THE ALGORITHMIC ARCHIVE: DATA SCIENCE AND BIG DATA ANALYTICS AS TRANSFORMATIVE FORCES IN HUMANITIES RESEARCH

Arin Jain

Grade - 12
School Name- Abhinav English School, Pune

Jitendra Vinayak Sandu

Director
Dr D. Y. Patil B-School, Pune

ABSTRACT

The intersection of Data Science and Big Data Analytics with Humanities scholarship marked a fundamental change in the research and knowledge of cultural heritage, historical stories, and textual traditions. The critical observation of the increasingly key role of computational methods in humanities studies is sifted in terms of the challenges and possibilities it poses for interdisciplinary research. This paper aims to demonstrate, based on the recent growth of data science in the humanities and computing approaches to literature studies, how data-driven perspectives have the potential to disclose patterns, analyse large quantities of data, and facilitate interdisciplinary work, and may also pose significant questions regarding data epistemology, algorithmic bias, and the preservation of humanistic values. This study draws on key research infrastructure investments, pedagogical advancements regarding data competence, emerging frameworks for ethically integrating AI, and a vision for a balanced approach. The major conclusions are as follows: interdisciplinary dialogue is necessary but not sufficient; data justice is a critical component of successful integration; and integration requires changes in scholarly practice. Finally, future research directions are presented that focus on critical studies of data, the notions of FAIR, and building workflows that integrate both computational and humanistic epistemologies.

Keywords: Data Science, Big Data Analytics, Digital Humanities, Computational Humanities, Interdisciplinary Research, Cultural Heritage, Text Mining, Machine Learning, Algorithmic Bias, Data Justice, Digital Archives, Computational Methods, Data Literacy, Humanities Research, FAIR Principles

A. INTRODUCTION

a. Contextualising Data Science and Big Data Analytics in the Humanities

The 21st century has witnessed unprecedented growth in the amount of digital data, which has revolutionised all domains of human knowledge. Although the fields of biomedicine, economics, and climate science have long moved towards computational methods, the humanities are now starting to contemplate the consequences of what has been called digital abundance (Cha, 2023). This has been enabled by decades of digitisation work, an explosion of born-digital material, and the emergence of powerful tools for processing and analysing cultural materials at scale.

Data Science is a separate interdisciplinary domain since its rise, and the dynamics between computer science and other humanities have been completely revised. Data science offers a method for determining the meaning of this stored treasure of digitalised cultural heritage

(Mayernik, 2023). This has transitioned from a service-provider mentality that has been the typical depiction of computer scientists making tools for humanists to a collaborative working mentality characterised by computer scientists and humanists working together in research design and knowledge creation. Data science is the interplay among the disciplines of statistics, informatics, computing, communication, management, and sociology, and it has opened new opportunities for a data-driven understanding of complex cultural phenomena from a computational perspective (Parti & Szigeti, 2021).

These developments are hugely important for the humanities. Today, large digital repositories are home to millions of historical documents, vast literary corpora, archaeological databases, and linguistic resources, allowing forms of enquiry that were never possible. Scholars can track changes in language over centuries, visualise social connections within historical communities, recognise stylistic structures across thousands of texts, and discover unexpected relations between cultural artefacts (Waters, 2022). These capabilities do not supplant classic humanistic ways but instead supplement them with another way of thinking about timeless questions regarding human culture, expression, and experience.

Computational methods have become widespread in cultural institutions for preservation, access, and communication with the public (cultural institutions). The adoption of data analytics in museums, libraries, and archives is growing, as outlined by Ncube and Ngulube (2025), to manage collections, guide preservation, and inform public programmes. Generating such massive datasets via digitisation allows for analysis, thereby changing the construction and maintenance of cultural memory in digital contexts, as Belteki et al. (2025) say.

This work is unique in its thorough description of the data science-humanities interface without succumbing to technological determinism or methodological scepticism and instead takes a sympathetic yet critical approach (Berry, 2022). This paper compiles and contextualises the ongoing debates about how to "Engage with Data in the Humanities" while maintaining the humanities' unique role in the creation of knowledge (Belyak, 2021) in view of recent advances in the computational humanities, ethics, and pedagogical innovation.

b. Research Questions and Objectives

The investigation and analysis of this study focused on three main research questions.

