Original Article

The Datafied Society: Challenges and Strategies in Big Data Research for Social Sciences and Humanities

Ebrahim Mohseni Ahooei

(Received 23 October 2023; accepted 25 June 2024)

Abstract

The advent of big data marks a profound shift in our epistemological framework, introducing a new knowledge paradigm where the social landscape is shaped by data processing, perceived as both comprehensive and natural. This transformative shift challenges traditional notions of human agency in societal understanding, positioning empirical quantification at the forefront of inquiry. Beyond philosophical implications, pragmatic challenges abound in big data research—from issues of commensuration and the influence of action grammars to the dominance of correlational over causal relationships, the prevalence of everyday data over historical archives, and the pervasive impact of algorithms on data ecosystems. This manuscript undertakes a comprehensive exploration of these challenges, proposing strategies for navigating them within emerging disciplines such as Digital Humanities, Social Computing, and Cultural Analysis. Methodologically anchored in constructivist principles and critical discourse analysis (CDA), the study investigates how socio-cultural contexts shape data and knowledge production. Drawing on extensive literature and meta-analyses, it synthesizes diverse perspectives to underscore the necessity for methodological innovation and reflexivity in addressing the complexities of big data research, ensuring the integrity and depth of social inquiry amidst evolving data-driven methodologies.

Keywords: big data, cultural analysis, datafied society, digital humanities, social computing.

Ebrahim Mohseni Ahooei: PhD in Communication and New Media studies, University of Vienna; Member of the Executive Committee of the UNESCO Chair in Cyberspace and Culture (Email: emohseni@ut.ac.ir; ORCID: https://www.orcid.org/0000-0003-4468-3571)

Introduction

For millennia, humans have used data to comprehend, evaluate, analyze, calculate, and define reality. From the recording of agricultural data, commercial transactions, and "administrative details on clay tablets by the Sumerians in 4000 BC" (Yeo, 2021: 17) to the "emergence of statistics as a formal discipline in the 19th century" (Porter, 2020: 193), data has consistently played a pivotal role in chronicling historical events, census figures, legal codes, and more. Indeed, the principles governing the nature of data and the logic of its relationships have been integral to human civilization.

As we move from the era of the internet of information to the age of the internet of data (Mohseni Ahooei, 2022; 2023) and witness the rise of the datafied society (Van & Schäfer, 2017), a new paradigm is emerging across all fields of science, including the humanities. This paradigm, driven by the phenomenon of big data, heralds what has been termed the "end of theory" (Anderson, 2008), where "the numbers speak for themselves" (Jablonka & Bergsten, 2021), and correlation is often equated with causation (Pietsch, 2021).

This paradigm shift is grounded in "a new order of knowledge" (Couldry, 2014; 883), where big data becomes a transformative force in human self-understanding, offering a purportedly superior form of intelligence and providing true, objective, and precise knowledge. This approach promises insights that were previously unattainable, potentially rendering obsolete past efforts to understand humanity through localized interpretations.

In the current era dominated by big data, it is imperative for academics, especially within the humanities and arts, to develop proficiency in data research methodologies. This skill set is essential for participating in significant public discussions about data science, which encompass topics such as accountability in data creation and application, ethical considerations, privacy concerns, the influence of data, and transparency. Furthermore, researchers must reassess their impact on public dialogue and policy formulation. For students, this means becoming discerning data practitioners capable of questioning the misconceptions about a data-driven society and interacting with data-centric practices outside the academic sphere.

From a rigorous scientific methodology standpoint, the fundamental inquiry pertains to the extent to which the evolving landscape of knowledge, influenced by the advent of big data, has substantiated its purported advancements, particularly within the realms of humanities and social sciences. Addressing this query necessitates demystifying

big data and assessing its analytical and descriptive capacities through the lens of established scientific methodology and statistical principles. Therefore, this article aims to scrutinize critical challenges posed by big data in social research and propose alternative approaches conducive to more effective scientific endeavors during this era dominated by big data.

To this end, the initial focus of this manuscript will be on delineating the characteristics that define the envisioned new knowledge paradigm. Emphasis will be placed on its phenomenological attributes and its epistemological departure from classical paradigms. Subsequently, a critical examination will be undertaken to evaluate key assertions regarding the efficacy of big data in knowledge generation. Finally, the discourse will pivot towards exploring more reasoned and viable pathways for scientific inquiry amid the prevalence of big data.

The New Order of Knowledge

The advent of big data has not merely introduced a new tool for knowledge development; it has ushered in a paradigm shift in the very nature of knowledge itself. As Van and Schäfer aptly observe, "Data have become ontological and epistemological objects of research – manifestations of social interaction and cultural production" (Van & Schäfer, 2017: 11).

The new order of knowledge, ontologically speaking, has heralded a fundamental shift in our understanding of knowledge, challenging the long-held notion of Kantian rationality as the bedrock of all scientific inquiry. This new paradigm, driven by computational power and data-driven methodologies, has given rise to an "ontological epoch" characterized by "destablising amounts of knowledge and information that lack the regulating force of philosophy" (Berry, 2011: 8). As Berry suggests, computationality has emerged as a new "ontotheology", shaping our perception of reality and establishing a new framework by "creating a new ontological epoch as a new historical constellation of intelligibility" (ibid: 12).

Referring to the famous quote of Karl Marx, "Philosophers have hitherto only interpreted the world in various ways; the point is to change it" (Marx, 1932[1845]: 123), this very new order is both responsible for imagining the world in a new way and creating the world in a novel sense. As Anderson states, "This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and

measure it with unprecedented fidelity. With enough data, the numbers speak for themselves" (Anderson, 2008).

The advent of big data has not only transformed the way we produce knowledge, but also brought about a fundamental shift in our understanding of human nature and agency. This new paradigm, characterized by its emphasis on computational methods and large-scale data analysis, often assumes a generic human nature, suggesting that the study of any individual can be generalized to the entire human population. This assumption, however, overlooks the inherent uniqueness and self-interpreting nature of human beings, as eloquently expressed by Taylor and Smith, "The human being is a self-interpreting animal condemned to meaning" (Taylor, 1986: 52; Smith, 2010: 126). Indeed, the very essence of our shared existence lies in our ongoing efforts to interpret and make sense of our interactions with one another.

In cognitive science, it is believed that research on neural pathways in one person can provide insights into common neural structures and functions across all humans (Kandel et al., 2013; Sporns et al., 2008). Similarly, in psychology, studying decision-making processes in a controlled group reveals universal patterns applicable to broader populations (Tversky & Kahneman, 1974). This approach aligns with trends in big data and machine learning, where large datasets are analyzed to identify generalizable patterns and behaviors, ignoring the attention to individual differences. In this way, the study of human cognition and behavior becomes just systematic and objective, paralleling methodologies in the natural sciences.

The second premise of the new order of knowledge is the reification of social processes and existences. The rise of big data has not only transformed our approach to knowledge production, but has also introduced a new set of assumptions about the nature of social reality. The collection and processing of vast amounts of data, often treated as objective "facts". have led to the reification of social processes and existences. This tendency to objectify social phenomena has resulted in a disregard for the interpretive and contextual dimensions of human behavior and social practices.

As Mayer-Schönberger and Cukier aptly observe, "We will no longer regard our world as a string of happenings that we explain as a natural or social phenomenon, but as a universe comprised essentially of information" (Mayer-Schönberger & Cukier, 2013: 96). This emphasis on information, or more accurately, on data, as the fundamental building block of reality overlooks the inherent complexity and subjectivity of social life.

The third premise of the new order of knowledge is the interpretability of social phenomena. In line with the methodologies of the natural sciences, the humanities and social sciences now aim to accurately interpret the relationships between variables and predict behavioral patterns. This shift signifies a movement towards a systematic approach to understanding social matters, recognizing the social realm "provable".

In economics, the use of econometric models enables researchers to analyze large datasets and predict economic trends with remarkable precision. The famous example of the Phillips Curve illustrates this point. By examining the relationship between inflation and unemployment, economists can interpret and predict economic conditions, providing a concrete example of how social phenomena can be measured and analyzed quantitatively (Gayo-Avello et al., 2011). Similarly, techniques such as network analysis enable social scientists to uncover complex relationships and predict patterns with unprecedented accuracy. Social media analysis, for example, can predict public opinion trends and even election results by interpreting vast amounts of user-generated data (Greene, 2003). This paradigm shift indicates that the focus of the human sciences is no longer merely to approach and describe human phenomena, but to rigorously prove and predict the patterns underlying these phenomena.

