4-2024

From the Editors: Mobilizing new sources of data: Opportunities and recommendations

Denis A. GREGOIRE

Anne L. J. TER WAL

Laura M. LITTLE

Sekou BERMISS

Reddi KOTHA
Singapore Management University, reddikotha@smu.edu.sg

See next page for additional authors

© *Academy of Management Journal*
2024, Vol. 67, No. 2, 289–298.
https://doi.org/10.5465/amj.2024.4002

FROM THE EDITORS

MOBILIZING NEW SOURCES OF DATA: OPPORTUNITIES AND RECOMMENDATIONS

In June 2008, the U.S.-based website Glassdoor.com began posting anonymous company reviews and salary data from current and former employees of various organizations. Doing so not only brought to the world information that had hitherto been restricted to private circles, it spontaneously prompted some organizations to alter their workplace practices (Dineen & Allen, 2016; Dube & Zhu, 2021). At the same time, Glassdoor's very activities gave rise to a completely new source of data for exploring a wealth of management and organizational phenomena (e.g., Bermiss & McDonald, 2018; Rhee, 2024). As this example illustrates, new data sources can not only transform managerial and organizational practices, they also invite the development of innovative theoretical explanations and can unlock opportunities to advance academic understanding of a broad range of management and organizational phenomena.

With the increased digitization that accompanies the transformational era we live in (Gruber, 2023), at least three major trends open up opportunities for mobilizing new sources of data. First, novel technologies such as wearable health monitors, sensory badges, and neuroimaging techniques offer new means of collecting data (Harari & Gosling, 2023; Laureiro-Martínez, Brusoni, Canessa & Zollo, 2015; Polzer, 2022). Second, the cost of conducting field experiments and interventions has dramatically plummeted, making firms and other organizations prominent players in generating experimental data (Cao, Koning & Nanda, 2023) and opening promising avenues for cooperation between research and practice (Kohavi & Thomke, 2017). Third, the digital footprints and traces inherent to the very functioning of social media platforms, artificial intelligence (AI) and blockchain technologies, as well as sensor-enabled cell phones and video-monitoring platforms, to name but a few applications, effectively create "big data," which can be analyzed with machine learning and other analytical techniques to potentially transform the landscape of management research (von Krogh, Roberson & Gruber, 2023).

Yet, realizing the promise of new data sources is no easy task. "Good" access to quality data is often difficult—the more so when investigating novel phenomena. When relevant data are not altogether absent, archival databases often lack measures that are specific enough for notions of theoretical interest—thus undermining validity. Similarly, validating survey instruments and experimental manipulations takes time and may be difficult when large swaths of potential participants are unfamiliar with the phenomenon being studied. And, while qualitative researchers often circumvent these challenges by collecting data directly from primary informants, the novelty of a given phenomenon and informants' limited familiarity with it may still place important limits on scholars' ability to uncover theoretically relevant insights about people's sensemaking efforts (Langley, Bell, Bliese, LeBaron & Gruber, 2023). In the same way that "engaging in research on novel phenomena (can be) quite challenging and risky due to its inherent uncertainty and the potential disorder that it can create" (Dencker, Gruber, Miller, Rouse & von Krogh, 2023: 1295), mobilizing new data sources poses unique difficulties. And, as often happens in the case of transformational or disruptive innovations, opportunities for mobilizing and interpreting these new data are arising faster than guidance about best practices can be developed. This raises the danger that apparent advances in academic understanding later prove to have "feet of clay."

To help advance understanding of how new sources of data can transform management and organization scholarship and to encourage authors submitting their work to *Academy of Management Journal* (AMJ) to embrace these new sources, we discuss below a few illustrative cases and offer recommendations. Although our editorial focuses primarily on mobilizing new sources of data (like that obtained from Glassdoor above), we sometimes found it necessary to also discuss new methods for analyzing some of these new sources of data (such as machine learning). Yet the two forms of novelty are in many cases independent; for instance, new sources of data can be appropriately analyzed with

"classical" means (e.g., ordinary least squares regression or qualitative analyses) and, likewise, authors can use machine-learning techniques to analyze large datasets obtained from "classical" sources (Salganik et al., 2020; Yarkoni & Westfall, 2017).

In the paragraphs that follow, we highlight the theoretical and methodological opportunities that new sources of data offer and the trade-offs and ethical challenges they may imply. We then provide actionable recommendations for authors interested in tackling relevant challenges for mobilizing new sources of data—and new means of analysis, if relevant.

UNLOCKING OPPORTUNITIES PRESENTED BY NEW DATA SOURCES

The societal changes and technological advances unfolding in our current transformational times both demand and enable the use of new data sources (Gruber, 2023). Reviewing all the studies published in AMJ in the last seven years, we found that data sources represented in the journal have remained fairly stable. From 2017 through 2023, AMJ published 531 articles. Because a large portion of these articles contain multiple studies and multiple data sources, the relative proportion of data sources has fluctuated over time, despite the number of published articles remaining fairly constant. Archival sources constitute the largest representation in the sample, with 43% of annual papers published in AMJ containing at least one study using archival data. Survey (21%) and interview (18%) data are the two next most common data sources represented in AMJ papers, followed by experimental data (14%).

