD offer a convenient method for tele-observation and provide valuable knowledge for municipal green-space planning and the deployment of targeted investments citywide.

Climate change adaptation and disaster-risk reduction is yet another emerging field of SMD-based sustainability research[29] with important implications for pressing real-world challenges. Geolocated tweets have allowed scientists to model social activity during earthquakes[30], wild fires[31], rainstorms[32] and floods[33,34] as well as combine tweets with remote-sensing data to assess flood-related transportation infrastructure damage[35]. SMD can also usefully support spatial planning for flood evacuation based on people's preferences of specific shelter locations[33]. Using Twitter data on the socio-spatial interactions in New York during Hurricane Sandy in 2012 (ref. [36]), researchers identified what seems to be an important property of SMD — even a small subset of big data can throw light on larger spatial patterns of social activities.

Gauging public sentiment and perception of climate change is yet another key sphere of sustainability research in which SMD is helping social scientists to make progress. Sentiment analysis using big data from Sina Weibo — the alternative to Twitter in China — has revealed a possible disconnect between perception of climate change as a global issue and its local implications[37]. In the United States, Twitter-based SMD analysis showed that climate change skepticism is overshadowed by climate change awareness discussions

Box 1 | Social-media data and sustainability planning at a global scale

Implementing the 2030 Agenda for Sustainable Development requires planning for sustainability at a global scale. Building more-efficient, equitable and sustainable urban systems — the focus of SDG 11 — on a globalized and networked planet warrants data collection and analysis that traverse political boundaries and local jurisdictions. At present, an adequate data infrastructure that can support global sustainability planning is lacking. Evidence from recent research suggests that big data from social media can aid the development of large-scale social–ecological analyses. Over 500 million tweets and 1.7 million Flickr images are posted around the world every day by 310 million and 90 million active users respectively.

At the global scale, researchers have successfully used geotagged Flickr data to measure visitation of recreation areas outside cities[6,104] and identify popular urban areas of interest[105] with datasets ranging between 7.6 million and 197 million images and sample populations between 13,900 and 1 million participants. Geographic and textual data from Twitter aided analyses of urban mobility resilience[106], global mobility patterns[41], and geographical awareness[107] with datasets between 3.7 million and 944 million tweets and sample populations between 213,000 and 13 million. Social-media data (SMD) have thus enabled the analysis of hundreds of naturalistic areas in tandem and the simultaneous comparative analysis of thousands of cities. Considering that tweet location can be derived for over 20 million Twitter users[59], there is a largely untapped potential to use SMD for global urban sustainability research.

At the national and regional scales, researchers have used SMD from Flickr, Instagram and Panoramio to assess highly valued recreation locations in Europe and the United States[7,108] and post-disaster tourism recovery in the Philippines[109]. While not at global scales, datasets in these studies range between 71,000 and 78 million geotagged photos. SMD from Twitter and Sina Weibo (the parallel microblogging platform to Twitter in China) have been used to examine nationwide air pollution in China, Italy and the United States[110–112], healthy food access in the United States[24,26], attitudes toward climate change in China and the United States, pace of life in cities[113] and sentiments in shrinking and expanding urban areas in the United States[114]. Sample sizes in these studies range from 148,000 to 2.5 million tweets (or Sina Weibo posts).

and that climate activists have played a key role in making this sentiment dominant within the platform[38]. These analyses have also underscored that climate change can be perceived as a larger threat than a single event which triggers a social media discussion, and that some topics, such as Hurricane Sandy, can persist in social-media discussions long after the one-off event.

Public health

Evaluating how the interplay between social and physical environments affects public health in cities remains a core research challenge[39]. Geographers have already begun to use SMD from Twitter to map the prevalence of unhealthy food outlets in the United States[40], explore the extent to which food environments influence food choices[41], and predict obesity rates in the United States[42]. Analysing SMD within a single US state, medical professionals were also able to detect a strong positive relationship between alcohol-related tweets and alcohol-related emergency visits[43], pointing to the potential for real-time prediction (using SMD) of surges in emergency visits to hospitals. Additionally, in China, researchers have been able to infer air quality and particulate matter concentrations from microblogging activity and posts' semantic content, with the strongest correlations found for the megacity of Beijing.

SMD also appear to lend themselves well to analyses of psychological wellbeing in urban areas. For instance, public-health researchers have used data-mining and content-analysis techniques to examine the correlation between place and happiness, based on specific words found in tweets[44], and used GIS and advanced text-mining algorithms to uncover spatial clusters of Twitter users expressing feelings associated with clinical depression in the metropolitan area encompassing New York, Newark and Jersey City[13] and in metropolitan areas across the United States[13,45]. At a finer scale, geotagged Twitter data have also helped scientists to uncover composite geographies of happiness, diet and physical activity. In New York City, this revealed East Harlem and Chinatown as the census tracts with the lowest happiness scores[46], while in Brisbane, Australia, this technique enabled the study of physical activity patterns of mothers of primary-school-aged children[47].

Social equity

Urban inequality research has been useful but not sufficient in attempts to fully understand how deepening socio-economic and ethnic divides[48] are entrenched and socially reproduced through physical space and everyday life. The continuous stream of people's shared thoughts, trajectories and experiences in urban space make SMD a useful asset for social-equity research in cities.

GIS scientists have used geotagged Twitter data to re-examine the socio-spatial divide between east and west neighbourhoods in Louisville, United States[49]. Findings from the SMD-based spatial analysis revealed less-rigid physical boundaries between wealthy and disadvantaged population enclaves and challenged monolithic demarcations purported in local discourses. Geotagged tweets also proved useful for mapping the spatial distribution of different ethnic groups in London[50] and how this geography varies during day and night throughout the city. This newly uncovered spatial distribution of the city's ethnic communities added useful information to traditional, static data available from electoral registers. Also in London, SMD from Foursquare — a social-media platform that allows users to share their current location — have effectively been used to predict neighbourhood deprivation[51].

By offering an up-to-date glance over evolving social practices in inner cities and changing perceptions of danger and decay, SMD can help anticipate, and potentially plan for, creative-class-driven gentrification and displacement. In Johannesburg, researchers have used SMD from Instagram to reveal how suburban youth and creative-class individuals are increasingly accessing the city through neighbourhood exploration practices such as 'Instawalks'[14], driven by and driving the expansion of urban-redevelopment projects.

Urban mobility

Transportation planning is a major branch of city and regional planning and has a central role in the development of healthier and more-equitable urban regions. With the advent of low-cost global positioning system (GPS) navigators and ubiquitous mobile technology, real-time transportation planning is coming within reach. GIS scientists in Beijing[52] have, for instance, devised a method for the automated extraction of road maps from user-generated geographic information, combining GPS trajectories from special vehicles and geolocated Flickr data to derive road attributes. This approach promises a rapid, low-cost method to keep urban road-maps up-to-date and amend them in real time.

The surge of geolocated SMD over the past decade points to new possibilities to investigate human mobility. By combining crowdsourced content from Flickr, TripAdvisor and Uber, researchers have developed navigation tools suggesting customized route