Following this new approach, a wide set of justifying concepts appeared under the influence of the new worldview. The first one is "posthumanism" (Haraway, 2011; Hayles, 2000). It seeks to decenter humanity, arguing that human identity and significance are not intrinsic but contingent upon our interactions with other species, machines, and the environment. This perspective diminishes the traditional, exceptionalist view of human beings as the pinnacle of evolution or the primary agents of meaning. By recognizing humans as entities without inherent supremacy, posthumanism redefines the human condition, emphasizing our interconnectedness and interdependence with all forms of life and matter, thus stripping humanity of its unique, semantic pedestal.

Subsequently, the actor-network theory (ANT) (Latour, 2007) emerged as a framework that considers objects, ideas, processes, and other related factors to be as vital as humans in shaping social situations. This theory serves as both a theoretical and methodological approach to social theory, asserting that all entities in the social and natural realms exist within constantly evolving networks of relationships. According to the ANT, all elements in a social situation hold equal importance.

The object-oriented ontology (Harman, 2018) proposes a framework where all entities, whether human, non-human, animate,

or inanimate, possess equal ontological significance. This perspective aligns with contemporary scientific approaches that emphasize the interconnectedness and agency of diverse objects within complex systems and networks.

These new approaches indicate a paradigm shift in the academy and the emergence of systemic alternatives to human-centered culture. This shift not only redefines human identity and its place in the world but also fosters a more inclusive consideration. Through this lens, the academy is embracing a more systemic approach to knowledge and culture.

Method

The methodological approach of this study is anchored in a critical examination of the ontological and epistemological shifts precipitated by the advent of big data in social sciences. Recognizing the transformation from traditional paradigms to a data-driven knowledge order, this research adopts a constructivist epistemology, which posits that knowledge and data are inherently contextual and constructed by human agents (Crotty, 1998). This perspective is crucial in addressing the decontextualization crisis, where data is often stripped of its contextual meaning, leading to potential misinterpretations and oversimplifications of social phenomena (Boyd & Crawford, 2012).

To navigate the complexities associated with big data, the study employs a qualitative approach. This strategy ensures a comprehensive analysis of the phenomena under investigation, leveraging the strengths of methodological tradition which involves an in-depth critical discourse analysis (CDA) of the narratives and contexts surrounding data production and utilization. CDA is particularly suited for examining how language and power dynamics influence the construction and interpretation of data, providing insights into the socio-cultural underpinnings of big data practices (Fairclough, 2013). This approach is instrumental in uncovering the implicit assumptions and biases that shape data-centric knowledge production.

To ensure methodological rigor and transparency, the study draws on a diverse range of data sources. These include extensive literature reviews encompassing theoretical discussions and empirical studies on big data's implications for social sciences. The selection criteria prioritize relevance, credibility (preferably peer-reviewed sources), and diversity of perspectives to capture a comprehensive spectrum of opinions and findings.

Additionally, the study conducts meta-analyses of existing research to deepen insights into big data methodologies and their socio-cultural

impacts. This involves synthesizing and critically evaluating studies that examine the application and implications of big data in social research. The analytical procedures are guided by principles of contextual integrity and commensuration. Contextual integrity, as proposed by Nissenbaum (2011), underscores the importance of preserving the context within which data is generated and interpreted, thus maintaining its intrinsic meaning and relevance amidst aggregation and analysis processes. The analytical framework of this study is rooted in the principles of contextual integrity and commensuration. Contextual integrity, as proposed by Nissenbaum (2011), emphasizes the importance of preserving the context within which data is generated and interpreted. This concept is pivotal in addressing the decontextualization crisis, ensuring that data retains its intrinsic meaning and relevance. Commensuration, defined as the transformation of different qualities into a common metric (Espeland & Stevens, 1998), is critically examined to highlight its potential pitfalls in homogenizing diverse social phenomena. By interrogating these processes, the study aims to reveal the limitations and biases inherent in big data methodologies.

By employing these methodological approaches, the study aims to provide a nuanced and critical examination of big data's role in social research. This holistic approach enhances the transparency, reproducibility, and theoretical grounding of the research findings, enabling readers to comprehensively evaluate its contributions to understanding contemporary socio-technological dynamics influenced by big data.

Demystifying the Big Data: Pitfalls of Data-Driven Social Research

Despite the prevailing hype surrounding the transformative potential of big data in knowledge production, a critical examination of both big data analysis processes and research methodologies reveals a multitude of shortcomings. These limitations are so profound that they call into question the validity of many social research studies conducted using big data approaches.

The pervasive adoption of big data in social research has sparked a lively debate, with a spectrum of perspectives ranging from radical critiques to enthusiastic endorsements. I aim to present a comprehensive overview of the key criticisms leveled against big data research methodologies and findings, categorizing them along a continuum of critical stances.

At the extreme end of the spectrum lies Couldry's (2014) philosophically grounded critique, which challenges the very epistemological foundations of big data-driven knowledge production

in cultural research. Couldry argues that big data research cannot replace the rigor and depth of traditional qualitative studies. Moving towards the center of the spectrum, we encounter perspectives like that of Boyd and Crawford (2012), who acknowledge the inherent challenges of big data research but ultimately maintain that these challenges can be overcome through methodological advancements and careful consideration of ethical implications. At the opposite end of the spectrum, we find proponents of big data who view the criticisms as misguided and rooted in a misunderstanding of the nature and potential of big data. These scholars, such as Resnyansky (2019), argue that the focus should be on adapting and refining "social scientific theories and conceptual frameworks that may inform the analysis of the social in the age of Big Data."

While I refrain from making definitive pronouncements on the merits or demerits of big data in social research, I emphasize the importance of critically evaluating the methodological pitfalls and potential invalidity of big data-driven research. Ultimately, the decision to employ big data approaches lies with individual researchers. However, it is crucial to be aware of the inherent limitations and potential biases associated with big data analysis to ensure the integrity and validity of research findings.

In my articulation of the shortcomings of big data research, I have drawn heavily from my personal research experiences. This has inevitably led to the inclusion of more specialized discussions that may pose challenges for readers unfamiliar with advanced social research methodologies. While these sections may not be essential for non-specialists, I have made a conscious effort to incorporate such complex analyses to enhance the practical value of this work for those deeply engaged in the field. Conversely, I have intentionally refrained from dwelling on widely acknowledged challenges that are already familiar to those interested in big data research.

The Objectivity Crisis

The debate surrounding objectivity and subjectivity in research is not a new one. Immanuel Kant is often credited with initiating this discourse by introducing the concept of "the thing-in-itself (Ding an sich)" (Kant, 1908[1781]: 26), emphasizing the importance of recognizing the inherent objectivity of reality. In the realm of social sciences, Durkheim (1982[1895]) made a significant contribution by establishing objectivity as a fundamental principle of sociological methodology. Since then, the pursuit of objectivity and the critiques of its limitations have remained central to both the justification and evaluation of scientific research.

Epistemological perspectives underscore the inherent subjectivity and contextuality of data, arguing that it cannot be separated from the knowledge systems in which it is embedded. Constructivist epistemologies posit that all knowledge, including data, is constructed by human beings and is thus inherently subjective and contextual (Crotty, 1998). Phenomenological approaches further emphasize that data is always perceived and interpreted through the lens of human experience and consciousness, highlighting the role of individual perception in shaping understanding (Moran, 2002).

Contextual dependence also plays a crucial role in the interpretation and meaning of data. Cultural and social contexts influence data collected in different settings, as seen in the varying implications of income data across different economic environments (Geertz, 1973).

Subjectivity in data collection is another significant factor, as the choices and biases of those who collect data inherently influence the results. Researchers' decisions on what variables to measure and how to measure them frame the data within their theoretical perspectives (Kuhn, 1962). Additionally, the measurement tools used, whether qualitative or quantitative, can affect outcomes.

With the advent of big data, the question of objectivity has once again taken center stage in scientific research. The inherent challenges of maintaining objectivity in big data research cast a shadow over all stages of the research process, from data collection to the "discovery" of correlational relationships.