The first research question investigates the (r)evolution of research methodologies and practices in the humanities considering the advent of data science and of big data analytics. This question examines how computational approaches are currently used in the humanities field and in each discipline, ranging from literary studies to history and archaeology. It explores the tools, techniques, and processes that have developed from this intersection and how they result in both a variety of questions that can be asked and new answers that can be found in the future.

The second research question explores the epistemological, methodological, and ethical issues related to embedding data-driven research in the humanities. This question delves into the conflict between quantitative and qualitative paradigms, the type of data applied to humanistic fields, and the ethical concerns regarding such algorithmic analyses. This reflects the use of computational approaches that might illuminate or obscure cultural phenomena and what it means for researchers to take responsible measures with computational methods.

The specific Objectives are as follows:

- Assess existing uses of data science for humanities research
- Understand unmet gaps and barriers to current methods
- Discuss Tom Trueblood's case studies on successful interdisciplinary collaboration
- Formulate principles of ethical and good integration
- Identify paths of further research and practice

These questions are important because they define how the humanities interact with the computational turn. As data-laden methods are becoming a growing part of university research, cultural heritage preservation, and public discourse, humanists will not only need to have technical facility in using such tools but also to engage in important thinking about the way they can be applied.

B. LITERATURE REVIEW

a. Overview of Data Science and Big Data Analytics

Data science is a combination of statistics, computer science, and domain expertise to derive findings from data. It includes machine learning, natural language processing, network graph analysis, and data visualisation for all types of data, whether structured or unstructured. Big data analytics refers to large and complex datasets with the characteristics of volume, velocity, and variety.

Data science has come a long way towards becoming a democratic technology. Machine learning and visualisation libraries, such as Python and R, are also accessible in open-source programming languages. These tools enable scholars without extensive computational training, especially in the humanities, to experiment with these techniques, as formal training in data analysis is often not taught. This accessibility, along with the ease of interfaces and long documentation, allows scholars to study computational methods without becoming computational specialists.

The interdisciplinary nature of data science, both in its added value and difficulty, lies at its heart. The power comes from the three domains of statistics, computer science, and domain experts. The inherent interdisciplinarity applies especially to the humanities, where a tradition of combining points of view exists (Bednarowska-Michaiel & Uprichard, 2025). However, it also brings its own set of challenges due to the different epistemologies, methods, and cultures of publication in various domains, which necessitate patience, respect for one another's ways of working, and eagerness to learn from others' traditions and cultures.

Three subfields of analytics are more relevant to the humanities. Descriptive analytics combines and integrates patterns and has been widely used to explore trends and anomalies in cultural data (yet routine with text mining, topic modelling, and network visualisation). Predictive analytics: Why are cultural changes predicted in the future using historical data, but less widely used in humanities research? Prescriptive Analytics for the optimisation of decision-making processes, such as digital preservation (Ibrus et al., 2021) and collection management.

There has been tremendous development in the technical infrastructure that powers data science in the humanities. In addition to scalable storage using cloud computing platforms and digital repositories for storing and sharing data, collaborative platforms support distributed work within institutions and disciplines. This improvement contributes to the feasibility and sustainability of large-scale computational studies (Sendra et al. 2024).

b. Application of Data Science in the Humanities

In the past 20 years, the humanities have witnessed a huge growth in Data Science activities. It follows in the footsteps of academic computing with its move from mainframe projects to personal computing to cloud-based collaborative work environments (Barbecho et al., 2023).

Table 1: Computational Methods Across Humanities Disciplines

Method	Application	Example
Text Mining/NLP	Authorial attribution, conceptual evolution	Tracking philosophical terms across centuries
Network Analysis	Social relations, character networks, idea diffusion	Mapping Renaissance correspondence networks
Machine Learning	Handwriting recognition, artefact classification	Neural networks for archaeological pottery identification
Geographic Information Systems	Spatial analysis, historical mapping	Visualising trade routes and cultural exchange
Computational Demography	Historical population analysis	Studying family structures in historical census data

The power of data science in historical demography is demonstrated. Projects that combine enormous and complex datasets from historical sources offer a way to address new research questions regarding population, social inequalities, and historical change, as well as pose recent problems about data quality, standardisation, and interoperability (Edvinsson et al., 2023). These are projects that need to pay attention to: how and what data researchers gather and what assumptions researchers make.