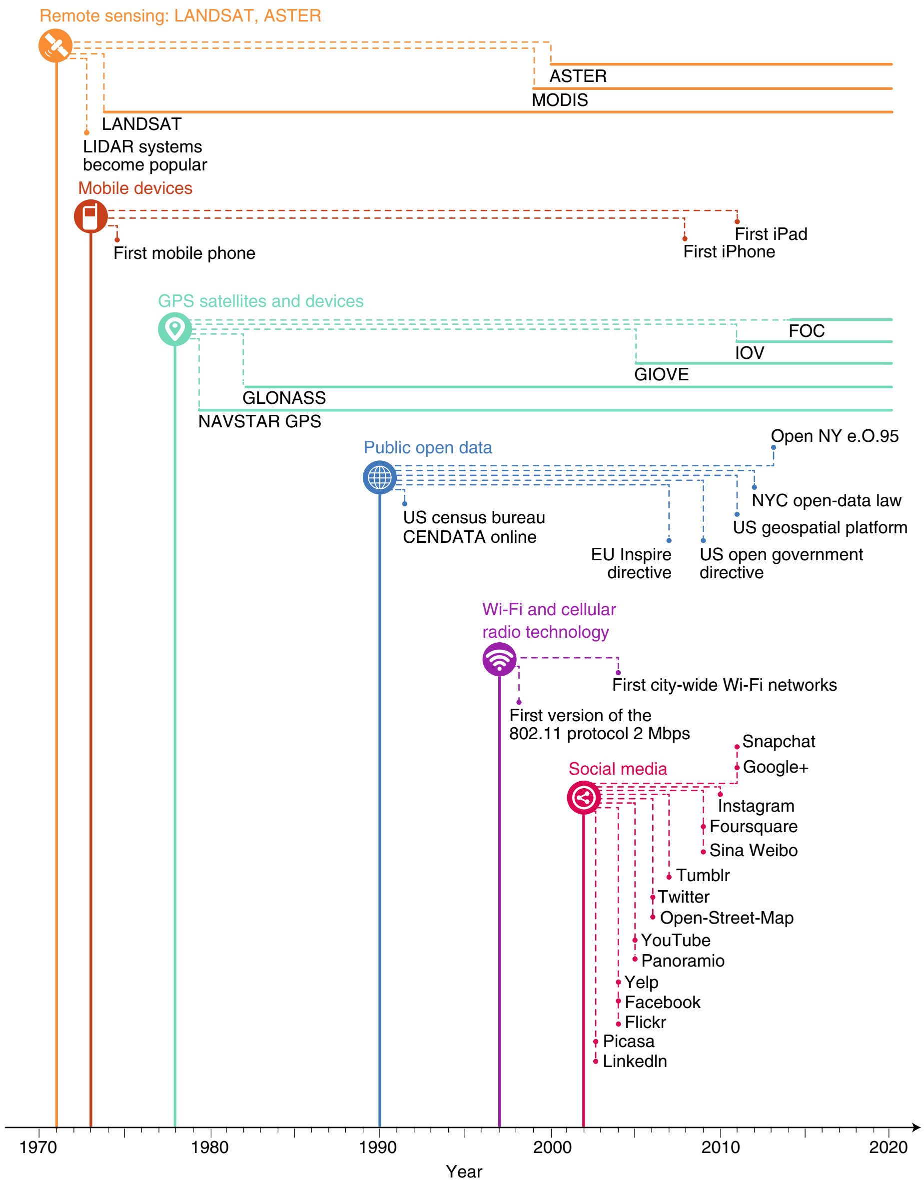

Fig. 2 | Evolution of key big-data sources and technologies, and the rise of social-media data. The evolving data landscape over the past few decades demonstrates the increasing availability of location-based social, infrastructural, and landscape and biophysical data. Data availability is key to tracking progress on the global SDGs and on local targets and goals in urban-sustainability plans, increasingly requiring high-resolution data both temporally and spatially. SMD represent a major new phase in our ability to understand links between human behaviour, values and preferences and infrastructural, climatological and other core components of urban, peri-urban and rural systems that are important for driving transformative sustainability changes. LANDSAT, Land Remote-Sensing Satellite (System); LIDAR, Light Detection and Ranging; ASTER, Advanced Spaceborne Thermal Emission and Reflection Radiometer; MODIS, Moderate Resolution Imaging Spectroradiometer; GPS, Global Positioning System; FOC, full operational capability satellites; IOV, In-Orbit Validation satellites; GIOVE, Galileo In-Orbit Validation Element; GLONASS, Global Navigation Satellite System; NAVSTAR, Navigation Satellite Time and Ranging.

sequences to travellers in Atlanta and Chicago (both in the United States)[53]. Geotagged Twitter posts have also allowed transportation analysts to develop an agent-based model of urban activity patterns (for example, home, work, eating, shopping and so on)[54] and derive the trajectories of individual social-media users, identifying their home and activity centers[55]. By analysing textual SMD from Twitter through data mining and machine learning, urban-planning scholars have shed light on how the way transit agencies communicate about public transit can reinforce stigma and negative perceptions and affect future infrastructure investments[56].

Furthermore, geolocated Flickr posts have been used to identify and predict the most popular routes to major points of interest by tourists in Florence (Italy), Glasgow (United Kingdom) and San Francisco (United states)[57]. This has led to a better understanding of how and why people choose different routes and transportation modalities to reach their destinations in cities, and thus has the potential to inform broader pedestrian- and public-space planning efforts. Using a large two-year dataset of geolocated Twitter activity in the Borough of Manhattan in New York City, computer scientists have effectively extracted crowd-mobility patterns and determined how these shape different functional regions at different hours during the day and over the week[58]. Based on first applications to the metropolitan areas of Pittsburg and Philadelphia (both in Pennsylvania), urban transportation researchers have also underscored how big data from Twitter can provide an effective low-cost tool to monitor traffic incidents on urban highways and arterial roads.

Box 2 | Emergent opportunities and challenges for sustainability research based on social-media data

1. Opportunities

Human-environment relationships. Social-media data (SMD) can help researchers to assess the value people place on nature and capture non-material benefits, such as aesthetic or spiritual enrichment, derived from proximity to healthy ecosystems. A combination of geolocated text and picture SMD enables measurement of green spaces' popularity and visitation drivers across city, regional and international scales. Analyses of SMD during extreme weather events and natural disasters can support the fine-tuning of climate resiliency planning and disaster-response models and interventions.

Public health. Geolocated SMD reveal how people feel, move and make choices in space. Content and location of microblogging activity can throw light on how the built environment affects healthy food choices and physical activity, as well as on where the highest incidence of depression and alcohol abuse among the urban population is.

Social equity. The granularity of SMD enables the study of how social inequalities and uneven access to urban amenities play out in space and time and change over the course of the day and night in a city. SMD have proved useful in uncovering how boundaries of socio-spatial segregation are socially constructed and why the spatial contours of socioeconomic and ethnic divides may be less fixed than is commonly thought.

Mobility. SMD with location information offer new insights into pedestrian mobility, allowing the detection of people's travel trajectories as well as the path most travelled between two destinations. International SMD-generated travel trajectories can reveal users' home locations, helping distinguish social-media users who are locals from those who are tourists. Geolocated SMD can also support the real-time production, revision and update of city road maps.

Economic development. SMD can be used to track consumption practices (for example, travel, fitness, food, entertainment and so on) during the day and week and detect recurrent patterns of commercial activity. SMD can assist researchers in mapping regional shopping flows to inform future location of retail centres. Geolocated SMD can also support investigations of built environment and economic determinants of tourism and the identification of popular urban attractions, public plazas and restaurants.

2. Challenges

Population. Participation in social media continues to grow but is not universal. The makeup of microblogging communities greatly varies across demographics (for example, age, gender, ethnicity, nationality), resulting in unrepresentative SMD sample populations. Hyperactive bloggers overshadowing the voice of less active ones, unaccounted share of posts from nonhuman accounts, or unknown link between number of accounts and real-world individuals can skew SMD samples.

Location. The proportion of SMD with associated geographic coordinates is still small, while manually added locations can be erroneous or ambiguous. SMD is, thus, not a straightforward proxy for human activities in space and requires corroboration by other sources. Inferring places' importance from the availability or density of geolocated content is also problematic. Not all places and activities are suited for microblogging and the interests of habitual visitors can greatly differ from those of newcomers.

Sentiments. Use of sarcasm, metaphors and colloquial language can lead to misinterpretation of SMD content. The extraction of meaning from textual SMD is challenging due to possible mismatches between people's real and broadcasted opinions and actions. Statements classified as 'neutral' make up the majority of SMD, which hinders the extrapolation of positive or negative sentiments.

Time. The time-frame within which SMD are retrieved influences both data content and data quantity. Some social trends are ephemeral and related to one-off events, while others are tied to weekly or seasonal activities. Trends of sentiments and opinions can also be volatile and diverge from one sampling period to another.

Ethics. Threats to privacy in SMD research can arise when collating user metadata from multiple sites, using the support of data analysts unfamiliar with ethical norms in human-subjects research, or sharing un-anonymized data with third parties. Selective access to SMD can affect who does SMD research and the type of questions being asked.

Data analysis. As other big-data sources, SMD research can be affected by patternicity, or the detection of false patterns. Corroboration of findings can be difficult since the same dataset is rarely accessible to different independent research teams. The massive, networked, unstructured, real-time nature of SMD also poses computational challenges.

At the global scale, Twitter SMD have been effectively deployed to map international mobility flows between countries[59], and Flickr SMD have been used to capture mobility flows between global cities[60], allowing researchers to distinguish between residents and local and international tourists. This is important because inferring local societal needs and mobility patterns based on SMD, without considering the sociodemographic characteristics and provenance of SMD users, can lead to misplaced needs assessments and interventions in global cities such as New York and London.

Economic development

In many cities, especially large globally connected urban hubs, tourism makes up an important part of the local economy and thus mayors are keen on understanding which cultural programmes and activities attract the most visitors. SMD can be a promising source for developing decision-support tools to this end. Foursquare data has, for example, allowed researchers to identify the nineteen most successful urban plazas in eight cities in Alicante, Spain[61] and related determinants, such as proximity to historical centres and major transportation axes. Twitter data have showed promise in informing urban smart tourism in the United States. In San Francisco, in the United States, semantic and geographic analysis of tweets and their visual attachments has allowed assessment of where tourists are in the city and what attractions and events they find appealing and talk most about[62].