Moving beyond technical considerations and adopting a phenomenological lens, it becomes evident that the concept of "raw data" in the context of big data is an oxymoron (Gitelman, 2011). Even before data collection, there are pre-existing frameworks and constructs that shape what data is deemed worth collecting. These frameworks are informed by cultural, social, and scientific norms, which influence the way data is framed and understood (Bowker & Star, 2000). As Badiou argues, "What counts, in the sense of what is valued – is that which is counted. Conversely, everything that can be numbered must be valued" (Badiou, 2008: 1). As such, data cannot be considered raw because it is always already shaped by these constructs. This issue has far wider implications in the field of big data.

The mediation of data collection and processing by technology introduces many layers of orientation. Software algorithms can introduce biases through the data processing and analysis stages (Couldry, 2020). Machine learning models are trained on specific datasets that reflect the orientations and biases inherent in those datasets, often leading to

skewed outcomes (Pessach & Shmueli, 2022). Similarly, the way data is stored, indexed, and retrieved can significantly impact its accessibility and usability, which in turn shapes its interpretation (Berman, 2013), therefore data cleaning process is necessarily biased by some subjective filter (Bollier, 2010).

The "Big" Ideology Crisis

The acceptance of big data is mainly justified by the ideology of "the bigger, the better" (see, for example, Mayer-Schönberger & Cukier, 2013; Kitchin, 2014). But this ideology in the field of scientific research is just a myth. Among the many arguments that have addressed the challenges of big data, I will address four of the more pernicious, yet more technical ones.

One of the fundamental claims of big data is that its sheer volume makes it a more accurate representation of society compared to traditional sampling methods. However, for several reasons, big data does not reflect the entirety of society. Big data research often treats all data points from social media platforms equally, regardless of their diverse functions and usage patterns among different social groups. This homogenization of data points overlooks the heterogeneous nature of user actions and interpretations, leading to oversimplified and potentially misleading analyses. Moreover, big data often fails to account for the contextual factors and socio-economic disparities that influence individual behavior and online interactions. By treating all data points as equally representative, big data analyses risk perpetuating biases and misrepresenting the nuances of social dynamics. In addition, the inherent biases and limitations of the algorithms used to collect and process big data can further distort the accuracy of its representation of society. These algorithms may reflect the biases of their creators or the underlying data they are trained on, leading to systematic misinterpretations and perpetuation of stereotypes.

Drawing generalizations about the global population based on Twitter data, for example, is fraught with limitations. The Twitter user base does not accurately represent the world's demographics, and equating accounts with individual users is a flawed assumption. The prevalence of multiple accounts per user and the shared use of single accounts complicate the assessment of individual behavior and preferences. This heterogeneity in account usage patterns means that each tweet cannot be directly attributed to a single unique individual and a significant portion of the population does not engage with Twitter at all, limiting the platform's ability to capture a comprehensive picture

of global perspectives and opinions. Those who do not use Twitter may have different viewpoints and experiences than those who are active on the platform.

User activities on social media platforms, recorded via social media clients, cross-platform integrations, and automated software, are frequently converted into standardized metrics. This quantification process allows for the aggregation of diverse actions into singular data points, thereby concealing the varied interpretations and practices that exist beneath the surface. Optimization algorithms, including genetic algorithms and simulated annealing, further shape the data and determine the importance assigned to its different elements (Boyd & Crawford, 2012; Gillespie, 2014).

The notion that an increase in the volume of data inherently ensures representativeness is fundamentally flawed due to the intricacies of advanced statistical methodology and the profound challenges posed by high-dimensional data analysis. In statistical parlance, the assumption of representativeness is often scrutinized through the lens of sampling bias and non-probability sampling (Lohr, 2021). Traditional probability sampling techniques, underpinned by the Central Limit Theorem, assure that

$$\sqrt{n}(\bar{x} - \mu) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma^2),$$

where n is the sample size, the sample mean, and μ the population mean (Ziegel, 2002: 408). However, big data typically eschews these methodologies, favoring convenience sampling, thus violating the assumption $P(Xi) > 0$ for all i, leading to substantial undercoverage and selection bias (Meng, 2018).

Furthermore, the presence of heteroskedasticity, characterized by $Var(\epsilon_i | x_i) = \sigma_i^2$, disrupts the homoscedasticity assumption of classical linear models. The Generalized Least Squares (GLS) method attempts to rectify this by transforming the data via $\hat{\beta} GLS = (x' \Omega^{-1} x)^{-1} x' \Omega_y^{-1}$, where Ω represents the covariance matrix of the error terms (Greene, 2003). Despite this, estimating Ω accurately in large datasets is computationally intensive and often infeasible, leading to biased parameter estimates and inefficiency (Hansen, 2022).

Compounding these issues is the curse of dimensionality, a phenomenon where the feature space increases exponentially with the number of dimensions, denoted as $O(n^p)$, where n is the number of observations and p the number of features (Bellman, 1961). This exponential growth renders many traditional statistical techniques, such as Maximum Likelihood Estimation (MLE), impractical due to the sparsity of the data (Fan & Li, 2006). Consider the K-nearest neighbors (KNN) algorithm, where the distance metric

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

becomes less informative as p increases, leading to the "empty space" phenomenon and high variance in model predictions (Beyer et al., 1999). Moreover, Simpson's paradox further complicates the interpretation of aggregated data, where an observed relationship within the aggregate data, $\Pr(Y \mid A) \neq \Pr(Y \mid A, B)$, can be reversed when disaggregated, indicating a severe misrepresentation if subgroup heterogeneity is ignored (Pearl, 2009).

In the realm of classification problems, Imbalanced Classes pose significant challenges. The skewness in class distribution, where $n_1 > n_2$ (with n_1 and n_2 representing the sizes of the majority and minority classes, respectively), biases the classifier towards the majority class (He & Garcia, 2009). Techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and the use of cost-sensitive learning algorithms $L = \sum_{i=1}^{n} \omega_i L(y_i, \hat{y}_i)$, where ω_i adjusts the weight for minority class samples are employed to mitigate this bias, yet they introduce additional complexity and require meticulous parameter tuning (Chawla et al., 2002).

Missing data mechanisms, categorized as MCAR, MAR, and MNAR, present further complications. The MCAR assumption, $P(M_i = 1) = P(M_i = 1 \mid X)$, rarely holds in practical scenarios, leading to biased estimates when applying techniques like Expectation-Maximization (EM) for data imputation (Little & Rubin, 2019). For MNAR data, where $P(M_i = X,Y)$, advanced models like Heckman's two-step correction $y_i^* = X_i\beta + \epsilon_i$; observe $y_i = y_i^*$ if $Z_{iy} + v_i > 0$ must be employed, significantly complicating the modeling process and often relying on unverifiable assumptions (Heckman, 2013).

A second critical challenge arises from the misconception that correlational relationships in big data represent genuine causal connections. As I have previously noted, the new epistemological order shaped by big data often equates correlations with causal relationships (Pietsch, 2021). However, as we will further explore, even the purported correlations identified in big data analyses are not always real.

This assertion can be rigorously substantiated by delving into the complex interplay between high-dimensional statistical theory and the inherent limitations of traditional methodologies when scaled to the context of big data. First, consider the pervasive issue of dimensionality reduction, where techniques like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (*t*-SNE) are often

employed. The eigenvalue decomposition in PCA, denoted as $X = U\Sigma V^T$ where $X \in R^{n \times p}$, $U \in R^{n \times n}$, $\Sigma \in R^{n \times p}$, and $V \in R^{p \times p}$, becomes computationally infeasible and loses interpretability as p grows exponentially relative to n (Johnstone & Lu, 2009).

Over and above that, the stochastic nature of t-SNE, governed by the joint probabilities

$$P_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma^2\right)}{\Sigma_{k \neq 1} \exp(-\|x_k - x_L\|^2 \sigma^2)},$$

introduces significant instability and sensitivity to parameter tuning, which is exacerbated in high-dimensional settings, rendering the derived correlation structures highly unreliable. Moreover, the phenomenon of spurious correlations in high-dimensional data is particularly pernicious. Given p variables, the number of pairwise correlations is

$$\binom{p}{2} = \frac{p(P-1)}{2},$$

leading to a combinatorial explosion of hypothesis tests. Under the null hypothesis of no association, the probability of observing at least one significant correlation purely by chance is approximated by

$$1 - (1 - \alpha)^{\binom{p}{2}},$$

where α is the significance level (Fan, 2008). This probability rapidly approaches 1 as p increases, necessitating the use of stringent multiple testing corrections like the Benjamini-Hochberg procedure or the Bonferroni correction, both of which suffer from severe power limitations in high-dimensional contexts (Benjamini & Hochberg, 1995).