However, there is still more work to be done, and there are undoubtedly important gaps. A significant obstacle is that humanities data are heterodox, context-bound, and frequently unstructured. Unlike scientific data, humanities data can only be understood in historical and social processes, and the data must be understood before reading and interpretation can be performed (Li et al., 2023). There is a tension between quantitative and qualitative researchers regarding epistemological issues (what is evidence) and analysis (how to analyse).

These problems were intensified by skill deficits. In research projects, collaboration between humanities scholars and computer scientists reveals that some humanities scholars are not trained in computational research methods, and computer scientists do not know the specialised language of humanities research. Common terminology and collaborative practices require time and effort to be established (Schwandt, 2020). Institutional settings and mechanisms do not provide incentives for such types of collaboration, especially when it comes to promotion and tenure processes that prioritise traditional publications.

The prospects of this study are promising. Simultaneously, computational methods can augment close reading to find patterns that have never been spotted at the far larger scale of interdisciplinary studies. There is a need to be aware of the limitations. Algorithms are far from neutral; they reflect assumptions and biases that can amplify or sustain inequality. Computational abstraction constitutes a kind of "onion-ness", where the specificity and nuance cherished by humanists can be lost. Digital data infrastructures reflect institutional priorities and power relations in their local practices of digitisation and storage, shaping what is knowable and how (Belteki et. al, 2025).

C. METHODOLOGY

a. Research Design and Approach

The methodology used in this study is interpretive content analysis and critical data study, which is in the field of algorithmic archives and digital cultural heritage. This qualitative data allows the analysis of epistemological, methodological, and ethical issues that would otherwise only be raw numbers.

A case-based, qualitative design was chosen to learn about the current practice of computational techniques and the conditions, effects, and workings of such practice. This contextual, complex understanding is complementary to what quantitative methods offer. This research investigates various use cases of interdisciplinary data science, patterns, and idiosyncrasies.

Three Layers of Analysis:

Layer	Focus	Data Sources
Infrastructure	Projects, platforms, initiatives enabling data-driven research	Research centres, funding programs, technical documentation
Methodological Practices	Adaptation and use of computational approaches	Specific analytics, workflow development
Pedagogical Innovation	Data literacy and computational training	Syllabi, courses, programs

It is a multi-level approach that recognises that successful integration must address every issue. Infrastructure alone is insufficient; the right methods and training are also necessary. An approach based on anything other than education is unsustainable. Pedagogy that is untethered from research cannot prepare students for real-world practice. Considering all three aspects helps define synergies and gaps which may not otherwise be found.

Ethical aspects were emphasised from the beginning of this research. Special emphasis was given to the issues of data justice, algorithmic biases, and the responsibility of researchers regarding the data representation of cultural heritage. Equal attention was paid to voices that were hitherto absent from archives and to the potential of data-driven methods to either (re)produce or upset power relations. The ethical aspects of the choices regarding the data to gather and analyse and how to report the data are explained clearly.

b. Data Collection and Analysis

Various data sources were used to achieve a detailed snapshot of current data:

Table 2: Data Collection Sources

Source Type	Specific Materials	Purpose
Published Literature	Peer-reviewed articles, conference papers, book chapters	Establish state of the art and key discussion points
Project Documentation	Project reports, technical documentation, research outputs	Understand practical design and implementation
Pedagogical Materials	Syllabi, course descriptions, teaching materials	Examine computational skills and data literacy instruction

Literature Search Strategy:

- Databases: Scopus, Web of Science, JSTOR, Project MUSE
- Search terms: "Digital Humanities," "Computational Humanities," "Big Data Analytics," "Text Mining," "Machine Learning," "Cultural Analytics."

- Time period: 2015-2025
- Language: English
- Geographic focus: Inclusive of but not limited to Western/European contexts.

Inclusion Criteria:

1. Relevance to research questions
2. Representation of diverse disciplinary and institutional contexts
3. Sufficient documentation for analysis
4. Peer-reviewed sources for literature (articles, conference papers, books)
5. Project documentation from publicly funded research initiatives

Analytical Strategy:
Interpretive content analysis was conducted through:

1. **Open Coding**: Initial reading and line-by-line coding of all materials
2. **Axial Coding**: Identification of relationships between initial codes
3. **Selective Coding**: Development of core themes and patterns
4. **Synthesis**: Integration of themes into coherent findings

The process was iterative, allowing emergent themes beyond the predetermined research questions to emerge. This responsiveness to the material is appropriate for humanities research, where context and nuance are valued.