SMD have proved useful for examining commercial activity in cities by enabling the tracking of chains of everyday consumption activities (for example, travel, fitness, food, cinema and so on)[54] as well as use of specific subsets of commercial establishments, such as restaurants. Researchers in China, for example, used GIS and geolocated crowdsourced data from a consumer review website (dianping.com) to identify popular restaurant destinations in the city of Hangzhou[63]. Interestingly, most popular restaurants were found in mixed-use, easily accessible urban neighbourhoods. This kind of research is meaningful not just for tourism purposes but also for optimizing catering logistics, reducing unnecessary food waste, and informing overall strategies of urban economic development.

Table 1 | SMD crosscutting challenges and possible strategies to address them

Challenge	Possible strategies
Representation of population	Use multiple sources of data Correct platform-specific and proxy population biases Derive metadata from profile descriptions Use demographic collators Develop models able to isolate the most-active users contributing to a specific activity pattern Use facial recognition algorithms Use filters or algorithms for detecting non-human accounts Adopt new statistical methods for generalization from unrepresentative samples Use local-authority censuses and total empirical annual visitor user-days for parks visitation
Representation of activities	Activity pattern-recognition model to predict missing activities Rich record on both sayings and doings Use of datasets from multiple social-network sites Make explicit the limitations of each sample used Check for correspondence between surveyed and SMD-derived visits Untangle psychosocial from platform-driven behaviour
Representation of space	Distinguish different cities with the same name and dissimilar semantics Individually analyse images with upload location errors Produce results for the whole city and for individual neighbourhoods Collapse all single-user photos taken within a given radius to a single arithmetically centred point Consider new units of analysis, estimating the visitation probability of a place Use images, textual metadata and full-text geocoders Use the image user's traces on other social-media platforms Use social-network analysis, considering the density of retweets Borrow search techniques from artifical intelligence, information retrieval and natural-language processing
Representation of sentiments	Use statistical cluster analysis Use an opinion-bias detection model Analyse words out of context Collect contextual information about users' choices, goals and activities Include participatory research methods, mixed-method research design and action research Use social-network analysis for level and frequency of connections between individuals
Temporality	Develop new tools to measure and monitor the resilience of social and geographic ties Treat Social Media Data (SMD) as an imperfect continuous panel survey Ensure consistent availability and update of large datasets
Ethics and privacy	Anonymize data and ensure that no interaction with the individuals in the sample takes place Use only geo-tagged and time-stamped information without accessing personal details Assess how SMD research complies with Institutional Review Board (IRB) guidelines or other country- or discipline-specific codes of ethics Secure data and resolve ownership and Internet protocol (IP) addresses issues Host data-download applications on a restricted access website Use more-integrated analytical tools with privacy protection Re-conceptualize privacy
Data access and data quality	Integrate qualitative studies of cultures of large-scale, quantitative data Identify specific aspects of the behaviour of proprietary systems to begin reporting biases Quantify biases, showing results for more than one platform and for time-separated datasets from the same platform Combine user-generated with other datasets Use metadata to assess data quality
Data analysis	Make explicit the limits of possible questions and interpretations Ensure transparent research methods Use metadata management to evaluate the quality and trustworthiness of SMD Set uniform units of analysis and data formats across institutions and software Use a system integrating heterogeneous social-media feeds Build multidisciplinary research teams combining expertise in geography, computational social sciences, linguistics and computer science Deploy automatic semantic analysis to support manual coding Have independent raters score multiple records, across various categories and data sources Use distributed file-storage systems and NoSQL (Not only Structured Query Language) databases to overcome storage challenges Use parallel computing and data indexing for large-scale spatial queries Use machine learning and massive social-network analysis to reduce SMD complexity
Validation	Use mixed-methods approaches including ethnographic, statistical and computational Validate land-use and social-dynamics clustering models with personal interviews and surveys Validate land-use maps derived from geotagged tweets by using real land-use data provided by city planning departments, and test them in different cities Replicate studies and try to falsify and test them in different cities Make failed studies visible Compare results obtained using new and existing research methods (on the same data set) Use two or more distinct data sets to corroborate findings on new social phenomena Create a framework for sharing, citing, and reusing big data

Source: The authors, based on review of current literature on the topics (for a detailed analysis of each strategy please refer to the Supplementary Information). Some strategies can be deployed to address multiple challenges.

Box 3 | Critical insights for future research

Emergent research using social-media data (SMD) is charting a new terrain in urban sustainability science. Our review of the first wave of SMD explorations, however, suggests that this field has yet to reach the degree of maturity necessary to noticeably influence the outcomes of sustainable-development actions. We offer five critical insights for advancing applied SMD scholarship:

- Data versus outcomes. The bulk of existing SMD research is focused on describing behaviour and human–environmental interactions rather than changing them. To more squarely put SMD in service of the global 2030 Sustainable Development Agenda and its implementation, the next phase of SMD studies should explore new kinds of research designs aimed at interventions and evaluations of SMD-aided sustainability planning practices.
- Research design. Few studies combine multiple data-collection methods to corroborate SMD-derived findings. While SMD allows for generalizations at an unprecedented scale, explanatory sequential mixed-method research designs — whereby quantitative-data collection is followed by qualitative-data collection and analysis — would help SMD researchers to test assumptions about how urban systems work and behave in a more robust fashion. This will allow researchers to identify entry points for sustainability transitions with improved confidence.
- Impact on planning practices. Existing SMD research rarely gauges the impact of SMD-derived knowledge on actual urban policies and plans. This makes it difficult to appreciate the value that SMD adds to city and regional planning efforts. The next phase of SMD studies should critically explore and compare multiple instances of SMD adoption in real-world city planning and local resilience action planning processes.
- Focus on interactions. While most SMD publications exhibit elements of interdisciplinary science, few use a systemic perspective as an organizing principle of research. Current SMD research helps to amplify and advance knowledge on specific topics, but misses opportunities to use SMD to expose the interactions between different, and often competing, sustainable development goals in the context of complex urban systems. Future SMD works can enter this space by building on recent scholarship on the interplay between the 17 SDGs constituting the 2030 Agenda for Sustainable Development.
- Images versus tools. Scholars experimenting with SMD in different domains of sustainability research have yielded stunning new representations of human activity in space and probed new ways of mapping complex social webs. Although images have the power to unsettle and reorient collective attitudes or beliefs, they fall short on providing equitable long-term structures for planning and decision-making. What new planning, design, and decision-support tools can be developed using SMD and how SMD can be woven into existing decision-support systems (for example, Hazus in the United States) are questions SMD researchers can already address.

At the regional scale, geolocated tweets have proved a reliable source of data to study shopping flows in the city of Leeds, United Kingdom[64]. SMD results were validated through consumer survies and geolocated data from a mobile-phone service provider, and issues of data sparsity were addressed by allowing for multiple trips per SMD user. Among other applications, this kind of SMD research can help local administrations plan for future retail centres and their related transportation networks. In Shenzhen, China, SMD have showed promise in aiding housing-market research. Analysis of Sina Weibo posts enabled prediction of housing prices based on the type and number of points of interest (for example, green spaces, public transit stations, hospitals and business facilities) located nearby.

Crosscutting challenges

SMD show clear promise for advancing urban-sustainability research and planning, but it is important to acknowledge limitations of this source of geographic data as well. Based on our comprehensive literature review (see Supplementary Information for methods and further details on results), these limitations include unequal representation of real population, inaccurate geolocation, mismatches between stated opinions and real-world behaviour, linguistic ambiguities, temporality of phenomena broadcasted through social media, privacy protection, research ethics, data access, data interoperability, and cultural resistance to the use of SMD in social research, among other limitations. Here we discuss each of these challenges separately and highlight possible solutions.

Sample population. Equating numbers of social-media accounts with numbers of people is one common pitfall that scholars using big data from social media face in their research[65]. The issue is problematic because a single user can have more than one account, and a single account can be used by more than one individual. In addition, not all social-media accounts are equally active. For instance, inactive Twitter accounts represent about 47% of all registered users, with 330 million monthly active users in 2017 (ref. [66]). Potential influence of automated accounts known as botnets is yet another source of concern since some datasets might contain large volumes of non-human posts such as advertisements or election-campaign retweets[67,68]. Reliable techniques that can help researchers to isolate user-generated data[69] from other content, and assess their credibility[70], are needed in future SMD research.

Knowing who produces SMD is essential if SMD are to advance sustainability research, yet social-media accounts are often devoid of basic demographic data such as age, ethnicity, gender, education, residency, income and occupation. Software known as demographic collators offer one way to retrieve these data by combining information from multiple sites and accounts for a given user[65,71]. This approach, however, can lead to privacy concerns and the retrieval of sensitive information that was not intended for public dissemination[72].