Additionally, the issue of noise accumulation, quantified through the spectral norm of the noise matrix $||E||_2$, where E represents random noise, results in the contamination of the signal, leading to inflated eigenvalues and distorted principal components. This is encapsulated in the eigenvalue perturbation theory, where $\Delta\lambda_i \leq ||E||_2$ for the i-th eigenvalue, highlighting the sensitivity of high-dimensional data to noise and further undermining the validity of inferred correlations (Johnstone & Lu, 2009). High-dimensional regression models, such as Least Absolute Shrinkage and Selection Operator (LASSO) and Ridge Regression, introduce regularization to mitigate overfitting. LASSO, which solves

$$\min_{\beta}(||y - X\beta + \lambda||\beta||_1),$$

imposes sparsity by shrinking coefficients to zero, yet its performance is highly dependent on the choice of the regularization parameter λ, typically selected via cross-validation (Tibshirani, 1996). In the context of big data, the curse of dimensionality exacerbates the variance-bias tradeoff, often leading to unstable and non-generalizable models.

Fur thermore, the presence of multicollinearity, quantified by the condition number $\kappa(X)=\sigma_{max}(X)/\sigma_{min}(X)$', inflates the variance of coefficient estimates, rendering traditional correlation metrics ina dequate. Consider the intricacies of the high-dimensional Gaussian graphical models (GGM), where the precision matrix $\Theta=\Sigma^{-1}$ is estimated to infer the conditional independence structure among variables. The gra phical lasso, which maximizes the penalized log-likelihood log det $(\Theta) - tr(S\Theta) - \lambda||\Theta||_1$, is computationally intensive and sensitive to the penalty parameter λ. Moreover, the sparsity pattern of Θ is highly sensitive to the sample size and the underlying distribution of the data, oft en leading to erroneous inferences about the network structure in hig h-dimensional regimes (Koller & Friedman, 2009). The temporal dynamics inherent in big data necessitate advanced time series models such as Vector Autoregression (VAR) and State Space Models (SSM). The VAR model, expressed as

$$y_t = \Sigma_{i=1}^{p} A_i y_{t-i} + u_t,$$

whe re y_t is a vector of observations and A_i are coefficient matrices, encounters significant challenges in high-dimensional settings due to the proliferation of parameters, leading to overfitting and unstable forecasts. State Space Models, defined by the observation equation $y_t = Z_t\alpha_t + \varepsilon_t$ and the state equation $\alpha_{t+1} = T_t\alpha_t + \eta_t$, where α_t represents the latent states, involve complex estimation procedures such as the Kalman filter, which bec ome computationally prohibitive as the state dimension increases (Durbin & Koopman, 2012).

The third challenge lies in the low signal-to-noise ratio inherent in big data, which significantly compromises the reliability of inferential conclusions. This low signal-to-noise ratio can be dissected through the lens of advanced statistical techniques and methodological frameworks, rev ealing a plethora of complexities that render the extraction of mea ningful information exceedingly difficult.

Fir stly, consider again the concept of the curse of dimensionality, whi ch exacerbates the noise accumulation effect. In high-dimensional spa ces, the Euclidean distance

$$d(x_i, x_j)= \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2}$$

becomes increasingly dominated by noise as the number of dimensions p increases (Bellman, 1961). This phenomenon leads to the concentration of distances, where the relative difference between the nearest and farthest neighbor diminishes, making it challenging to distinguish signal fro m noise. Theoretical underpinnings from random matrix theory

further illustrate this, where the eigenvalues of the sample covariance matrix

$$S = \frac{1}{n} X^T X$$

converge to the Marčenko-Pastur distribution as $p/n \to c$, with $c > 0$, indicating that the largest eigenvalues (representing signal) are not significantly separated from the bulk of smaller eigenvalues (representing noise) (Marchenko & Pastur, 1967).

Moreover, the application of high-dimensional statistical methods such as Principal Component Analysis (PCA) and its variants encounter significant challenges. In the high-dimensional regime, the principal components themselves become noisy. Specifically, Johnstone and Lu (2009) demonstrate that the empirical principal components deviate substantially from the population principal components, leading to distorted signal extraction. This is exacerbated by the presence of spiked eigenvalues, where the top eigenvalues of the covariance matrix are inflated due to noise, further obfuscating the true signal structure (Johnstone, 2001).

Another critical issue is the application of regularization techniques such as LASSO and Ridge Regression in high-dimensional settings. While these methods aim to mitigate overfitting by imposing sparsity or penalizing large coefficients, they are highly sensitive to the tuning parameters λ. The choice of λ significantly impacts the bias-variance tradeoff, and in the context of big data, the optimal λ is often difficult to estimate accurately due to the low signal-to-noise ratio. This sensitivity can lead to model instability and poor generalizability (Tibshirani, 1996).

Furthermore, consider the phenomenon of multiple hypothesis testing in high-dimensional data. With the vast number of variables *p*, the likelihood of encountering spurious correlations increases dramatically. The family-wise error rate (FWER) and the false discovery rate (FDR) must be controlled using methods such as the Bonferroni correction or the Benjamini-Hochberg procedure. However, these corrections introduce their own set of problems, such as reduced statistical power and increased Type II errors, which further diminish the signal-to-noise ratio by making it harder to detect true associations (Benjamini & Hochberg, 1995).

Additionally, the notion of intrinsic dimensionality, which refers to the minimal number of parameters needed to accurately describe the data structure, often reveals that many big data sets are intrinsically low-dimensional. This implies that the high-dimensional representation contains a significant amount of redundant and noisy information,

further lowering the signal-to-noise ratio. Techniques such as manifold learning, including Isomap and Locally Linear Embedding (LLE), attempt to uncover the low-dimensional manifold. However, these methods are computationally intensive and sensitive to the choice of parameters, often leading to unstable and non-robust embeddings (Tenenbaum et al., 2000).

While big data analysts often emphasize statistically significant correlations within data, Leinweber (2007) cautions against interpreting these correlations as anything more than "apophenia"– the tendency to perceive patterns and meaning in random or unrelated data. Therefore, as the third fatal challenge, the propensity of big data to overfit is a deeply entrenched issue that arises from the complex interplay between model complexity, high-dimensionality, and the limitations of traditional validation techniques in adequately assessing model performance. Overfitting occurs when a model captures the noise and idiosyncrasies in the training data rather than the underlying signal, leading to poor generalizability to new, unseen data.

To understand this phenomenon, consider the Vapnik-Chervonenkis (VC) dimension, a fundamental concept in statistical learning theory which measures the capacity of a model to fit a variety of functions (Vapnik, 2013). In high-dimensional settings typical of big data, the VC dimension tends to be extremely large, reflecting the model's ability to fit an enormous number of configurations. This high capacity, while potentially allowing for very accurate fitting of the training data, increases the risk of overfitting, as the model may essentially memorize the training data rather than learning a generalizable pattern.

The bias-variance tradeoff further elucidates the overfitting problem. In high-dimensional data, models with high complexity (low bias) tend to have high variance, which manifests as sensitivity to the specific training data points. This is mathematically described by the decomposition of the mean squared error (MSE) into bias, variance, and irreducible error components (Hastie et al., 2009):

$$\text{MSE} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}.$$

In the context of big data, where the dimensionality p can vastly exceed the number of observations n, the variance term can dominate, leading to models that perform exceedingly well on the training data but fail to generalize to new data.

Regularization techniques, such as LASSO and Ridge Regression, aim to mitigate overfitting by introducing a penalty term to the loss function. LASSO, for instance, adds an ℓ_1 penalty to the sum of absolute coefficients (Tibshirani, 1996). However, in the context of big data, the choice of the

regularization parameter λ is crucial and often determined via cross-validation. Cross-validation itself becomes computationally intensive and prone to overfitting if not properly managed, as the validation sets may not fully represent the variability in the data.