At AMJ, we believe there is substantial untapped potential for scholars—both in qualitative and quantitative as well as micro and macro research—to embrace new data sources in their research designs. New data sources play a key role in unlocking opportunities to ask innovative questions in light of a changing society and to shed new light on preexisting questions at the heart of management research. Below, we illustrate how new technologies and methodological advances generate new sources of data, and how existing data are a critical input to the new technologies and methods that management scholars can deploy to process and interpret data.

New Technologies That Offer New Sources of Data for Quantitative and Qualitative Research

The transformational era enables the use of new data sources, with tremendous opportunities for advancing quantitative and qualitative research at both micro and macro levels. First and foremost, the greater digitalization of data has generated a myriad of digital traces and footprints that open up promising avenues for leveraging new data sources in management research (Edelmann, Wolff, Montagne & Bail, 2020; Matz, 2022). Among other research opportunities, the widespread adoption of wearable health monitors such as the Apple Watch and Oura Ring offers a chance to employ mobile sensing methodologies to obtain more precise and objective indicators of well-being and physiological responses (Harari & Gosling, 2023). Similarly, the pervasive use of smartphones offers unprecedented opportunities to leverage app usage data (Sust, Talaifar & Stachl, 2023) and aggregate patterns of GPS data available from providers such as SafeGraph (Li, Ning, Jing & Lessani, 2024).

At the interface between big data and content-analysis techniques, advances in computer-assisted content analysis (McKenny, Aguinis, Short & Anglin, 2018) and natural language processing algorithms have allowed for the use of topic-modeling techniques to interpret large datasets of text data (Lockwood, Glynn & Giorgi, 2023; Taeuscher, Bouncken & Pesch, 2021). In very much the same thrust, greater use of video data—in conjunction with improved capabilities to process such data (Langley et al., 2023)—offers exciting opportunities for the study of collective decision-making (Veltrop, Bezemer, Nicholson & Pugliese, 2021) or the use of video diaries in qualitative research (de Rond, Holeman & Howard-Grenville, 2019). In this context, analysis of video-recorded boardroom conversations may provide path-breaking insights into the role that narratives and emotions play in how leaders make difficult decisions about sustainability challenges. Perhaps more broadly, advances in communication and immersive technologies open up new possibilities for qualitative researchers to collect data remotely with respondents facing difficult circumstances (such as natural disasters or wars) that might otherwise put researchers at such high risks that it would render the research impossible.

In addition to unlocking new sources of data, technological advances also raise new research questions. At a micro level, new technologies in the workplace not only result in the greater availability of new sources of data that can then be used for research purposes but could also yield new research opportunities in their own right. For example, wearable sensor badges allow for unobtrusively tracking people's movements within an organization, as well as whom

they meet with, for how long, and even the emotional tone of their exchanges (see Chaffin et al., 2017). More broadly, email data can be used as a source of data about, for example, employees' cultural fit with their organization (Srivastava, Goldberg, Manian & Potts, 2018). Social media data can be analyzed using natural language processing methodologies and linguistic analysis to offer unique insights into the communication styles of leaders from traditionally underrepresented groups (Dupree, 2024) or identity dynamics in collective action (Chiang, de Rond & Lok, 2023).

At a macro level, the transformational era raises new questions about the innovative products, services, business models, and organizational practices brought about by technological advances. For example, blockchain technologies have enabled new ways of automating the verification of a host of transactions between organizations as well as new ways of organizing, such as decentralized autonomous organizations (e.g., Hsieh & Vergne, 2023; Lumineau, Wang & Schilke, 2021). The timestamped transaction and data records—inherent to the high transparency of this organizational form—are a valuable complement to widely used archival data in the study of new organizational forms (e.g., Arslan, Vasudeva & Hirsch, 2024; Li & Khessina, 2024). Similarly, advanced imaging techniques may be used to gather fine-grained data on firms' sustainability practices; for example, using data that maps deforestation patterns, emissions, or carbon credit usage, which, in conjunction with text mining of firms' public claims regarding their sustainability initiatives, may be a critical input into research on greenwashing (Cenci, Burato, Rei & Zollo, 2023).

Taken together, opportunities for mobilizing new data sources for management scholarship abound. Although the above examples are but a selection of the broad array of exciting possibilities spurred by recent trends in the availability of data, at AMJ, we hope they will serve as a catalyst for the management scholarly community to embrace them enthusiastically (and also cautiously) in their research designs.