The usability of SMD as a reliable proxy for the general population also presents a challenge for SMD-based research. For instance, Instagram and Pinterest users are mostly young women, with Instagram appealing more to urban, African-American and Latina users[73]. Similarly, the most-active Twitter accounts tend to belong to young, African-American, and urban or suburban residents[74]. The emerging field of signature social science[71] and user profiling allows the weighting of social-media content in relation to different demographic parameters and this is one approach to reducing bias[70] in SMD samples. Validation through non-SMD records, when available, has likewise proven effective when testing SMD as a proxy for visitation rates of public parks in the United States[22] and worldwide[75]. In addition, combining multiple sources of SMD[76], as well as SMD with other sources of big data[77], can help overcome the bias that affects each micro-blogging platform individually.

More broadly, SMD population bias is related to access to and use of the Internet, which is negatively correlated with age and poverty[78]. With the increase in mobile-device ownership and its decoupling from socio-economic status[51], disparities in Internet access have, however, greatly diminished. In the United States, Internet users now represent over 88% of the population[79], and people who use at least one social-media platform make up more than 69% of all Internet users[80]. In less-industrialized nations, big data from microblogging are still at the early stages of proliferation. In some countries, there are parallel local platforms, such as Sina Weibo, which replace international ones and limit opportunities for global SMD research.

Locating human behaviour

Part of the allure of big data, including that of SMD, is the prospect of asking a wider range of spatially explicit sustainability questions at an unprecedented scale. Most social media, in fact, allow one to either manually select a place label from where one posts a message or a picture, or have it automatically added through geolocation services via GPS, cell towers, Wi-Fi points or the device's internet protocol (IP) address.

Microblogging posts with latitude and longitude coordinates still represent a small fraction of the total volume of SMD; 1–3% of all tweets[67,81] and approximately 3% of all Flickr images[82] are geolocated. Paucity of geolocated SMD could be due to the fact that people are able or willing to share only a limited subset of locations[83] and that location services are turned off by default on most social-media applications. Emerging approaches to increase the number of geolocated texts and pictures include automated text-based geolocalization that extrapolates location from toponyms appearing in tweets or picture tags[84], user-profile information[85], user uploading and tagging patterns[86], friends' locations[87], time zones[88] or a combination of multiple geo-indicators[89]. So far, most methods have proven effective for low-resolution analyses, accurate to the city or state level, with localization accuracies ranging between 30 km (ref. [89]) and 160 km (ref. [90]). Some recent studies have been able to achieve neighbourhood-scale granularity[91] and resolutions as high as 145 metres (ref. [92]).

Besides possible content-ambiguity challenges with user-input locations, acquired coordinates can also be inaccurate. This may be due to a voluntary or involuntary interruption of connectivity at the time a post is shared. Such mismatches may affect the validity of SMD-based spatial analyses. One strategy to circumvent inaccurate locations is to set a buffer accounting for nearby upload locations rather than use only precise locations[27]. One can also combine information from multiple platforms by, for instance, using textual information from tweets to estimate geolocation of Flickr images[93].

Gaps between real and stated locations can be captured by, for instance, comparing user-selected labels with GPS data or predicted locations. These gaps can be approached not as an inaccuracy to rectify but as an important source of cognitive and behavioural information. Locations of tags and user-provided metadata can thus be viewed not so much in terms of precise locations but, rather, as forms of different expressions of one's 'spatial self' through social media[94].

As discussed previously, geolocated SMD offer a reliable proxy for real-world visitation for some tourist and recreation destinations around the world. As such, researchers may be tempted to make inferences about the popularity of certain points of interest within a larger area based on SMD points density. When considering SMD pictures, for instance, this assumption can be misguiding for at least three reasons: there may be large volumes of pictures taken by a single photographer[75], some activities are less suited to the taking of pictures[21] and familiarity with a place or the distance travelled to visit it influences one's propensity to photograph it[21,75]. Although researchers have no control over the place of origin of SMD users, the single-photographer (and by extension single-tweeter) challenge can be dealt with through GIS, by collapsing the photos taken by a single Flickr user within a given radius to a single arithmetically centred point[75]. Researchers can also set thresholds to exclude the least-active and most-active tweeters from their data set at the outset of the analysis[64], to reduce representation bias.

Mapping using demographically unrepresentative data can also reproduce spatial segregation and provide an unfair picture of the places that matter citywide[83]. This holds true both for local- and global-scale analyses. New York City is one of the top cities in the world in terms of number of photographs uploaded on Flickr per year — over 1 million — however, even if social-media users in Chicago uploaded less than half of the photographs that New York City did, the city scored higher in terms of Flickr photograph uploads per capita[60]. The volume of geocoded tweets still greatly differs across nations worldwide. The United States, Brazil and Indonesia are the top three countries in terms of volume of geolocated tweets produced, whereas Brazil, Malaysia and Kuwait are the top three countries in terms of geolocated tweets produced as proportion of the total population of Internet users in each country. Conversely, Denmark, Norway and China (because of censorship laws) exhibit significantly lower volumes of total geolocated tweets produced alongside low penetration of Twitter among Internet users[95].

Overall, although efforts to assess the pitfalls of using SMD in geographic research are still rare, to improve current methods for geographic information retrieval, SMD research would benefit from borrowing search techniques from artificial intelligence, information-retrieval science and natural-language processing[96].

Sensing of sentiments and perceptions

How people make sense of and value different places, infrastructures and events in their everyday lives has profound implications for sustainability planning and implementation efforts. But people's stated opinions can greatly differ from their actual thoughts and actions[97,98]. This is particularly critical to acknowledge when using sentiment analysis to examine the drivers behind certain social behaviours such as the dominant preferences for specific venues such as parks and recreational areas[24,75], museums[99], restaurants[51] or shopping centers[64] as well as tracking the dynamics of the escalation of protests[100] and public response to natural disasters[36] in cities.

Interpreting opinions can be extremely difficult due to the biases of social desirability — altering one's statements and publicly stated positions to gain broader social acceptance — or ethical ideology, which means altering one's true position to appear as a person of higher ethical standards[101]. What people say can slightly or substantially differ from what they do, and self-imposed filters can cause people to say only what is commonly considered appropriate and safe[97].

People can also 'cheat' in SMD through the diffusion of misinformation or the production of high volumes of tweets and retweets on purpose[70]. Researching the relevance of a given topic or place can further be compromised by the invisibility of censored tweets, muted through the filtration systems that social media have in place. The ability of SMD to reveal what drives public behaviour is also limited by the lack of contextual information of the broader set of choices, goals, and activities of social-media users, and the assumption of universal laws behind sentiments.

Despite these caveats, tweets, pictures and their tags offer a novel entry point for uncovering public interests and perceptions of space, place and human–nature interactions. Scholars in different fields have begun exploring their potential — Flickr tags have been probed as a means for studying perception of borders[102], the modelling of loosely demarcated regions such as 'Lower East Side' or the 'Alps'[103], or the perception of urban landmarks such as the Golden Bridge in San Francisco and the High Line of New York City[75].

Extrapolating and codifying information through tags, however, can be particularly challenging due to linguistic idiosyncrasies, diversity of language[104] and the use of metaphors and sarcasm[40]. Ambiguity can compromise the effectiveness of research techniques such as sentiment analysis, which still lack sound validation procedures. The majority of posts tend to be scored as 'neutral', which further hinders analytical efforts[105].

To classify vague or ambivalent words, researchers can consider and classify the context within which these are placed[45] or perform correlations through automatically derived dictionaries through unsupervised learning; these capture unconventional words that may not be part of dictionaries due to incorrect spelling[106]. The uneasy issue of sentiment scoring and polarity detection (positive–negative posts) can be addressed through the combination of multiple sources of information[107,108]. For instance, emojis can be used to train datasets for text-based analyses or established lexicons can be supplemented with other sentiments scales such as the Happiness Index based on the Affective Norms for English Words (ANEW) study[109]. Each approach to sentiment analysis has unique advantages and limitations and thus researchers need to be aware of their variable suitability in relation to the question asked and the specific data[106].

Visual-sentiment analysis remains one of the most challenging SMD analysis techniques. Analytical strategies include focusing on the mid-level attributes of a picture[110] and combining visual and textual information (for example, tags and descriptions) in supervised[111] or unsupervised[112] techniques.

Temporality of content

The temporality of social-media content (for example, meanings, opinions and trends) is yet another potential source of bias in SMD research. This is especially the case when studying events that are seasonal in nature (for example, summer festivals, farmers' markets, and visitation and use of natural areas) or related to a one-off episode (for example, strikes, protests, hurricanes or earthquakes), all of which unfold within a limited time frame. In addition, social norms and opinions can also vary relatively quickly and so can the regulatory and filtering practices of social-media companies.

Time is a critical variable affecting the relation between space and the meanings attached to it. The specific time of the week, or even the time of day, when data are gathered can affect both the intensity and typologies of activities observed in a given place[113]. The choice of the sampling period thus simultaneously influences data quantity and data content[40] and this must be acknowledged when considering SMD as a basis for long-term planning strategies or policy interventions.