Another sophisticated approach to combat overfitting in big data is the use of ensemble methods, such as Random Forests and Gradient Boosting Machines. These methods build multiple models and aggregate their predictions to improve generalizability. The Random Forest algorithm, which constructs multiple decision trees and averages their predictions, theoretically reduces overfitting by decorrelating the individual trees through random feature selection (Breiman, 2001). However, in high-dimensional settings, the trees themselves can become overly complex, and the aggregation may still suffer from high variance due to the underlying data noise.

The concept of model selection criteria, such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), also plays a pivotal role in addressing overfitting. These criteria introduce a penalty for model complexity, aiming to balance the goodness-of-fit with model simplicity: AIC = $2k - 2\log(L)$ and BIC = $\log(n)k - 2\log(L)$, where k is the number of parameters and L is the likelihood function (Akaike, 1974; Schwarz, 1978). In the context of big data, however, the computation of these criteria can become infeasible due to the high dimensionality, and their effectiveness diminishes as the parameter space grows exponentially.

Moreover, the stability selection method, which combines subsampling with selection algorithms, offers another layer of robustness against overfitting. Stability selection operates by repeatedly sampling the data and applying a selection algorithm to identify stable features, thus enhancing the reliability of the selected model (Meinshausen & Bühlmann, 2010). Despite its advantages, stability selection can be computationally prohibitive in big data contexts and requires careful calibration of the subsampling and selection parameters.

Lastly, the advent of deep learning models, particularly neural networks with numerous layers and parameters, exacerbates the overfitting challenge. While these models are capable of capturing highly complex patterns, their training involves minimizing a loss function through gradient-based optimization methods, which can easily lead to overfitting if the model complexity is not adequately regularized. Techniques such as dropout, where a fraction of the neurons are randomly ignored during training, and batch normalization, which normalizes the input of each layer, have been proposed to address this issue (Srivastava et al., 2014;

Ioffe & Szegedy, 2015). However, the sheer scale and flexibility of these models mean that overfitting remains a significant concern.

The inherent statistical and methodological complexities in big data reveal significant challenges in extracting reliable and meaningful information. These issues necessitate rigorous skepticism and advanced methodologies to address the high model capacity, increased variance, and computational constraints, ultimately challenging the prevailing ideology that merely increasing data volume leads to better insights.

The Decontextualization Crisis

The Internet of Data's purported superiority over the Internet of Information hinges on the assumption that data, as the smallest unit of the digital ecosystem, can be abstracted from its context and other units (Mohseni Ahooei, 2022). However, this notion is fundamentally flawed due to the intrinsic properties of data and the complex interplay of contextual factors.

The notion of decontextualization crisis in the context of big data highlights profound epistemological and methodological challenges. Data, when abstracted from its context, ceases to be a "thing-in-itself" (Kant, 1908[1781]). Data cannot exist in a vacuum; it is inherently bound to the context of its creation and interpretation. Boyd and Crawford (2012) emphasize that data devoid of context loses its intrinsic meaning. This concept challenges the fundamental assumption of big data analytics, which posits that larger volumes of data inherently yield greater insights. Data abstraction, or the process of decontextualizing data, strips it of the essential contextual factors that impart significance and relevance. Advanced statistical techniques, such as hierarchical linear modeling (HLM), attempt to account for nested data structures but often fall short when the data is fundamentally decontextualized (Raudenbush & Bryk, 2002). HLM is designed to handle data with multiple levels of context, but when these contexts are not adequately captured, the results can be misleading.

Moreover, the premise of interpretability in big data hinges on the concept of "commensuration", which Espeland and Stevens define as "the transformation of different qualities into a common metric" (Espeland & Stevens, 1998: 314). This process is assumed to render data generic and universally interpretable. However, the inherent uniqueness of each data unit, due to its dependence on its physical counterpart, fundamentally contradicts this notion. Data normalization techniques, such as z-score standardization or min-max scaling, are often employed to achieve

commensuration, yet these methods introduce interpretative flexibility that can distort the original meaning of the data (Jain et al., 1999). Van Dijck (2012) argues that this "interpretative flexibility" undermines the reliability of big data analytics.

Digital media platforms frequently pre-structure user interactions into specific formats or "action grammars," including actions like liking, sharing, or commenting (Van Dijck, 2013). These standardized actions generate pre-defined data points, converting user behaviors into quantifiable metrics. Nevertheless, the meanings users attach to these actions can vary significantly, and the widespread use of third-party applications further complicates the interpretation of these data points (Gillespie, 2010).

Therefore, the generalization of relationships between data points to physical social relationships represents a significant misunderstanding in big data studies. The virtual network, composed of digital interactions, is not equivalent to the network of personal relationships. Granovetter's (1973) theory of weak ties and strong ties demonstrates that social relationships in physical contexts are fundamentally different from those in virtual environments. Advanced network analysis techniques, such as social network analysis (SNA) and graph theory, are used to map and analyze these relationships (Wasserman & Faust, 1994). However, these methods often fail to capture the depth and complexity of interpersonal relationships. For example, eigenvector centrality and betweenness centrality, commonly used SNA metrics, may not accurately reflect the strength and quality of social ties in virtual networks (Freeman, 1977). Additionally, the use of big data to analyze social networks often relies on assumptions of homophily and transitivity, which do not necessarily hold true in digital interactions (McPherson et al., 2001). Thus, equating virtual network metrics with real-world social dynamics is a flawed approach that overlooks the nuanced and context-dependent nature of human relationships.

The presumption that strong and weak ties in virtual networks can be equated with their counterparts in real-world social contexts is deeply problematic and fundamentally flawed. Granovetter's (1973) seminal work on the strength of weak ties posits that weak ties, or acquaintances, serve as crucial bridges in social networks, facilitating the flow of information and resources across otherwise disconnected groups. Strong ties, or close relationships, are characterized by frequent interactions, emotional intensity, and mutual confiding, serving as sources of support and solidarity. In the digital realm, however, the nature and quality of interactions diverge significantly from these

classical definitions due to the medium's inherent limitations and affordances.

The distinction between "articulated networks" and "behavioral networks" in social media research is pivotal for understanding the limitations and potentials of digital data in representing personal relationships. Articulated networks are those that users explicitly create, such as friendship lists or followerships on platforms like Facebook and Twitter. Behavioral networks, on the other hand, are inferred from user activities, such as interactions, likes, comments, and shares. While both types of networks offer valuable insights for researchers, they fall short of capturing the nuanced dynamics of personal networks, a limitation underscored by the concept of "tie strength" proposed by Granovetter (1973). Granovetter's theory differentiates between "strong ties", which are characterized by frequent, emotionally intense, and reciprocal interactions, and "weak ties", which are less frequent and emotionally distant but crucial for bridging different social groups and facilitating the flow of information across networks. Articulated networks may reflect strong ties to some extent, as individuals often list their close friends and family. However, these networks can be misleading, as they are subject to social desirability bias and strategic self-presentation. Users may include or exclude connections for reasons unrelated to the actual strength of their relationships. Behavioral networks provide a more dynamic view by capturing interactions that occur naturally over time. Yet, these networks are not immune to misinterpretation. High frequency of online interactions does not necessarily indicate strong ties, as people may interact frequently with acquaintances or even strangers due to shared interests or network algorithms that promote certain content. Moreover, the lack of context in behavioral data means that the quality and depth of relationships are often obscured. Therefore, while articulated and behavioral networks derived from social media offer substantial data for analysis, they cannot fully substitute for the richness of personal networks.

Furthermore, the context-dependent nature of human relationships poses a significant challenge to the direct application of virtual network metrics to physical social dynamics. Studies have shown that digital interactions often lack the depth and emotional resonance of face-to-face communication (Turkle, 2011). This discrepancy is particularly evident in the formation and maintenance of strong ties, which rely heavily on non-verbal cues, shared experiences, and the ability to provide immediate support and feedback (Wellman & Wortley, 1990).

Moreover, the reliance on big data to analyze social networks introduces significant methodological biases. Big data analytics often prioritize quantifiable interactions, such as likes, comments, and shares, while neglecting the subtleties of relational dynamics. The "datafication" of social interactions reduces complex human behaviors to simplistic metrics, which can lead to misinterpretations and overgeneralizations (Van Dijck, 2013).