New Technologies That Rest On (and Yield) Big Data

A second area of opportunity revolves around how new data may not be the product of societal and technological changes but serve as an input to them. Unprecedented opportunities are emerging in light of the rapid development and diffusion of new technologies that rest on big data as input, and that in turn yield processed and interpreted data that can be mobilized in management research. In this context, the growing utilization of blockchain technology as well as generative AI and other machine-learning approaches within the workplace invite efforts to examine how and when these developments can be used as meaningful sources of data in how individuals, teams, and organizations take decisions. For instance, the use of machine learning in hiring decisions (Kelan, 2021; Raisch & Krakowski, 2021) necessarily rests on the availability of a wide swath of data about the candidates. To test whether such technologies may introduce unintended biases, however, it would seem relevant to compare their recommendation against benchmarks obtained from human judgment (see Yeomans, Shah, Mullainathan & Kleinberg, 2019).

Among the rapidly developing use cases of AI in organizational settings is the growing number of platforms that "automate" common organizational tasks. This includes tools for managing projects and communications among team members (e.g., monday.com, Taskade.com) or for inventory management and supply chain applications (e.g., Alloy.ai, Mely.ai). In human resources (HR), AI-enabled platforms exist not only to automate the sifting of prospective employees but also to handle common queries about HR benefits and policies (e.g., Airudi.com, Humanly.io, Workable.com). Perhaps even more intriguing is the case of generative AI platforms that allow managers to create advanced digital versions of themselves to handle increasing volumes of routine tasks, like answering low-priority communications and managing one's schedule (e.g., ALBISAI.com). Similarly, new data formats such as blockchain's unique timestamp records of data creation, removal, and exchange provide an unprecedented opportunity for the study of phenomena wherein blockchains are widely applied, such as in supply chains (e.g., OpenPort.com, OriginTrail.io). Common to all these tools is that their ongoing operations not only rest on prior communications and documents but also readily capture all digital exchanges across team members, employees, and their managers. As such, these tools offer ready access to large sets of process and text data that directly reflect the evolution of various organizational phenomena.

Meanwhile, advances in machine learning are unlocking new possibilities to analyze data formats such as large bodies of unstructured text that were previously unsuitable for large-scale analysis (Hannigan et al., 2019; von Krogh et al., 2023), which may be especially useful for capturing insightful

relationships from unstructured text including corporate reports, 10K corporate filings, patents and other legal documents, and social media data.

Generative AI methodologies may also have a role to play in experimental research. Researchers have already started to explore whether and how large language models can replace human subjects in pilot experiments (Bisbee, Clinton, Dorff, Kenkel & Larson, 2023) or replicate moral judgment (Dillion, Tandon, Gu & Gray, 2023)—not to mention how such tools can considerably speed up the "writing" and "correcting" of experimental instructions, manipulations, and other prompts. Similarly, the relentless pace of advancement in virtual and augmented reality has the potential to transform experimental research designs (Hubbard & Aguinis, 2023). For example, an enhanced sense of realism in simulated experimental tasks in decision-making in crisis situations may be a game changer in capturing, in minute detail, how participants respond to emergencies that may be difficult, if not impossible, to observe in the real world. Moreover, the data collection potential of augmented- and virtual-reality headsets is purported to be staggering. The new Apple headsets, for example, continuously track individuals and surroundings in three dimensions, capturing every hand gesture, eyeball movement, and environmental detail, collecting an unprecedented amount of data compared to other personal devices (O'Flaherty, 2024).

ADVICE FOR MOBILIZING NEW DATA SOURCES

Notwithstanding these opportunities' import, the large untapped potential of new data formats and sources for management research comes with its own set of unique challenges. Given readers' and reviewers' potentially limited familiarity with new data sources, not to mention the lack of established best practices in how best to use them for management research, the onus is on authors to make a compelling case for how their new data can help advance management theory and how they can best be integrated into a compelling overall research design. Although the nature and extent of the challenges will strongly depend on the type of new data being used, we posit that three broad areas of concern are of particular importance in the scholarly effort to unlock the opportunities presented by new data sources and, relatedly, new analytical methods for research published in AMJ: (1) data context, (2) data transparency, and (3) the alignment between theory and method.

Recommendation #1: Take Data Context Seriously

A common pitfall around the use of new data sources in research submitted to AMJ is a lack of detail on data context. Whereas established data sources may require little introduction, as authors can assume a degree of familiarity with commonly used databases and data formats, it becomes of critical importance for researchers using new and unfamiliar data to provide detailed background information on how, when, where, and why the data were compiled. This is equally important for primary data intentionally collected for research purposes as it is for "organic" data used as a secondary source (e.g., to corroborate or triangulate findings).