Scholars studying the life of open spaces in cities and naturalistic areas thus ought to consider that the amount and nature of SMD available in those spaces can substantially vary with changes in weather, seasons, and special programmes and events. Given that one of the important added values of SMD is that they offer near-real-time observation of social phenomena, a key task for SMD-based sustainability research is to ensure the consistent availability and update of large datasets[114] through which researchers could corroborate or refute preliminary findings worldwide.

Obsolescence of data, however, is a challenge for traditional data-collection methods as well, and thoughts and intentions stated in surveys, focus groups, or interviews can significantly differ from future thoughts and behaviour[115]. Distinct advantages of SMD in terms of temporality are the relative ease of devising longitudinal studies, overcoming some of the limitations of traditional cross-sectional research, and informing an adaptive style of environmental management — one cognizant of shifts in social behaviour following specific public policies or interventions.

Ethical dilemmas

The risk of violating social-media users' privacy[116] is arguably among the most debated ethical challenges in the field. Threats to privacy could occur either directly — for example, through user-generated geolocated data[94] — or indirectly, through the coupling of multiple data sources revealing information that users had no intention of disclosing[72] or sharing data with third parties in collaborative research projects. Social-media users do not grant consent for the extraction and circulation of the data they author in any formal way, nor do they picture researchers as the intended audience of the content they share with relatives, friends and colleagues[94]. Devising effective restrictions and practices that could prevent unethical use of spatial and spatially embedded network data is, therefore, key[117]. A useful starting point is the Code of Ethics for GIS professionals[118], which provides specific indications on privacy protection and undue intrusion into the lives of individuals, including through the combination of multiple databases.

Ethical challenges stem also from the uncertainty of who is (or will be) excluded from accessing SMD. The fact that SMD providers are primarily large private corporations is troublesome for at least three reasons. First, because profit maximization constitutes a core principle for those businesses[99], price conditions can change and make data inaccessible for many researchers and public and not-for-profit organizations. Second, researchers can be excluded altogether if companies restrict data access to company insiders or private businesses only. And, third, restrictions on the sharing of data with other research groups can preclude the validation of findings through replication, and undermine the reliability of SMD research at large[69,73].

SMD access is consequential for equity in SMD-based sustainability research since the scope and nature of sustainability questions is determined by who is asking them[65]. Deploying action research to bridge the digital gap between private and not-for-profit community and public organizations is one promising strategy for addressing this structural challenge[119].

At the macro level, the uncertain social–ecological impacts of SMD research are an additional source of ethical concerns. The physical infrastructures and energy needed to sustain ever greater volumes of big data from social media and related research will be significant. Yet, by providing a finer-grained understanding of social and environmental change, SMD can also lead to more effectively planned and coordinated human activities, thus enabling us to better preserve and restore the environment[117].

Data analysis

Aggregating large volumes of data, such as user-generated data from social networks, could provide new insights into questions previously impossible to examine at larger scales, but it can also be misleading. Scholars should be aware that SMD research can suffer from a 'majority' bias — missing or misinterpreting behaviours of minority groups[83], discussed at the beginning of this section, or from 'apophenia', the identification of patterns where no patterns exist[21].

Additional analytical challenges include the lack of validation studies of SMD research — partly due to a diffused fear-based resistance to the use of SMD in the social sciences, the unverified quality of do-it-yourself research tools and designs, and the uncertainty of whether data analysts are professionals or amateurs unfamiliar with ethics codes for human-subject research[105].

Further, intelligible units of analysis[69] and interdisciplinary approaches[120] are needed to ensure interoperability between SMD research and the work of public and private organizations advancing sustainability planning in cities. Future SMD research requires that researchers from different disciplines — computer scientists, computational social scientists, linguists, geographers, architects, urban planners and urban ecologists, among others — address SMD challenges in a transdisciplinary fashion[30].

Finally, in addition to validation, big data from social media still present a few computational challenges that scientists will have to address. New data storage, transmission and analytics approaches are needed to tame the massive streams of largely unstructured, networked, and real-time SMD. Cloud computing, and its related functionalities, such as distributed file-storage systems (for example, Hadoop, Google Drive and Dropbox), NoSQL databases, parallel computing (for example, MapReduce), data indexing for large-scale spatial queries (for example, SpatialHadoop and B-tree), machine learning and visual analytics, offer emerging solutions to some of (spatial) SMD's pervasive computational challenges[121]. Complexity of social networks can become less taxing in terms of software and hardware requirements through deploying new tools for massive social-network analysis[122] and models for spatio-temporal SMD analysis[123].

Future directions

Our Review shows that experts in the emerging field of SMD research are in accord. To cope with current limitations of SMD, advances in methodology are needed. Reviewing existing research techniques and devising new ones is of critical importance, while also paying attention not to overestimate what can actually be done with SMD. Yet, some scholars contend that it is not only a matter of new methods, but also of new theories. To harness the many opportunities that the use of SMD in geography and spatial sciences opens up, for instance, there is a need to overcome the inherent limits of our cognitive capabilities through the development of new network-based ontologies[117]. A greater public discussion of SMD research and its challenges is also warranted to ensure that SMD research effectively serves the public good[119].

However, the opportunity to take advantage of new and massive data streams on human behaviour and perceptions is exciting. Ancillary to the growing volumes of publicly available urban data, SMD can offer city planners and engineers around the globe the opportunity to fill topical gaps and overcome the time lags affecting traditional census sources. Needs assessments and future-oriented planning can use SMD as a starting point for established forms of community engagement and participatory GIS science[124,125]. Because the pursuit of economic, environmental and social goals entails multiple trade-offs and compromises[126], cities can leverage SMD tools to facilitate discussion about how weights and values are to be assigned to different aspects of city life (for example, affordable housing versus open space). Additionally, if SMD are to help cities to build a much-needed global urban science[8], SMD-literate public officials and new spaces for big-data governance will have to be established.

There is much work remaining to overcome the existing challenges that face SMD and make the sector more amenable to the wide variety of research needs. Our Review suggests that the use of SMD for sustainability research is still confined to a subset of sustainability domains. Food insecurity, clean energy, quality education and gender equality are among the spheres where the contribution of SMD is yet to be fully tested. Culture, the fourth pillar of sustainability[127], has also received scant attention by SMD researchers. There is ample room for original SMD research on how cities shape and are shaped by culture, which includes the arts as well as the meanings, perceptions and beliefs of urban citizens that must guide public action and governance strategies for sustainability.

Understanding the value of urban green space, as well as the use and social benefits associated with urban parks, is one additional area where SMD could provide new and important opportunities. Researchers have struggled for the last decade or more to bring values of nature more fully into planning, policymaking, and design and development opportunities[6]. Despite the rise of urban tree planting[128,129] and slow but steady progress integrating green space into urban infrastructure[130,131], fully accounting for the monetary value of urban green space for public physical and mental health, recreation, spiritual and aesthetic value, climate-change adaptation, disaster risk reduction and more has been hampered by lack of temporally and spatially comprehensive data on people's use of green space.

Though SMD is not necessarily a silver bullet for solving data gaps, it has the potential to play a strong part in providing new data streams that are temporally and spatially rich. To advance solutions to challenges raised in early SMD research, further research is required to develop new analytical methods and bring SMD into wider use in research.

Some of the strategies for dealing with SMD challenges suggested in current literature point to the combination of SMD analysis with more-traditional research techniques such as focus groups and surveys[105] or the incorporation of action research and mixed-methods research designs in general[97,119]. Other strategies present specific tactics that researchers can deploy to mitigate or resolve a variety of issues related both to the representation and the representativeness of human thoughts and behaviour and how they play out in time and space. Researchers have already shown the merits of using SMD from multiple platforms simultaneously — such as Twitter and Flickr[132]; Twitter, Flickr, Instagram and Picasa[31]; or Facebook, Twitter and Foursquare[47] — to achieve greater representativeness of their samples. Further, when using SMD to study people's opinions and behaviours, analysts have demonstrated that time-related inconsistencies of population representation can be effectively addressed by conceiving of the data as the result of a continuous opt-in panel rather than a 'perfect' cross-sectional survey[85]. Missing or inaccurate locations of SMD posts have also been shown to be a tractable problem, solvable through both high-tech approaches, such as using language-model algorithms to derive the location of a Flickr image through a user's textual meta-data on Twitter[93], and low-tech solutions, such as the manual verification of images with upload location errors[27].

Overall, SMD researchers across different domains share the belief that, as with early polling techniques (for example, the 'Dewey Defeats Truman' headline incident in 1948), research tactics and techniques in social-media research will be increasingly improved and refined over time, leading to greater credibility and confidence in the field. The direct, almost real-time input that local administrations can obtain from urban dwellers via SMD has the potential to establish new, more-meaningful connections between citizens and institutions, thus bringing improved synergies to the management of the urban ecosystem and the vital services it provides to humans and other living organisms in the city. Whose voices are being heard and how all urban communities can benefit from the new insights that SMD research provides are key questions that scholars will have to attend to as we lay the groundwork for a new domain in sustainability research.