Theoretical advancements in network science and sociology further highlight the limitations of equating virtual and physical social networks. Borgatti and Halgin (2011) argue that networks must be understood as multiplex, consisting of multiple types of ties (e.g., friendship, kinship, professional connections) that intersect and influence one another. The reduction of these multifaceted relationships to single-dimensional metrics in digital networks fails to capture the intricacies of social life.

The application of virtual network metrics to understand physical social relationships is fundamentally flawed due to the inherent differences in the nature of interactions, the context-dependent nature of human relationships, and the methodological limitations of big data analytics.

The New Digital Gap Crisis

Access to big data correlates tightly with socio-economic standing, exacerbating systemic disparities within the digital realm. The unequal and constrained access to vast data reservoirs creates a distinct digital schism, vividly demarcating global disparities across institutional lines. This divide is acutely pronounced within academia, where elite universities command substantial resources to furnish students with fundamental knowledge and state-of-the-art instruments essential for rigorous big data exploration. Proficiency in navigating and exploiting expansive data ecosystems demands mastery of intricate computational paradigms such as distributed computing architectures, parallel processing frameworks, and cloud-based infrastructures. Statistical acumen is equally imperative, involving advanced techniques including Bayesian inference, machine learning algorithms, and sophisticated data visualization methodologies. These technical proficiencies predominantly reside within the purview of computer science specialists, a discipline historically skewed towards male predominance, further complicating gender parity issues (Hellberg, 2024).

Furthermore, the ramifications of unequal access to big data transcend mere technical aptitude, profoundly influencing institutional capabilities to shape knowledge domains, inform evidence-based policies, and steer

societal advancements. Differential access to comprehensive and high-fidelity data repositories confers significant advantages in scientific inquiry, enabling privileged institutions to dictate research agendas, monopolize innovation pathways, and wield considerable influence over public discourse (Andrejevic, 2013).

Effective democratization", as Derrida posits, "is gauged by participation in and access to the archive—its formation and interpretation" (Derrida, 1996: 4). This principle assumes critical relevance in the domain of big data, where access to expansive datasets and the capacity to conduct comprehensive analyses are predominantly confined to privileged entities. Manovich (2011) categorizes the hierarchical structure within the big data milieu into three primary roles: data generators encompassing both deliberate contributors and inadvertent digital footprint creators, developers of analytical tools, and data analysts. This hierarchical segmentation underscores the inherent power dynamics within the big data ecosystem, where individuals possessing specialized expertise and substantial resources wield disproportionate influence over scientific inquiry and societal comprehension.

The digital divide significantly influences the questions posed in scientific research and the subsequent answers derived. The current big data ecosystem exacerbates the divide between the "Big Data rich" and the "Big Data poor", leading to a skewed landscape where a handful of institutions dominate knowledge production. The power to decide which questions are asked, what methodologies are employed, and how findings are interpreted resides with those who have access to extensive data repositories and the means to analyze them (Andrejevic, 2013). This concentration of power raises critical questions about legitimacy and governance within the realm of big data: Who has the authority to make decisions? Who owns and controls the data? Who operates the analytical infrastructure? These questions, as Derrida suggests, are paramount in contemplating our digital future.

When systemic inequalities are ingrained in the architecture of big data access and analysis, they entrench resistant class structures. The inequities in data access not only reinforce existing societal disparities but also create barriers to entry for less privileged institutions and individuals (Eubanks, 2018). This dynamic fosters a class of data elites who dictate the trajectory of research, policy, and innovation, thereby marginalizing diverse perspectives and perpetuating a cycle of exclusion. The resultant class structures are resistant to change, as those in positions of power have little incentive to democratize access or share resources equitably.

The IRB Crisis

In the realm of scientific research, Institutional Review Boards (IRBs) serve as critical gatekeepers, ensuring ethical practices, safeguarding human subjects, and maintaining public trust in scientific endeavors. Their operations are guided by the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979), which emphasizes respect for persons, beneficence, and justice. However, as scientific research evolves, particularly in the domain of big data, IRBs have faced significant challenges in maintaining their regulatory influence.

First, the sheer volume and variety of data available in big data research do not inherently justify its ethical use. The principle of informed consent, a cornerstone of ethical research, is often overlooked in big data studies. Informed consent necessitates that individuals are fully aware of and agree to the specific uses of their data. However, in the context of big data, individuals who provide their personal data on social media platforms and other digital services typically do not consent to their data being used for research purposes. This misuse of data can be likened to unauthorized surveillance, which contravenes ethical standards set forth by IRBs.

Advanced algorithms and artificial intelligence (AI) systems exacerbate this issue by mining data from various sources without explicit consent, raising serious ethical concerns. For instance, predictive algorithms used in big data analytics can infer sensitive information about individuals, even if such data were never explicitly provided (Metcalf & Crawford, 2016). This practice violates the principle of autonomy, as individuals lose control over their personal information and its uses.

Moreover, privacy concerns are paramount in the use of big data for research. The aggregation and analysis of vast amounts of data can lead to the discovery of intricate connections between data points, effectively enabling researchers to make detailed claims about individuals. This process, facilitated by sophisticated machine learning algorithms and cloud computing technologies, can inadvertently expose personal information and violate privacy rights. The re-identification of anonymized data sets, for instance, is a well-documented risk. Studies have shown that even de-identified data can be traced back to individuals through cross-referencing with other data sources (Narayanan & Shmatikov, 2008). The use of cloud computing for data storage and analysis further complicates privacy concerns, as data breaches and unauthorized access to sensitive information can occur despite advanced security measures.

Moreover, the statistical analyses resulting from big data explorations often involve correlations that can misrepresent or oversimplify complex social phenomena. When these analyses pertain to subjugated social groups, including children, women, and racial, religious, or cultural minorities, the results can perpetuate harmful stereotypes and exacerbate social inequalities. For instance, biased algorithms can reinforce existing prejudices, leading to discriminatory outcomes in areas such as education, healthcare, and criminal justice (O'Neil, 2016).

Furthermore, the application of quantum computing in big data research introduces new dimensions of complexity and risk. Quantum computing has the potential to break traditional encryption methods, thus exposing data to unprecedented vulnerabilities (Shor, 1994). These technological advancements necessitate rigorous oversight to ensure that privacy is not compromised in the pursuit of knowledge.

Finaly, the potential for big data research to harm vulnerable populations is a significant ethical concern as well. Children and teenagers, who are prolific generators of digital data, are particularly at risk. The analysis and publication of data derived from these groups can lead to unintended consequences.

The ethical implications of these practices are profound. The principle of justice, which mandates the fair distribution of research benefits and burdens, is often violated in big data research. Vulnerable populations may disproportionately bear the risks of data misuse, while the benefits of such research accrue to more privileged groups.

Conclusion

The advent of big data has undoubtedly transformed the landscape of social research, offering unprecedented opportunities for insight. However, it also brings to the fore significant challenges related to objectivity, contextualization, and access. To mitigate these challenges, it is imperative to adopt a holistic approach that integrates robust scientific methodologies, ethical considerations, and a critical perspective on the socio-political dimensions of data. By doing so, we can harness the potential of big data while safeguarding the integrity and depth of social research, ensuring that it contributes meaningfully to our understanding of human society.

Addressing the crises and challenges of big data research necessitates a multifaceted approach, integrating both technological advancements and methodological shifts. A critical starting point involves the de-reification of social processes. This requires moving beyond the perception of data as mere objective entities and recognizing their

embeddedness within social, cultural, and political contexts. Such an approach demands a nuanced understanding of how data points are generated, interpreted, and utilized, acknowledging that they reflect diverse human behaviors, intentions, and meanings.

A crucial aspect of big data research lies in its foundational conditions. The continuous collection of data across various domains and the context-aware interpretation of aggregated data. For big data to truly be effective, it is essential that data is not only amassed in real-time from diverse sources—ranging from what people do and say to their physical positions—but also meticulously analyzed with an understanding of the contexts from which it originates. This dual requirement ensures that the data reflects a comprehensive and dynamic snapshot of reality, capturing the multifaceted nature of human behavior and interactions. The continuous collection of data allows for the observation of trends and patterns over time, providing a more robust and granular understanding of social phenomena. Meanwhile, the aggregation and contextual interpretation of this data enable researchers to draw meaningful correlations and insights that are grounded in the specific circumstances of the data's origin. Without these conditions, big data would risk becoming a collection of disconnected and potentially misleading fragments, rather than a coherent and insightful resource for understanding complex social dynamics.