First, a rich descriptive account of the data context provides the fundamental background to key underlying assumptions. This is critical to ensuring a tight match between theory and data and for building a solid foundation for correct interpretation (Gruber & Bliese, 2024). For instance, leadership scholars leveraging video data to examine the role of emotions in organizational decision-making will need to make a compelling case for why emotions in one aspect of their communication—that is, their recorded speeches—should be representative of their broader approach to leadership to make a credible link to theorized outcomes of their leadership approach. A detailed data description should convey with precision what types of emotions one can expect in such speeches, what sources of potential variation there would be, and why such variation can justifiably be understood to be reflective of broader patterns in the presence of emotions in different leadership styles. Similarly, company reports or 10K corporate filings will be highly suitable for text-mining techniques and machine-learning approaches. Still, without details on who compiled these reports, what the underlying incentives and motivations for information provision are, what is legally required, or even what information may have been deliberately left out, it will be difficult for the reader to gain confidence in the data's suitability for a given research purpose. In quantitative research, a few well-targeted interviews with people knowledgeable about how data were compiled can go a long way in putting to the test key assumptions about what the data represent. Conversely, in qualitative research, the provision of broader contextual

information about the geographical, industry, or temporal context in which a particular data source or format is situated can be highly effective in convincing readers about the data's appropriateness.

Second, contextual information on new datasets is essential to provide clarity on the extent to which the data contained within them are representative of the broader phenomenon they are intended to illustrate. In a similar way that patent data are not necessarily representative of the overall body of inventions in a given firm or industry (Criscuolo, Alexy, Sharapov & Salter, 2019), most new data sources will selectively capture certain aspects of a phenomenon of interest, while omitting others. For instance, when using long time series of data, authors should reflect on whether the logic by which the data were compiled or accumulated may have shifted over time, due to changes in regulatory requirements, which may imply the data capture a different phenomenon at the beginning of the covered time period than at the end. Similarly, scholars using audio- or video-recordings as sources of data and those who use other forms of digital data will need to consider how the recording of data may have changed or manipulated the behaviors or decisions being studied. It cannot be simply assumed that recorded boardroom meetings are as frank as unrecorded ones, or that people will not change their interaction patterns when they are aware of being tracked. Certain projects may only exist informally in organizations (Criscuolo, Salter & ter Wal, 2014), meaning that digital project repositories may only partially capture an organization's project portfolio. Although not a problem per se, acknowledgment of such limitations is critical to ensuring theoretical arguments are closely tied to what is captured empirically and avoid a mismatch between theoretical claims and empirical reality.

Contextual considerations are also crucial when employing machine-learning techniques. Simplifying assumptions made during sampling from a dataset and partitioning it into training and testing sets can render the analysis ineffective. For instance, disregarding shocks such as the global financial crisis of 2007–2008, the constraints on decision-makers' choices, individual and firm idiosyncrasies, and the temporal sequence in the data can all lead to results with limited normative value. Accordingly, authors need to draw the contours of the historical, institutional, or geographical context in which data inputted to machine-learning models was first created, and perhaps provide samples of the inputted materials to help readers and reviewers get a more concrete and comprehensive picture of the nature of the data that fed into the modeling approach.

Recommendation #2: Take Data Transparency and Ethical Considerations Seriously

A second area of attention in the adoption of new data sources relates to transparency and ethical concerns. Alongside other leading academic journals, AMJ strongly encourages and supports authors' efforts to transparently provide a full set of relevant methodological details (DeCelles, Howard-Grenville & Tihanyi, 2021; Grimes, von Krogh, Feuerriegel, Rink & Gruber, 2023). In terms of the transparency of how data were obtained, processed, and analyzed, the expected standards to be upheld by management scholarship published in AMJ and elsewhere are increasing, yet the burden on authors to be fully transparent on this front is amplified in cases in which they adopt new sources of data. In addition to adopting general best practices in empirical research (such as preregistration of study design and hypotheses, ethical review, and making study materials and anonymized subsets of data available for blind peer review), it is imperative for authors to pay attention to transparency and ethical considerations that are specific to the new data sources and formats they are using.

First, transparency in terms of how data were accessed or generated is critical. For example, when using data not originally compiled for research purposes, it becomes imperative to establish that data used for research is not in violation of any data use agreements, ownership rights, or consent procedures that enabled the original compilation of the data. In the absence of clear guidelines or precedents on this front, we recommend that authors proactively engage with officers from their local ethical review boards to determine how best to establish why it may be acceptable to use anonymized non-identifiable data (including text and actual quotes) even if the people "behind the data" did not explicitly authorize their use for a specific research purpose. Similarly, when using social media data or crowd-sourced data, authors may discuss how potential risks of data inference and manipulation or the presence of fake accounts have been mitigated. When using data based on generative AI methodologies, transparency on the inputs used to generate data will be critical to allow for a degree of reproducibility. When space constraints would otherwise prohibit such disclosures, one may consider uploading this information in an online appendix on the Open Science

Framework (osf.io), which authors can make anonymously available through a hyperlink in the manuscript.

Second, it is important for authors to be transparent about how data are handled. With machine-learning methodologies, it is important to demonstrate how observed patterns may be sensitive to the specific choice of approach or the properties of the initial training dataset used in supervised learning models. Relatedly, it is critical to be aware of how using unrepresentative training data may introduce unintended biases. The currently available large language models have been shown to perform well in emulating the properties of people from Western, industrialized, rich, educated, and democratic societies, but significantly worse in other contexts (Atari, Xue, Park, Blasi & Henrich, 2023)—and to reproduce human tendencies for in-group solidarity and out-group hostility (Hu, Kyrychenko, Rathje, Collier, van der Linden & Roozenbeek, 2023).