SMD have the potential to bring social data in line with more-extensive environmental data such as satellite-derived spatial data and other sensor-based data already being used in social- and environmental-systems research. We suggest that pairing spatially and temporally extensive environmental data with SMD has the potential to revolutionize our ability to understand social–ecological interactions and feedbacks in complex systems such as cities and urbanized regions[6].

Data availability

The data that support the findings of this study are available from the authors on reasonable request.

Received: 28 September 2017; Accepted: 11 September 2018;
Published online: 15 October 2018

References

1. Elmqvist, T. et al. (eds) *The Urban Planet: Knowledge Toward Sustainable Cities* (Cambridge Univ. Press, 2018).
2. Parnell, S., Elmqvist, T., McPhearson, T., Nagendra, H. & Sörlin, S. in *The Urban Planet: Knowledge Towards Sustainable Cities* (eds T. Elmqvist et al.) (Cambridge Univ. Press, 2018).
3. Berger, M. The unsustainable city. *Sustainability* **6**, 365–374 (2014).
4. Acuto, M., Parnell, S. & Seto, K. C. Building a global urban science. *Nat. Sustain.* **1**, 2–4 (2018).

Provides an overview and synthesis of the state of global urban science and underscores the need to develop better tools for data-driven urban policy at the global level.

5. McPhearson, T. Scientists must have a say in the future of cities. *Nature* **538**, 165–166 (2016).
6. McPhearson, T. et al. Advancing urban ecology toward a science of cities. *Bioscience* **66**, 198–212 (2016).
Calls for a science of cities and advances in the use of 'Big Data' to address urban sustainability challenges. It highlights the need for an integrated social–ecological–technological systems approach.
7. Ürge-Vorsatz, D. et al. Locking in positive climate responses in cities. *Nat. Clim. Change* **8**, 174–177 (2018).
8. Acuto, M. Global science for city policy. *Science* **359**, 165–166 (2018).
9. Langemeyer, J., Baró, F., Roebeling, P. & Gómez-Baggethun, E. Contrasting values of cultural ecosystem services in urban areas: The case of park Montjuïc in Barcelona. *Ecosyst. Serv.* **12**, 178–186 (2015).
10. *Transforming Our World: The 2030 Agenda for Sustainable Development A/RES/70/1* (United Nations, 2015).
11. Hamstead, Z. A. et al. Geolocated social media as a rapid indicator of park visitation and equitable park access. *Comput. Environ. Urban Syst.* **72**, 38–50 (2018).
12. Brouwer, T. et al. Probabilistic flood extent estimates from social media flood observations. *Nat. Hazards Earth Syst. Sci.* **17**, 735–747 (2017).
13. Yang, W., Mu, L. & Shen, Y. Effect of climate and seasonality on depressed mood among twitter users. *Appl. Geogr.* **63**, 184–191 (2015).
This paper provides a good example of how 'big data' from social media can help researchers to study the relationship between depression and specific aspects of climate and seasonality in different urban areas.
14. Hoogendoorn, G. & Gregory, J. Instagrammers, urban renewal and the Johannesburg inner city. *Urban Forum* **27**, 399–414 (2016).
15. Yang, F., Jin, P. J., Cheng, Y., Zhang, J. & Ran, B. Origin-destination estimation for non-commuting trips using location-based social networking data. *Int. J. Sustain. Transp.* **9**, 551–564 (2015).
16. Wu, C. et al. Spatial and social media data analytics of housing prices in Shenzhen, China. *PLoS One* **11**, 1–19 (2016).
17. Gómez-Baggethun, E. & Barton, D. N. Classifying and valuing ecosystem services for urban planning. *Ecol. Econ.* **86**, 235–245 (2013).
18. *Urbanization, Biodiversity and Ecosystem Services: Challenges and Opportunities: a Global Assessment* (Springer, 2013).
19. de Groot, R. et al. Global estimates of the value of ecosystems and their services in monetary units. *Ecosyst. Serv.* **1**, 50–61 (2012).
20. Small, N. & Munday, M. & Durance, I. The challenge of valuing ecosystem services that have no material benefits. *Glob. Environ. Change* **44**, 57–67 (2017).
21. Wood, S. A., Guerry, A. D., Silver, J. M. & Lacayo, M. Using social media to quantify nature-based tourism and recreation. *Sci. Rep.* **3**, 2976 (2013).
This is a pioneering study to ground-truth the use of social-media data (Flickr) to test its potential to predict visitation rates in recreational sites worldwide.
22. Sessions, C., Wood, S. A., Rabotyagov, S. & Fisher, D. M. Measuring recreational visitation at US National Parks with crowd-sourced photographs. *J. Environ. Manage.* **183**, 703–711 (2016).
23. Sonter, L. J., Watson, K. B., Wood, S. A. & Ricketts, T. H. Spatial and temporal dynamics and value of nature-based recreation, estimated via social media. *PLoS One* **11**, 1–16 (2016).
24. Keeler, B. L. et al. Recreational demand for clean water: evidence from geotagged photographs by visitors to lakes. *Front. Ecol. Environ.* **13**, 76–81 (2015).
25. Donahue, M. L. et al. Using social media to understand drivers of urban park visitation in the Twin Cities, MN. *Landsc. Urban Plan.* **175**, 1–10 (2018).
26. Roberts, H. V. Using Twitter data in urban green space research: A case study and critical evaluation. *Appl. Geogr.* **81**, 13–20 (2017).
27. Guerrero, P., Møller, M. S., Olafsson, A. S. & Snizek, B. Revealing cultural ecosystem services through Instagram images: the potential of social media volunteered geographic information for urban green infrastructure planning and governance. *Urban Plan.* **1**, 1–17 (2016).
This article shows how volunteered geographic information from social media (Instagram) can be used to gain knowledge on how people interact with and perceive urban green spaces and inform future green infrastructure planning.
28. Schwartz, R. & Hochman, N. in *Locative Media* (eds. R. Wilken & G. Goggin) 52–65 (Routledge, 2015).
29. Murakami, D., Peters, G. W., Yamagata, Y. & Matsui, T. Participatory sensing data tweets for micro-urban real-time resiliency monitoring and risk management. *IEEE Access* **4**, 347–372 (2016).
30. Stefanidis, A., Crooks, A. & Radzikowski, J. Harvesting ambient geospatial information from social media feeds. *GeoJournal* **78**, 319–338 (2011).
31. Kent, J. D. & Capello, H. T. Spatial patterns and demographic indicators of effective social media content during the Horsethief Canyon fire of 2012. *Cartogr. Geogr. Inf. Sci.* **40**, 78–89 (2013).
32. Wang, Y., Wang, T., Ye, X., Zhu, J. & Lee, J. Using social media for emergency response and urban sustainability: a case study of the 2012 Beijing rainstorm. *Sustainability* **8**, 25 (2015).
33. Kusumo, A. N. L., Reckien, D. & Verplanke, J. Utilising volunteered geographic information to assess resident's flood evacuation shelters. Case study: Jakarta. *Appl. Geogr.* **88**, 174–185 (2017).
34. Tkachenko, N., Jarvis, S. & Procter, R. Predicting floods with Flickr tags. *PLoS One* **12**, e0172870 (2017).
35. Cervone, G., Schnebele, E., Waters, N., Moccaldi, M. & Sicignano, R. in *Seeing Cities Through Big Data* (eds. P. Thakuriah, N. Tilahun, & M. Zellner) 443–457 (Springer, 2017).
36. Shelton, T., Poorthuis, A., Graham, M. & Zook, M. Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'. *Geoforum* **52**, 167–179 (2014).
37. Liu, J. C.-E. & Zhao, B. Who speaks for climate change in China? Evidence from Weibo. *Clim. Change* **140**, 413–422 (2017).
38. Cody, E. M., Reagan, A. J., Mitchell, L., Dodds, P. S. & Danforth, C. M. Climate change sentiment on Twitter: An unsolicited public opinion poll. *PLoS One* **10**, 1–18 (2015).
39. Morris, G. P., Beck, S. A., Hanlon, P. & Robertson, R. Getting strategic about the environment and health. *Public Health* **120**, 889–903 (2006).
40. Widener, M. J. & Li, W. Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. *Appl. Geogr.* **54**, 189–197 (2014).
41. Chen, X. & Yang, X. Does food environment influence food choices? A geographical analysis through 'tweets'. *Appl. Geogr.* **51**, 82–89 (2014).
This is one of the first studies using social-media data (Twitter) to assess how people's food-related activities are influenced by their local food environment.
42. Gore, R. J., Diallo, S. & Padilla, J. You are what you tweet: Connecting the geographic variation in America's obesity rate to twitter content. *PLoS One* **10**, 1–16 (2015).
43. Ranney, M. L. et al. Tweet now, see you in the ED later? Examining the association between alcohol-related tweets and emergency care visits. *Acad. Emerg. Med.* **23**, 831–834 (2016).
44. Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S. & Danforth, C. M. The geography of happiness: connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS One*. https://doi.org/10.1371/journal.pone.0064417 (2013).
45. Yang, W. & Mu, L. GIS analysis of depression among Twitter users. *Appl. Geogr.* **60**, 217–223 (2015).
46. Nguyen, Q. C. et al. Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity. *Appl. Geogr.* **73**, 77–88 (2016).
47. Ben-Harush, O., Carroll, J.-A. & Marsh, B. Using mobile social media and GIS in health and place research. *Continuum* **26**, 715–730 (2012).
48. Hurst, C. E., Gibbon, H. M. F. & Nurse, A. M. *Social Inequality: Forms, Causes, and Consequences* (Routledge, 2017).
49. Shelton, T., Poorthuis, A. & Zook, M. Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landsc. Urban Plan.* **142**, 198–211 (2015).
This paper offers important insights into how geotagged social-media data (Twitter) can help study questions of socio-spatial inequality and contribute to the advancement of critical GIScience in cities.
50. Adnan, M., Lansley, G. & Longley, P. A geodemographic analysis of the ethnicity and identity of Twitter users in Greater London. *Proc. 21st Conf. GIS* **44**, 1–6 (2013).
51. Quercia, D. & Saez, D. Mining urban deprivation from Foursquare: Implicit crowdsourcing of city land use. *IEEE Pervasive Comput.* **13**, 30–36 (2014).
52. Li, J., Qin, Q., Han, J., Tang, L. A. & Lei, K. H. Mining trajectory data and geotagged data in social media for road map inference. *Trans. GIS* **19**, 1–18 (2015).
53. Zhou, X., Wang, M. & Li, D. From stay to play — A travel planning tool based on crowdsourcing user-generated contents. *Appl. Geogr.* **78**, 1–11 (2017).
54. Hasan, S. & Ukkusuri, S. V. Urban activity pattern classification using topic models from online geo-location data. *Transp. Res. Part C Emerg. Technol.* **44**, 363–381 (2014).
This paper uses social-media data (Twitter) to extract typologies and patterns of daily urban activities with the potential to inform activity-based modelling and future urban transportation planning.
55. Luo, F., Cao, G., Mulligan, K. & Li, X. Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. *Appl. Geogr.* **70**, 11–25 (2016).
56. Schweitzer, L. Planning and social media: A case study of public transit and stigma on twitter. *J. Am. Plan. Assoc.* **80**, 218–238 (2014).