To navigate the complexities of big data, scholars must adopt new methodologies that transcend traditional paradigms. The continuous collection of data across various domains, coupled with sophisticated aggregation and context-aware interpretation, forms the bedrock of effective big data analysis. This entails not only tracking what individuals do and say but also understanding the subtleties of their interactions and behaviors. As Marres and Weltevrede (2013) suggest, metrics in this context should be considered "lively metrics"—dynamic and reflective of the evolving nature of social phenomena. These metrics, as Gerlitz and Lury (2014) argue, are inherently variable and animated, capturing the multifaceted ways in which individuals engage with digital platforms.

Furthermore, fostering a culture of critical engagement with data is essential. Researchers and students must be equipped to interrogate the assumptions and implications of data-driven findings, particularly regarding their roles in establishing value and credibility across different fields. This involves not only technical proficiency but also a critical understanding of the philosophical underpinnings of data interpretation and utilization. Public discussions should aim to demystify data points, opening up debates about their significance and impact on society.

Ethical considerations

The author has completely considered ethical issues, including informed consent, plagiarism, data fabrication, misconduct, and/or falsification, double publication and/or redundancy, submission, etc.

Conflicts of interests

The author declares that there is no conflict of interests.

Data availability

The dataset generated and analyzed during the current study is available from the corresponding author on reasonable request.

References

Akaike, H. (1974). "A new look at the statistical model identification". *IEEE Transactions on Automatic Control.* 19(6): 716-723. doi: 10.1109/TAC.1974.1100705.

Anderson, C. (2008). "The end of theory, will the data deluge makes the scientific method obsolete?:. *Edge*. retrieved at 12 September 2023 [Online] from: http://www.edge.org/3rd_ culture/anderson08/anderson08_index.html.

Andrejevic, M. (2013). *Infoglut: How too Much Information Is Changing the Way We Think and Know*. Routledge.

Badiou, A. (2008). *Number and Numbers*. Cambridge: Polity.

Bellman, R. (1961). *Adoptive Control Processes: A Guided Tour*. University Press.

Benjamini, Y. & Hochberg, Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society: Series B (Methodological).* 57(1): 289-300. https://www.jstor.org/stable/2346101.

Berman, J.J. (2013). *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information*. Newness.

Berry, D. (2011). "The computational turn: Thinking about the digital humanities". *Culture Machine*. 12. Retrieved at 08 October 2023 [Online] from: http://www.culturemachine.net/index.php/cm/article/view/440/470.

Beyer, K.; Goldstein, J.; Ramakrishnan, R. & Shaft, U. (1999). "When is 'nearest neighbor' meaningful?". *Database Theory—ICDT'99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7:* 217-235. Springer Berlin Heidelberg.

Bollier, D. (2010). "The promise and peril of big data". retrieved at 10 September 2023 [Online] from: http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf.

Borgatti, S.P. & Halgin, D.S. (2011). "Networktheorizing". *Organization Science.* 22(5):1168-1181.https://doi.org/10.1287/orsc.1100.0641

Bowker, G.C. & Star, S.L. (2000). *Sorting Things out: Classification and its Consequences*. MIT press.

Boyd, D. & Crawford, K. (2012). "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon". *Information, Communication & Society*. 15(5): 662-679. https://doi.org/10.1080/1369118X.2012.678878

Breiman, L. (2001). "Random forests". *Machine Learning.* 45: 5-32. http://dx.doi.org/10.1023/A:1010933404324.

Chawla, N.V.; Bowyer, K.W.; Hall, L.O. & Kegelmeyer, W.P. (2002). "SMOTE: Synthetic minority over-sampling technique". *Journal of Artificial Intelligence Research.* 16(2002): 321-357. https://doi.org/10.1613/jair.953.

Crotty, M.J. (1998). *The foundations of social research: Meaning and perspective in the research process*. London: Routledge.

Couldry, N. (2014). "Inaugural: A necessary disenchantment: Myth, agency and injustice in a digital world." *The Sociological Review, 62*(4): 880-897. https://doi.org/10.1111/1467-954X.12158.

----------. (2020). Recovering critique in an age of datafication. *New Media & Society, 22*(7): 1125-1336. https://doi.org/10.1177/1461444820912536.

Derrida, J. (1996). *Archive Fever: A Freudian Impression*. Translated by Prenowitz E. University of Chicago Press, Chicago.

Durbin, J. & Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*. Vol. 38. OUP Oxford.

Durkheim, E. (1982[1895]). *Rules of Sociological Method*. New York: The Free Press.

Espeland, W.N. & Stevens, L.M. (1998). "Commensuration as a social process". *Annual Review of Sociology*. 24(1): 313–343.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press. https://www.jstor.org/stable/223484.

Fairclough, N. (2013). *Critical Discourse Analysis: The Critical Study of Language*. Routledge.

Fan, J. (2008). "Sure indepedence screening for ultra-high dimensional feature space". *JR Stat Soc B.* 70(5): 849-911. https://doi.org/10.1111/j.1467-9868.2008.00674.x.

Fan, J. & Li, R. (2006). "Statistical challenges with high dimensionality: Feature selection in knowledge discovery". *arXiv preprint math/0602133*. 595-622.

Freeman, L.C. (1977). "A set of measures of centrality based on betweenness". *Sociometry.* 40(1):35-41. https://doi.org/10.2307/3033543.

Gayo-Avello, D.; Metaxas, P.T. & Mustafaraj, E. (2011). *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment*. Proceedings of the International Conference on Weblogs and Social Media (ICWSM).

Geertz, C. (1973). *The Interpretation of Cultures: Selected Essays*. Basic Books.

Gerlitz, C. & Lury, C. (2014). "Social media and self-evaluating assemblages: On numbers, orderings and values". *Distinktion: Scandinavian Journal of Social Theory.* 15(2): 174-188. https://doi.org/10.1080/1600910X.2014.920267.

Gillespie, T. (2014). "The relevance of algorithms". Edited by Gillespie T.; Boczkowski P.J. & Foot, K.A. *Media Technologies: Essays on Communication, Materiality, and Society:* 167-194. MIT Press.

----------. (2010). "The politics of 'platforms'". *New Media & Society.* 12(3): 347-364. https://doi.org/10.1177/1461444809342738.

Gitelman, L. (2011). *Notes for the Upcoming Collection 'Raw Data' is an Oxymoron*. retrieved at 10 October 2023 [Online] from: https://files.nyu.edu/lg91/public/.

Granovetter, M.S. (1973). "The strength of weak ties". *American Journal of Sociology.* 78(6): 1360-1380. https://www.jstor.org/stable/2776392.

Greene, W.H. (2003). *Econometric Analysis*. 8th ed. Pearson Education India.

Hansen, B. (2022). *Econometrics*. Princeton University Press.

Haraway, D. (2011). "A cyborg manifesto (1985)". *Cultural Theory: An Anthology.* Edited by Szeman I.; Kaposy, T.: 454-471. WILEY Blackwell.

Harman, G. (2018). *Object-Oriented Ontology: A New Theory Of Everything*. Penguin UK.

Hastie, T.; Tibshirani, R.; Friedman, J.H. & Friedman, J.H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Vol. 2: 1-758. New York: Springer.

Hayles, N.K. (2000). "How we became posthuman: Virtual bodies in cybernetics, literature, and informatics". Chicago: Chicago University Press.

He, H. & Garcia, E.A. (2009). "Learning from imbalanced data". *IEEE Transactions on Knowledge and Data Engineering.* 21(9): 1263-1284. doi: 10.1109/TKDE.2008.239.

Heckman, J. (2013). "Sample selection bias as a specification error". *Applied Econometrics.* 31(3): 129-137. https://doi.org/10.1007/s11747-021-00816-9.

Hellberg, L. (2024). *Reduce the Gender Gap in Computer Science Education Using Creative Programming.* Master's Programme, Interactive Media Technology. KTH/Skolan för elektroteknik och datavetenskap (EECS).