Third, field experimenters—in academia and in private organizations and platforms—must be transparent about how they balance the demands of science, such as providing causal evidence, with the potential harm that may befall participants (Rahman, 2024; Rahman, Weiss & Karunakaran, 2023). For example, attempting to eliminate alternative explanations by designing experimental conditions that vary in only one factor can result in interventions that are unusual and atypical. For example, training entrepreneurs with expert instructors or having them learn from the life lessons and "war lessons" of successful entrepreneurs are two common practices. Yet these practices vary in more than one dimension: the content (life lessons vs. frameworks) and instructors (successful entrepreneurs vs. instructors) differ. Therefore, differences between the two training methods cannot be causally attributed to a single factor. Creating control conditions wherein instructors share war stories of successful entrepreneurs or where successful entrepreneurs impart frameworks may be counterproductive and may even adversely affect participants. Hence, the need for experimental control must be balanced with the potential for harm. Experimenters should specify stopping rules when adverse effects are detected.

Recommendation #3: Minding the Alignment of Theory and Method

An insidious challenge of mobilizing new data sources and formats lies in convincing readers that a study's methods offer valid means to represent the constructs of theoretical interest (Maupin, McCusker, Slaughter & Ruark, 2020). In our experience, many otherwise exciting manuscripts mobilizing big data came to falter on this front, typically because reviewers and editors expressed reservations with these studies' imprecise measures of their constructs of interest.

In a typical study mobilizing the computer-aided analysis of a very large corpus of documents, for instance, authors will sometimes use latent Dirichlet allocation (LDA)—a generative algorithmic technique resting on Bayesian network analysis—to automatically search and track the co-occurrence of specific words in these documents (see Hannigan et al., 2019; Taeuscher et al., 2021). By doing so, authors aim to uncover the most important "topics" emerging across these documents—as evidenced by a set of terms (whether individual words or groups thereof) that, taken together, reflect a common idea or theme. Building on the notion that the weighting of the co-occurring words distinguish each topic, authors must establish that, when taken together, the words identified by the algorithm as forming part of a "topic" offer valid representations of relevant constructs of interest. In similar fashion that a confirmatory factor analysis of responses to a series of survey items allows for assessing the extent to which these responses seem to "hang together" in a manner consistent with the intended measurement model, the coherence between the co-occurring words (with their associated weighting) allows for associating a vector of words in LDA with a topic of theoretical interest (or its variations).

As a challenging yet illustrative example, let us imagine a project in which a team of authors proposes to analyze the internal communications (e.g., emails and instant messaging) of a large sample of new ventures in order to examine the development of new organizational routines over a three-year period. To establish the validity of their topic observations, it would befall authors to explain how and why the different vectors of words identified by the algorithm uniquely capture different routines (e.g., internal team building and morale vs. business development) or the forms of these routines (e.g., simple or complex) at different stages of development (e.g., emergence, legitimation, institutionalization). In many such analyses, though, the word vectors identified are so broad, diverse, and overlapping with one another that it becomes difficult to convince reviewers that the obtained results uniquely capture the constructs and variations of interest. The challenges are exacerbated when the resulting word vectors appear very distant from extant conceptualizations of a topic of interest (like

with the concept of organizational routines above). In sum, establishing that the chosen measures and methods are rigorously aligned with the overarching theoretical frame forms a pivotal challenge to overcome for mobilizing new data formats and sources.

Minding the theory–method gap is especially pertinent in the world of big data, where potentially thousands of variables are available for analysis. Researchers adopting machine-learning software that sifts through these variables should not only provide comprehensive details on "feature importance scores" (to assess the relative importance of given variables in a model) but may also need to perform randomization tests or bootstrapping models to assess how preferred model specifications compare against alternatives (Sherman & Funder, 2009; Pargent, Schoedel & Stachl, 2023). A related issue concerns the choice of training datasets used for fine-tuning the machine-learning algorithms used to detect relevant patterns in the data. In supervised and semi-supervised forms of machine learning, one needs a training dataset of validly labeled cases, so that the algorithm can "learn" how best to classify different cases relevant for the target constructs of interest. For training a machine-learning classifier of skin lesion images (to eventually predict which one is most likely cancerous), for instance, traditional approaches call for "training" the algorithms with a large dataset of pre-labeled skin lesion images for which we already know which ones are associated with cancer (true positives) and which ones are not (true negatives). By designing a feedback-learning mechanism within the algorithm and letting it "train" itself with the labeled data, the algorithm can automatically "reinforce" the "heuristic routines" that it uses to correctly classify the images known to indicate cancer (or not), just as well as it can "downplay" the "heuristic routines" that it had "mistakenly" used to incorrectly classify as cancerous images known to be associated with healthy cells, and for incorrectly classifying as representing healthy cells those images known to be associated with cancer.