57. Lucchese, C., Perego, R. & Silvestri, F. in *Advances in Information Retrieval* (eds. Baeza-Yates, R. et al.) 195–206 (Springer-Verlag, 2012).
58. Assem, H., Buda, T. S. & O'Sullivan, D. RCMC : Recognizing crowd-mobility patterns in cities based on: location based social networks data. *ACM Trans. Intell. Syst. Technol.* **8**, 70–100 (2017).
59. Hawelka, B. et al. Geo-located Twitter as proxy for global mobility patterns. *Cartogr. Geogr. Inf. Sci.* **41**, 260–271 (2014).
60. Paldino, S., Bojic, I., Sobolevsky, S., Ratti, C. & González, M. C. Urban magnetism through the lens of geo-tagged photography. *EPJ Data Sci.* **4**, 1–17 (2015).
61. Martí, P., Serrano-Estrada, L. & Nolasco-Cirugeda, A. Using locative social media and urban cartographies to identify and locate successful urban plazas. *Cities* **64**, 66–78 (2017).
62. Brandt, T., Bendler, J. & Neumann, D. Social media analytics and value creation in urban smart tourism ecosystems. *Information Manag.* **54**, 703–713 (2017).
63. Zhai, S. et al. Mapping the popularity of urban restaurants using social media data. *Appl. Geogr.* **63**, 113–120 (2015).
64. Lovelace, R., Birkin, M., Cross, P. & Clarke, M. From big noise to big data: Toward the verification of large data sets for understanding regional retail flows. *Geogr. Anal.* **48**, 59–81 (2016).
This paper offers an approach to test the reliability of geotagged 'big data' from social media (Twitter) for mapping retail flows in an urban region, by using mobile phone and consumer-survey data.
65. Boyd, D. & Crawford, K. Critical questions for big data. *Information Commun. Soc.* **15**, 662–679 (2012).
66. Twitter now has 330 million monthly active users. *PC Tech Magazine* (2017).
67. Crampton, J. et al. Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. *Cartogr. Geogr. Inf. Sci.* **40**, 130–139 (2013).
68. Ratkiewicz, J. et al. Detecting and tracking political abuse in social media. *Int. Conf. Web Soc. Media* **11**, 297–304 (2011).
69. Housley, W. et al. Big and broad social data and the sociological imagination: A collaborative response. *Big Data Soc.* **1**, 1–15 (2014).
70. Schoen, H. et al. The power of prediction with social media. *Internet Res.* **23**, 528–543 (2013).
71. Sloan, L., Morgan, J., Burnap, P. & Williams, M. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PLoS One* **10**, e0115545 (2015).
72. Acquisti, A. & Gross, R. Predicting social security numbers from public data. *Proc. Natl Acad. Sci. USA* **106**, 10975–10980 (2009).
73. Ruths, D. & Pfeffer, J. Social media for large studies of behavior. *Science* **346**, 1–2 (2014).
This article offers an in-depth overview of the potential biases and flows in social-media data and how these can be reduced.
74. Maeve, D. & Brenner, J. *The Demographics of Social Media Users — 2012* (Pew Research Center, 2013).
75. Dunkel, A. Visualizing the perceived environment using crowdsourced photo geodata. *Landsc. Urban Plan.* **142**, 173–186 (2015).
76. Abel, F., Araujo, S., Gao, Q. & Houben, G.-J. Analyzing cross-system user modeling on the social web. in *Proc. 11th Int. Conf. Web Eng.* 28–43 (2011).
77. Jendryke, M., Balz, T., McClure, S. C. & Liao, M. Putting people in the picture: Combining big location-based social media data and remote sensing imagery for enhanced contextual urban information in Shanghai. *Comput. Environ. Urban Syst.* **62**, 99–112 (2017).
78. Zickuhr, K. *Who's Not Online and Why* (Pew Research Center, 2013).
79. *Internet/Broadband Fact Sheet* (Pew Research Center, 2017).
80. *Social Media Fact Sheet: Demographics of Social Media Users and Adoption in the United States* (Pew Research Center, 2016).
81. Leetaru, K. H., Shaowen, W., Guofeng, C., Padmanabhan, A. & Shook, E. Mapping the global Twitter hearbeat: The geography of Twitter. *First Monday* **18**, 5–6 (2013).
82. Catt, R. D. 100,000,000 Geotagged Photos (Plus) (2009); code.flickr.net/2009/02/04/100000000-geotagged-photos-plus
83. Cranshaw, J., Schwartz, R., Hong, J. & Sadeh, N. The livehoods project: utilizing social media to understand the dynamics of a city. in *Int. Conf. Web Soc. Media* 58–65 (2012).
84. Rattenbury, T. & Naaman, M. Methods for extracting place semantics from Flickr tags. *ACM Trans. Web* **3**, 1–30 (2009).
85. Diaz, F., Gamon, M., Hofman, J. M., Kiciman, E. & Rothschild, D. Online and social media data as an imperfect continuous panel survey. *PLoS One* **11**, 1–21 (2016).
86. Liu, B., Yuan, Q., Cong, G. & Xu, D. Where your photo is taken: Geolocation prediction for social images. *J. Assoc. Inf. Sci. Technol.* **65**, 1232–1243 (2014).
87. Ribeiro, S. & Pappa, G. L. Strategies for combining Twitter users geo-location methods. *Geoinformatica* **22**, 563–587 (2017).
This paper provides a comprehensive evaluation of sixteen different approaches for social-media user (Twitter) location inference and highlights those with the greatest accuracy.
88. Mahmud, J., Nichols, J. & Drews, C. Where is this tweet from? Inferring home locations of Twitter users. in *Proc. Sixth Int. AAAI Conf. Weblogs Soc. Media* 511–514 (2012).
89. Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J. & Mühlhäuser, M. A multi-indicator approach for geolocalization of tweets. in *Int. Conf. Web Soc. Media* 573–582 (2013).
90. Cheng, Z., Caverlee, J. & Lee, K. You are where you tweet: a content-based approach to geo-locating Twitter users. in *Proc. 19th ACM Int. Conf. Inform Knowledge Manag.* 759–768 (ACM, 2010).
91. Paraskevopoulos, P. & Palpanas, T. Where has this tweet come from? Fast and fine-grained geolocalization of non-geotagged tweets. *Soc. Netw. Anal. Min.* **6**, 89 (2016).
92. Oku, K., Hattori, F. & Kawagoe, K. Tweet-mapping method for tourist spots based on now-tweets and spot-photos. *Procedia Comput. Sci.* **60**, 1318–1327 (2015).
93. Hauff, C. & Houben, G. in *Advances in Information Retrieval* (eds. R. Baeza-Yates et al.) **7224**, 85–96 (Springer-Verlag, 2012).
94. Schwartz, R. & Halegoua, G. R. The spatial self: Location-based identity performance on social media. *New Media Soc.* **17**, 1643–1660 (2014).
95. Graham, M., Stephens, M. & Hale, S. Featured graphic. Mapping the geoweb: A geography of Twitter. *Environ. Plan. A* **45**, 100–102 (2013).
96. Naaman, M. Geographic information from georeferenced social media data. *SIGSPATIAL Spec.* **3**, 54–61 (2011).
97. Goodspeed, R. The limited usefulness of social media and digital trace data for urban social research. in *Proc. Sixth Int. AAAI Conf. Weblogs Soc. Media Underst.* 2–4 (2013).
98. Bacallao-Pino, L. M. Social media mobilisations: Articulating participatory processes or visibilizing dissent? *Cyberpsychology J. Psychosoc. Res. Cybersp.* https://doi.org/10.5817/CP2014-3-3 (2014).
99. Lovelace, R., Malleson, N., Harland, K. & Birkin, M. Geotagged tweets to inform a spatial interaction model: a case study of museums. Preprint at https://arxiv.org/ftp/arxiv/papers/1403/1403.5118.pdf (2014).
100. Cardullo, P. 'Hacking multitude' and big data: Some insights from the Turkish 'digital coup'. *Big Data Soc.* **2**, 1–14 (2015).
101. Grieve, R., Witteveen, K. & Tolan, G. A. Social media as a tool for data collection: examining equivalence of socially value-laden constructs. *Curr. Psychol.* **33**, 532–544 (2014).
102. Hollenstein, L. & Purves, R. Exploring place through user-generated content: Using Flickr to describe city cores. *J. Spat. Inf. Sci.* https://doi.org/10.5311/JOSIS.2010.1.3 (2010).
103. Schockaert, S. Vague regions in geographic information retrieval. *SigSpatial Spec.* **3**, 24–28 (2011).
104. Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J. & Tsioutsiouliklis, K. Discovering geographical topics in the twitter stream. *Proc. 21st Int. Conf. World Wide Web* https://doi.org/10.1145/2187836.2187940 (2012).
105. Pettit, A. The promises and pitfalls of SMR: Prevailing discussions and the naked truth. *Mark. Res.* 14–22 (2011).
106. Schwartz, H. A. & Ungar, L. H. Data-driven content analysis of social media: A systematic overview of automated methods. *Ann. Am. Acad. Pol. Soc. Sci.* **659**, 78–94 (2015).
107. Gonçalves, P., Araújo, M., Benevenuto, F. & Cha, M. Comparing and combining sentiment analysis methods. in *Proc. First ACM Conf. Online Soc. Networks* https://doi.org/10.1145/2512938.2512951 27–38 (ACM, 2013).
This paper assesses the merits of different existing approaches to sentiment analysisbased on social-media data and points to possible way to combine them to achieve better coverageand agreement results.
108. Jaewoo, K., Cha, M., Lee, W. & Sandholm, T. Identifying crime-prone areas based on tweet sentiments. *Telecommun. Rev.* **24**, 339–347 (2014).
109. Bradley, M. M. & Lang, P. J. *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings* (Citeseer, 1999).
110. Yuan, J., Mcdonough, S., You, Q. & Luo, J. Sentribute: Image sentiment analysis from a mid-level perspective. *Proc. Second Int. Work. Issues Sentim. Discov. Opin. Min.* https://doi.org/10.1145/2502069.250207 (2013).
111. Borth, D., Ji, R., Chen, T., Breuel, T. & Chang, S.-F. Large-scale visual sentiment ontology and detectors using adjective noun pairs. *Proc. 21st ACM Int. Conf. Multimed.* https://doi.org/10.1145/2502081.2502282 (2013).
112. Wang, Y., Wang, S., Tang, J., Liu, H. & Li, B. Unsupervised sentiment analysis for social media images. *Proc. 24th Int. Conf. Artif. Intell.* 2378–2379 (2015).
113. Godbey, G. C., Caldwell, L. L., Floyd, M. & Payne, L. L. Contributions of leisure studies and recreation and park management research to the active living agenda. *Am. J. Prev. Med.* **28**, 150–158 (2005).
114. Young, O. R. et al. *Earth System Challenges and a Multi-Layered Approach for the Sustainable Development Goals* (United Nations University Institute for the Advanced Study of Sustainability, 2014).
115. Perry, P. Certain problems in election survey methodology. *Public Opin. Q.* **43**, 312–325 (1979).
116. Barocas, S. & Nissenbaum, H. Big data's end run around procedural privacy protections. *Commun. ACM* **57**, 31–33 (2014).