Ioffe, S. & Szegedy, C. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift". *International Conference on Machine Learning.* Pmlr: 448-456.

Jablonka, E. & Bergsten, C. (2021). "Numbers don't speak for themselves: Strategies of using numbers in public policy discourse". *Educational Studies in Mathematics.* 108(3): 579-596. https://doi.org/10.1007/s10649-021-10059-8.

Jain, A.K.; Murty, M.N. & Flynn, P.J. (1999). "Data clustering: A review". *ACM Computing Surveys (CSUR)*. 31(3): 264-323. https://doi.org/10.1145/331499.331504.

Johnstone, I.M. (2001). "On the distribution of the largest eigenvalue in principal components analysis". *The Annals of Statistics.* 29(2): 295-327. doi: 10.1214/aos/1009210544.

Johnstone, I.M. & Lu, A.Y. (2009). "On consistency and sparsity for principal components analysis in high dimensions". *Journal of the American Statistical Association.* 104(486): 682-693. https://doi.org/10.1198/jasa.2009.0121.

Kandel, E.R.; Schwartz, J.H. & Jessell, T.M. (2013). *Principles of Neural Science.* 5th ed. McGraw-Hill Education.

Kant, I. (1781[1908]). *Critique of Pure Reason.* Modern Classical Philosophers. Cambridge, MA: Houghton Mifflin.

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences.* Sage.

Koller, D. & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT press.

Kuhn, T.S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Latour, B. (2007). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oup Oxford.

Leinweber, D. J. (2007). Stupid data miner tricks: overfitting the S&P 500. *Journal of Investing*, *16*(1), 15-22. https://doi.org/10.3905/joi.2007.681820.

Little, R.J. & Rubin, D.B. (2019). *Statistical Analysis with Missing Data.* Vol. 793. John Wiley & Sons.

Lohr, S.L. (2021). *Sampling: Design and Analysis*. Chapman and Hall/CRC.

Manovich, L. (2011). "Trending: The promises and the challenges of big social data". *Debates in the Digital Humanities*. Edited by Gold M.K.

The University of Minnesota Press, Minneapolis, MN. Retrieved at 18 September 2023 [Online] from: http://www.manovich.net/DOCS/Manovich_trending_paper.pdf.

Marchenko, V.A. & Pastur, L.A. (1967). "Distribution of eigenvalues for some sets of random matrices". *Matematicheskii Sbornik.* 114(4): 507-536. doi: 10.1070/SM1967v001n04ABEH001994.

Marres, N. & Weltevrede, E. (2013). "Scraping the social? Issues in live social research". *Journal of Cultural Economy.* 6(3): 313-335. https://doi.org/10.1080/17530350.2013.772070.

Marx, K. (1932[1845]). "Theses on Feuerbach". *The German Ideology.* Edited by Pascal, R. New York: International Publishers. Viktor and Cukier

Mayer-Schönberger, V. & Cukier K. (2013). *Big Data: A Revolution that Will Transform How We Live, Work, and Think*. New York: Houghton Mifflin Harcourt.

McPherson, M.; Smith-Lovin, L. & Cook, J.M. (2001). "Birds of a feather: Homophily in social networks". *Annual Review of Sociology.* 27(1): 415-444. https://doi.org/10.1146/annurev.soc.27.1.415.

Meinshausen, N. & Bühlmann, P. (2010). "Stability selection". *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 72(4): 417-473.https://doi.org/10.1111/j.1467-9868.2010.00740.x.

Meng, X.L. (2018). "Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election". *The Annals of Applied Statistics.* 12(2): 1-14. https://doi.org/10.1016/j.ijforecast.2024.04.008.

Metcalf, J. & Crawford, K. (2016). "Where are human subjects in big data research? The emerging ethics divide". *Big Data & Society.* 3(1): 2053951716650211. https://doi.org/10.1177/2053951716650211.

Mohseni Ahooei, E. (2023). "The end of information age society 5.0 and the L [e] ast man". *Journal of Cyberspace Studies.* 7(1), 45-66. doi: 10.22059/JCSS.2022.346205.1078.

----------. (2022). "Shifting from individualism to genericism: Personalization as a conspiracy theory. *Žurnalistikos Tyrimai*. 16: 14-38. https://doi.org/10.15388/ZT/JR.2022.1.

Moran, D. (2002). *Introduction to Phenomenology*. Routledge.

Narayanan, A. & Shmatikov, V. (2008). "Robust de-anonymization of large sparse datasets". *2008 IEEE Symposium on Security and Privacy:* 111-125. IEEE.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). "The belmont report: Ethical principles and guidelines for the protection of human subjects

of research". Retrieved at 21 September 2023 [Online] from: https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf.

Nissenbaum, H. (2011). "Privacy in context: Technology, policy, and the integrity of social life". *Journal of Information Policy.* 1: 149-151. https://doi.org/10.1145/3547299.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pessach, D. & Shmueli, E. (2022). "A review on fairness in machine learning". *ACM Computing Surveys (CSUR).* 55(3): 1-44. https://doi.org/10.1145/3494672.

Pietsch, W. (2021). *Big Data*. Cambridge University Press.

Pond, P. (2020). *Complexity, digital media and post truth politics: a theory of interactive systems*. Springer Nature.

Porter, T.M. (2020). *The Rise of Statistical Thinking, 1820–1900*. Princeton University Press.

Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Vol. 1. Sage.

Resnyansky, L. (2019). "Conceptual frameworks for social and cultural Big Data analytics: Answering the epistemological challenge". *Big Data & Society*. *6*(1): 1-12. https://doi.org/10.1177/2053951718823815.

Schwarz, G. (1978). "Estimating the dimension of a model". *The Annals of Statistics.* 6(2): 461-464. https://doi.org/10.1214/aos/1176344136.

Shor, P.W. (1994). "Algorithms for quantum computation: Discrete logarithms and factoring". *Proceedings of the 35th Annual Symposium on Foundations of Computer Science*: 124-134. IEEE. https://doi.org/10.1109/SFCS.1994.365700.

Smith, K.E. (2010). *Meaning, Subjectivity, Society: Making Sense of Modernity*. Leiden and Boston: Brill.

Sporns, O.; Bullmore, E. & Kaiser, M. (2008). "The human connectome: A structural description of the human brain". *PLoS Biology*. 6(7): 0245-0251. doi: 10.1371/journal.pcbi.0010042.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I. & Salakhutdinov, R. (2014). "Dropout: A simple way to prevent neural networks from overfitting". *The Journal of Machine Learning Research.* 15(1): 1929-1958. doi: 10.5555/2627435.2670313.

Taylor, C. (1986). *Self-Interpreting Animals. In Martin Heidegger*. Edited by Mulhall S. London: Routledge.

Tenenbaum, J.B.; Silva, V.D. & Langford, J.C. (2000). "A global geometric

framework for nonlinear dimensionality reduction". *Science.* 290(5500): 2319-2323. doi: 10.1126/science.290.5500.2319.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society Series B: Statistical Methodology.* 58(1): 267-288. https://www.jstor.org/stable/2346178.

Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books.

Tversky, A. & Kahneman, D. (1974). "Judgment under uncertainty: Heuristics and biases". *Science.* 185(4157): 1124-1131. doi: 10.1126/science.185.4157.1124.

Van Dijck, J. (2013). *The Clture of Connectivity: A Critical History of Social Media*. Oxford University Press.

----------. (2012). "Tracing Twitter: The rise of a microblogging platform". *International Journal of Media and Cultural Politics.* 7: 333–348. https://doi.org/10.1386/macp.7.3.333_1.

Van Es, K. & Schäfer, M.T. (2017). *The Datafied Society. Studying Culture through Data*. Amsterdam University Press.

Vapnik, V. (2013). *The Nature of Statistical Learning Theory*. Springer Science & Business Media.

Wasserman, S. & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.

Wellman, B. & Wortley, S. (1990). "Different strokes from different folks: Community ties and social support". *American Journal of Sociology. 96*(3): 558-588. https://doi.org/10.1086/229572.

Yeo, G. (2021). *Record-Making and Record-Keeping in Early Societies*. Routledge.

Ziegel, E.R. (2002). "Statistical inference". *Technometrics*. 44(4): 407–408. https://doi.org/10.1198/tech.2002.s94.