The same challenges of identifying relevant training datasets arise in management research leveraging machine-learning algorithms. Yet, in addition to the possibility of training an algorithm with the very same kind of data that one intends to analyze (as above, in a way akin to performing bootstrapping or split-sample analyses), one has the opportunity to train an algorithm with a "plausibly similar" dataset or to proceed with unsupervised learning techniques (akin to exploratory cluster or factor analysis). In the first case, the challenges inherent to leveraging a "plausibly similar" dataset for training purposes will be to convince reviewers (a) that the training dataset is sufficiently similar to that of the main study to generate an algorithm that will produce valid results for the target constructs of interest, and (b) that the main study dataset does not include idiosyncratic characteristics, endogenous dynamics, or exogenous differences that could unduly affect the algorithm's functioning to the point that it would yield biased results. By contrast, in the second case, the challenges inherent to leveraging exploratory approaches of unsupervised learning will likely be to convince reviewers that the obtained results can indeed be interpreted as rigorously representative of the claimed constructs and phenomena of theoretical interest (as with the LDA example above). Across both cases, our primary recommendation to authors is to be as transparent as possible about the reasons that motivated their initial choice of algorithmic approach (supervised, semi-supervised, unsupervised) and training dataset (when relevant, for supervised and semi-supervised learning). In turn, it will also befall them to transparently share relevant empirical observations to document the similarities across training datasets and between the raw data and their theoretical interpretation.

MOVING FORWARD WITH NEW DATA

Mobilizing new sources of data is challenging, yet it offers unique and exciting possibilities for advancing academic understanding of management and organizational phenomena. To help authors seize such opportunities, we summarize the gist of our recommendations in Table 1.

We began this editorial by highlighting that, even if it was not always or necessarily needed, mobilizing new data sources can sometimes call for also mobilizing new analytical techniques. We appreciate the extra challenges that innovating on both fronts might impose on authors. One might imagine that, when contemplating the daunting task of addressing reviewers' objections and additional requests on both fronts, many authors will prefer to shy away from pursuing projects that would require using new methods for analyzing data from new sources.

As a "broad tent" journal at the forefront of management and organization research innovations, however, AMJ benefits from a deep pool of expertise within its editorial team, editorial review board, and published authors to handle such innovative submissions. Already, we put in place an expanded panel of consulting methodological experts and we recently expanded our editorial team to handle

TABLE 1
Summary of Recommendations

#1	Take data context seriously	• Provide rich detail about the **assumptions** behind how, when, where, and why the data were originally compiled • Provide rich detail about the **motivations** and **incentives** that underpin the original provision of the data, and the implications for the nature and completeness of the data • Provide rich details about geographical, temporal, or other contextual factors to **situate the new data** in a specific empirical setting • Provide rich details about how sensing, recording, and tracking methodologies may bias or influence the observed behaviors and actions, and any precautions taken to **mitigate undue bias or influence**
#2	Transparency and ethical considerations	• Be transparent about **study design and hypotheses**; for example, through preregistration and making study materials and instructions available for blind peer review • Be transparent about **how data were accessed or generated**, providing details about data access agreements, ethical review, and consent procedures • Be transparent about **data inputs** into machine-learning models; for example, by "showing your data" in the paper and making anonymized data subsets available for blinded peer review • Be transparent about how experimental designs resolve any potential trade-offs in terms of precise causal inference with **ethical considerations** for participants
#3	Minding the alignment of theory and method	• Be vigilant about the captured data's representativeness relative to the broader concepts or phenomenon of interest, providing evidence of how empirical observations match theoretical concepts • Be vigilant about the **alignment** between contextual or generated labels and terms and the invoked theoretical constructs and mechanisms, illustrating in the paper how raw data maps onto final labels • Be vigilant about any differences in the nature of **data used for training and validation** in supervised machine learning vis-à-vis the data used for final analysis • Be vigilant about the results' sensitivity to choices made in terms of specific **training datasets and algorithms**, and discuss any biases these choices may introduce • Be vigilant about the **sensitivity** of **preferred data representations and model specifications** relative to alternative choices in the analysis of big data

methods-focused submissions that "broaden the repertoire" of management scholars (Gruber & Bliese, 2024; Langley et al., 2023). As AMJ's 23rd editorial team, we see ourselves a little like a team of venture capitalists willing to spend the effort, time, and editorial capital to commit pre-seed funds on highly promising submissions. As part of the transformational times theme that animates our team (Gruber, 2023), we hope that the few considerations discussed above will encourage and empower authors to seize the opportunities offered by new data sources!

Denis A. Grégoire
HEC Montréal

Anne L. J. Ter Wal
Imperial College London

Laura M. Little
University of Georgia

Sekou Bermiss
University of North Carolina

Reddi Kotha
Singapore Management University

Marc Gruber
École Polytechnique Fédérale de Lausanne

REFERENCES

Arslan, B., Vasudeva, G., & Hirsch, E. B. 2024. Public–private and private–private collaboration as pathways for socially beneficial innovation: Evidence from antimicrobial drug-development tasks. ***Academy of Management Journal***, 67: 554–582.