117. Sui, D. & Goodchild, M. The convergence of GIS and social media: challenges for GIScience. *Int. J. Geogr. Inf. Sci.* **25**, 1737–1748 (2011).
118. *A GIS Code of Ethics* (GIS Certification Institute, 2003).
119. Kennedy, H., Moss, G., Birchall, C. & Moshonas, S. Balancing the potential and problems of digital methods through action research: methodological reflections. *Inform. Commun. Soc.* **18**, 172–186 (2014).
120. Young, S. D. Behavioral insights on big data: using social media for predicting biomedical outcomes. *Trends Microbiol.* **22**, 601–602 (2014).
121. Yang, C., Huang, Q., Li, Z., Liu, K. & Hu, F. Big data and cloud computing: innovation opportunities and challenges. *Int. J. Digit. Earth* **10**, 13–53 (2017).
122. Ediger, D., Jiang, K., Riedy, J., Bader, D. A. & Corley, C. Massive social network analysis: Mining Twitter for social good. in *ICPP 2010 39th Int. Conf. Parallel Proc.* 583–593 (IEEE, 2010).
123. Cao, G. et al. A scalable framework for spatiotemporal analysis of location-based social media data. *Comput. Environ. Urban Syst.* **51**, 70–82 (2015).
124. McCall, M. K. Seeking good governance in participatory-GIS: A review of processes and governance dimensions in applying GIS to participatory spatial planning. *Habitat Int.* **27**, 549–573 (2003).
125. Dunn, C. E. Participatory GIS — a people's GIS? *Prog. Hum. Geogr.* **31**, 616–637 (2007).
126. Campbell, S. Green Cities, Growing Cities, Just Cities?: Urban Planning and the Contradictions of Sustainable Development. *J. Am. Plan. Assoc.* **62**, 296–312 (1996).
127. Hawkes, J. *The Fourth Pillar of Sustainability: Culture's Essential Role in Public Planning* (Common Ground, 2001).
128. Conway, T. M. Tending their urban forest: Residents' motivations for tree planting and removal. *Urban For. Urban Green.* **17**, 23–32 (2016).
129. Pincetl, S., Gillespie, T., Pataki, D. E., Saatchi, S. & Saphores, J. D. Urban tree planting programs, function or fashion? Los Angeles and urban tree planting campaigns. *GeoJournal* **78**, 475–493 (2013).
130. Jayasooriya, V. M., Ng, A. W. M., Muthukumaran, S. & Perera, B. J. C. Green infrastructure practices for improvement of urban air quality. *Urban For. Urban Green.* **21**, 34–47 (2017).
131. Newell, J. P. et al. Green Alley Programs: Planning for a sustainable urban infrastructure? *Cities* **31**, 144–155 (2013).
132. Li, L., Goodchild, M. F. & Xu, B. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartogr. Geogr. Inf. Sci.* **40**, 61–77 (2013).

Acknowledgements
T.M. was supported by the Urban Resilience to Extreme Weather-Related Events Sustainability Research Network (URExSRN; NSF grant no. SES 1444755). T.M.'s research was also carried out as part of the project ENABLE, funded through the 2015–2016 BiodivERsA COFUND call for research proposals, with the national funders The Swedish Research Council for Environment, Agricultural Sciences, and Spatial Planning, Swedish Environmental Protection Agency, German Aeronautics and Space Research Centre, National Science Centre (Poland), The Research Council of Norway and the Spanish Ministry of Economy and Competitiveness.

Author contributions
Both authors contributed to the writing of this manuscript and the ideas it contains.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41893-018-0153-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to R.T.I.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.