Limitations:

- Geographic concentration: Most literature focuses on Western/European contexts.
- Language constraints: Non-English sources limited.
- Temporal scope: Rapidly evolving field; findings may become dated.
- Access limitations: Few project documentations may be proprietary or unpublished.
- Interpretive bias: Researcher's disciplinary background influences analysis

D. RESULTS

a. Key Findings

The analysis resulted in four key findings with respect to the role of data science and big data analytics in humanities research.

One key finding was that sustaining interdisciplinary partnerships is crucial.

Finding time is essential for integration, for research questions to be co-formulated with the actors in charge, and for governance structures to provide adequate representation of all stakeholders. Large interdisciplinary projects are a testament to the fact that interdisciplinary collaboration can be very productive, provided there is interest in learning from others.

Table 3: Characteristics of Successful Interdisciplinary Projects

Characteristic	Description	Evidence
Co-located Research Teams	Physical proximity facilitates informal exchange	Chen et al., 2025
Extended Learning Cycles	Time for mutual understanding beyond initial	Parti & Szigeti, 2021

	phases	
Equitable Governance	Shared decision-making across disciplines	Newman, 2023
Regular Project Meetings	Sustained communication throughout project lifecycle	Sendra et al., 2024
Embedded Institutional Support	Cultural heritage institutions with dedicated teams	Belteki et al., 2025

Integrated teams in cultural heritage institutions have been particularly successful, where regular opportunities to meet have fostered the sustainability of the local research community (Chen et al., 2025). The projects demonstrate the requirement for ongoing communication, trust construction, and joint goals. In interdisciplinary work, the mere facilitation of this is not enough; it must be made possible by structures and processes that promote cooperation.

Case Example: The Living with Data Project
This collaborative initiative between historians, computer scientists, and demographers successfully integrated multiple historical datasets to analyse population patterns over three centuries. The success factors included:

Two-year preliminary phase for mutual training

- Co-located workspace with dedicated collaboration areas.
- Joint authorship with authorship contributions explicitly documented.
- Regular "translation" sessions where each discipline explained methods to others.

Finding 2: Humanities data are fundamentally different from the scientific ones.

Humanities data often contain textual and contextual elements that must be interpreted through schemes for meaningful analysis. Humanities data, whether in the arts or humanities, differ from numeric scientific data in that their organisation, categorisation, and encoding involve decisions made by scholars, contingent on the research context.

Table 4: Key Differences Between Humanities and Scientific Data

Aspect	Scientific Data	Humanities Data
Origin	Controlled experiments, instruments	Historical documents, cultural production
Structure	Typically numerical, structured	Textual, multimodal, often unstructured
Context	Context can be controlled	Context is integral to meaning
Collection	Planned data collection	Existing materials discovered
Interpretation	Primarily quantitative analysis	Requires qualitative interpretation

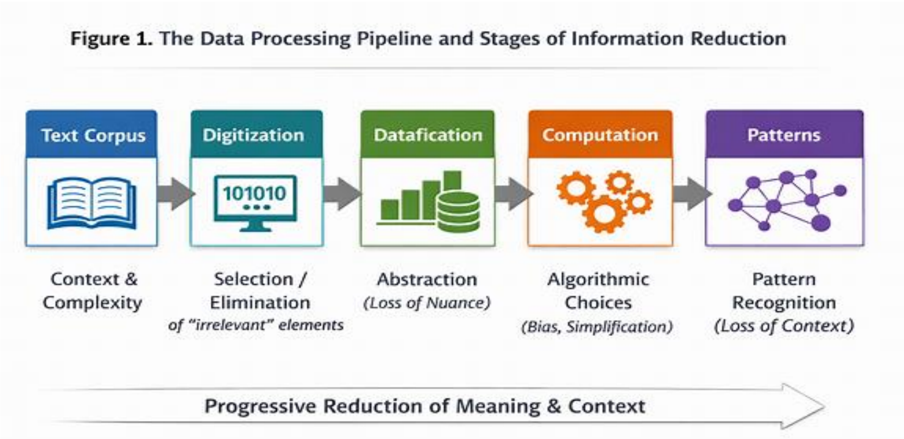
This is important for creating, labelling, and sharing datasets. The differentiation between Big Data (large volumes of digitised materials) and Smart Data (attention to the careful curation, annotation, and contextualisation of data collections) is of particular significance. Big data e-science can only be used after the organisation, description, curation, and annotation of the data; therefore, big data e-science demands strong humanistic skills. Humanities data have no inherent semantic meaning; rather, they acquire meaning through the interpretive processes of the researcher.