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. 2023. September 22: Which humans? ***PsyArXiv Preprints***. Retrieved from https://osf.io/preprints/psyarxiv/5b26t

Bermiss, Y. S., & McDonald, R. 2018. Ideological misfit? Political affiliation and employee departure in the private-equity industry. ***Academy of Management Journal***, 61: 2182–2209.

Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., & Larson, J. 2023. May 4: Synthetic replacements for human

survey data? The perils of large language models. ***SocArXiv Papers***. Retrieved from https://osf.io/preprints/socarxiv/5ecfa

Cao, R., Koning, R., & Nanda, R. 2023. December 14: Sampling bias in entrepreneurial experiments. ***Management Science***. Forthcoming.

Cenci, S., Burato, M., Rei, M., & Zollo, M. 2023. The alignment of companies' sustainability behavior and emissions with global climate targets. ***Nature Communications***, 14: 7831.

Chaffin, D., Heidl, R., Hollenbeck, J. R., Howe, M., Yu, A., Voorhees, C., & Calantone, R. 2017. The promise and perils of wearable sensors in organizational research. ***Organizational Research Methods***, 20: 3–31.

Chiang, E., de Rond, M., & Lok, J. 2023. July 6: Identity in a self-styled 'paedophile-hunting' group: A linguistic analysis of stance in Facebook group chats. ***Applied Linguistics***: amad034.

Criscuolo, P., Alexy, O., Sharapov, D., & Salter, A. 2019. Lifting the veil: Using a quasi-replication approach to assess sample selection bias in patent-based studies. ***Strategic Management Journal***, 40: 230–252.

Criscuolo, P., Salter, A., & ter Wal, A. L. 2014. Going underground: Bootlegging and individual innovative performance. ***Organization Science***, 25: 1287–1305.

DeCelles, K. A., Howard-Grenville, J., & Tihanyi, L. 2021. From the editors: Improving the transparency of empirical research published in AMJ. ***Academy of Management Journal***, 64: 1009–1015.

Dencker, J. C., Gruber, M., Miller, T., Rouse, E. D., & von Krogh, G. 2023. From the editors: Positioning research on novel phenomena: The winding road from periphery to core. ***Academy of Management Journal***, 66: 1295–1302.

de Rond, M., Holeman, I., & Howard-Grenville, J. 2019. Sensemaking from the body: An enactive ethnography of rowing the Amazon. ***Academy of Management Journal***, 62: 1961–1988.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. 2023. Can AI language models replace human participants? ***Trends in Cognitive Sciences***, 27: 597–600.

Dineen, B. R., & Allen, D. G. 2016. Third party employment branding: Human capital inflows and outflows following "best places to work" certifications. ***Academy of Management Journal***, 59: 90–112.

Dube, S., & Zhu, C. 2021. The disciplinary effect of social media: Evidence from firms' responses to Glassdoor reviews. ***Journal of Accounting Research***, 59: 1783–1825.

Dupree, C. H. 2024. Words of a leader: The importance of intersectionality for understanding women leaders' use of dominant language and how others receive it. ***Administrative Science Quarterly***. Forthcoming.

Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. 2020. Computational social science and sociology. ***Annual Review of Sociology***, 46: 61–81.

Grimes, M., von Krogh, G., Feuerriegel, S., Rink, F., & Gruber, M. 2023. From the editors: From scarcity to abundance: Scholars and scholarship in an age of generative artificial intelligence. ***Academy of Management Journal***, 66: 1617–1624.

Gruber, M. 2023. From the editors: An innovative journal during transformational times: Embarking on the 23rd editorial term. ***Academy of Management Journal***, 66: 1–8.

Gruber, M., & Bliese, P. 2024. From the editors: Expanding AMJ's manuscript portfolio: research methods articles designed to advance theory and span boundaries. ***Academy of Management Journal***, 67: 1–4.

Hannigan, T. R., Haans, R. F. J., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Kaplan, S., & Jennings, P. D. 2019. Topic modeling in management research: Rendering new theory from textual data. ***Academy of Management Annals***, 13: 586–632.

Harari, G. M., & Gosling, S. D. 2023. Understanding behaviors in context using mobile sensing. ***Nature Reviews Psychology***, 2: 767–779.

Hsieh, Y. Y., & Vergne, J. P. 2023. The future of the web? The coordination and early-stage growth of decentralized platforms. ***Strategic Management Journal***, 44: 829–857.

Hu, T., Kyrychenko, Y., Rathje, S., Collier, N., van der Linden, S., & Roozenbeek, J. 2023. October 24: Generative language models exhibit social identity biases. ***ArXiv Preprints***. Retrieved from https://arxiv.org/abs/2310.15819

Hubbard, T. D., & Aguinis, H. 2023. Conducting phenomenon-driven research using virtual reality and the metaverse. ***Academy of Management Discoveries***, 9: 408–415.

Kelan, E. K. 2021. July 26: Algorithmic inclusion: Shaping artificial intelligence in hiring. ***Academy of Management Proceedings***, 2021(1): 11338.

Kohavi, R., & Thomke, S. 2017. The surprising power of online experiments. ***Harvard Business Review***, 95(5): 74–82.

Langley, A., Bell, E., Bliese, P., LeBaron, C., & Gruber, M. 2023. From the editors: Opening up AMJ's research methods repertoire. ***Academy of Management Journal***, 66: 711–719.

Laureiro-Martínez, D., Brusoni, S., Canessa, N., & Zollo, M. 2015. Understanding the exploration–exploitation dilemma: An fMRI study of attention control and decision-making performance. ***Strategic Management Journal***, 36: 319–338.

Li, Y., & Khessina, O. M. 2024. Before birth: How provisional spaces shape the localized emergence of new organizational forms. ***Academy of Management Journal***, 67: 494–525.

Li, Z., Ning, H., Jing, F., & Lessani, M. N. 2024. Understanding the bias of mobile location data across spatial scales and over time: A comprehensive analysis of SafeGraph data in the United States. ***PLoS One***, 19: e0294430.

Lockwood, C., Glynn, M. A., & Giorgi, S. 2023. Polishing the gilt edge: Elite category endurance and symbolic boundaries in U.S. luxury hotels, 1790–2015. ***Academy of Management Journal***, 66: 9–42.

Lumineau, F., Wang, W., & Schilke, O. 2021. Blockchain governance: A new way of organizing collaborations? ***Organization Science***, 32: 500–521.

Matz, S. C. 2022. ***The psychology of technology: Social science research in the age of big data***. Washington, DC: American Psychological Association.

Maupin, C. K., McCusker, M. E., Slaughter, A. J., & Ruark, G. A. 2020. A tale of three approaches: Leveraging organizational discourse analysis, relational event modeling, and dynamic network analysis for collective leadership. ***Human Relations***, 73: 572–597.

McKenny, A. F., Aguinis, H., Short, J. C., & Anglin, A. H. 2018. What doesn't get measured does exist: Improving the accuracy of computer-aided text analysis. ***Journal of Management***, 44: 2909–2933.

O'Flaherty, K. 2024. February 2: Is Apple's new Vision Pro going to be a privacy nightmare? ***Forbes***.

Pargent, F., Schoedel, R., & Stachl, C. 2023. Best practices in supervised machine learning: A tutorial for psychologists. ***Advances in Methods and Practices in Psychological Science***, 6: 1–35.

Polzer, J. T. 2022. The rise of people analytics and the future of organizational research. ***Research in Organizational Behavior***, 42: 100181.

Rahman, H. A. 2024. March 10: Unethical online experiments risk real world harm. ***Financial Times***.

Rahman, H. A., Weiss, T., & Karunakaran, A. 2023. The experimental hand: How platform-based experimentation reconfigures worker autonomy. ***Academy of Management Journal***, 66: 1803–1830.

Raisch, S., & Krakowski, S. 2021. Artificial intelligence and management: The automation–augmentation paradox. ***Academy of Management Review***, 46: 192–210.

Rhee, L. 2024. CEO attentional vigilance: Behavioral implications for the pursuit of exploration. ***Academy of Management Journal***. Forthcoming.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., … & McLanahan, S. 2020. Measuring the predictability of life outcomes with a scientific mass collaboration. ***Proceedings of the National Academy of Sciences of the United States of America***, 117: 8398–8403.

Sherman, R. A., & Funder, D. C. 2009. Evaluating correlations in studies of personality and behavior: Beyond the number of significant findings to be expected by chance. ***Journal of Research in Personality***, 43: 1053–1063.

Srivastava, S. B., Goldberg, A., Manian, V. G., & Potts, C. 2018. Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations. ***Management Science***, 64: 1348–1364.

Sust, L., Talaifar, S., & Stachl, C. 2023. Mobile application usage in psychological research. In M. R. Mehl, M. Eid, C. Wrzus, G. M. Harari, & U. W. Ebner-Primer (Eds.), ***Mobile sensing in psychology: Methods and applications***: 184–213. New York: Guilford Press.

Taeuscher, K., Bouncken, R., & Pesch, R. 2021. Gaining legitimacy by being different: Optimal distinctiveness in crowdfunding platforms. ***Academy of Management Journal***, 64: 149–179.

Veltrop, D. B., Bezemer, P. J., Nicholson, G., & Pugliese, A. 2021. Too unsafe to monitor? How board–CEO cognitive conflict and chair leadership shape outside director monitoring. ***Academy of Management Journal***, 64: 207–234.

von Krogh, G., Roberson, Q., & Gruber, M. 2023. From the editors: Recognizing and utilizing novel research opportunities with artificial intelligence. ***Academy of Management Journal***, 66: 367–373.

Yarkoni, T., & Westfall, J. 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. ***Perspectives on Psychological Science***, 12: 1100–1122.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. 2019. Making sense of recommendations. ***Journal of Behavioral Decision Making***, 32: 403–414.