Finding 3: Pattern Identification and the Concealment of Contexts are Inherent Tensions

Computational techniques can uncover patterns, trends, and relationships that are not apparent using traditional techniques. However, abstraction is also an issue in these types of discoveries: when texts or artefacts (or culture) are rendered as data, some of it is lost or flattened.

Figure 1. The Data Processing Pipeline and Stages of Information Reduction

This flowchart shows the progressive transformation of a text corpus into computational patterns. In each stage, the process and associated epistemic loss were noted.



The algorithmic processing of cultural heritage materials is performative, thus constructing knowledge by privileging certain discourses over others. This is not an excuse to jettison the computational approach; rather, it is an opportunity to be critical of what is said and what remains unsaid. If a scholar's approach is not all-encompassing, he/she must state its nature, and the assumptions taken in the analysis must be identified.

Topic Modelling Victorian Novels: fifteen examples of using Topic Modelling with Victorian Novels:

Finds: Main overarching themes appearing in hundreds of novels; shift over time from domestic to imperial themes.

Difficulty reading between the lines; Character development, subtext, and aesthetic features are not present or salient.

Irony is not present or salient; Narrative voice is not present or salient.

Risk: Misinterpretation of statistical pattern(s) as literal significance.

Key finding 4: Data Literacy and Ethical Frameworks Must Go Hand in Hand with Technical Skills

Technical skills alone are insufficient for responsible data-driven approaches. There is also a need for data literacy, data ethics, and critical data studies to be the foundations of data skills to be learned by students and researchers. Essential Competencies for the Computational Humanities is a table outlining relevant skills that were deemed necessary for students working in the field.

Table 5: Essential Competencies for Computational Humanities

Competency Type	Specific Skills	Example
Technical	Data wrangling, analysis, visualisation	Python, R, SQL
Epistemological	Understanding data as constructed, not given	Recognising selection bias in archives
Ethical	Data justice, algorithmic bias, responsible AI	Evaluating representational harms
Methodological	Combining computational and traditional methods	Mixed-methods research design
Interpretive	Critical reading of algorithmic outputs	Questioning model assumptions

Small data approaches, with small, curated data sets, can be of pedagogic benefit in providing rich background information to model such skills while lessening technical challenges. Datasets can be humanised and messy, and approaches and assumptions can be undermined and challenged in ways that big data exercises may not allow.

Emerging Focus on FAIR Principles and Responsible AI

- **Findable**: Data discoverable with appropriate metadata
- **Accessible**: Data accessible for human and machine reading
- **Interoperable**: Data integration with other resources
- **Reusable**: Data clearly licensed for reuse

b. Interpretation of Results

These findings are important for research and work in the humanities. In addition to new tools, successful integration will require a transformation in the processes of conducting research, collaborating among scholars, and producing and accepting knowledge.

Need for institutional change:

- New organising structures (interdisciplinary centres, labs)
- New framework of rewards (rewards for teamwork)
- New quality criteria (Evaluation of computational methods)

Humanistic contributions to Data Science:

The scholars of the humanities are not only users of the technologies; they also produce knowledge, which illuminates:

- The way data are formed.
- How algorithms operate
- What constraints there have been in using techniques

Humanists can offer critical frameworks which can enhance computational analysis, not detract from it: attention to context, interpretation, power, and ethics. Within this context, the role of the humanities is also highly relevant to allay concerns and counterarguments concerning data and algorithms in the broader debate in society.

Pedagogical Challenges:

- Development of the curriculum that involves technical and critical skills.
- Setting up the programs by each institution is the context for facilitating interdisciplinary education.
- Development of a computational methods faculty.
- Technical skills and critical reflection in balance.

Broader Implications:

However, computational humanities approaches and the digital archives/algorithmic systems in which they are located have implications for:

- Cultural memory

- Public history
- Cultural policy
- Algorithmic governance
- Data justice

E. DISCUSSION

a. Comparison with Existing Literature: This study fits into and builds on the growing discourse of humanities and computer sciences. Over the years, the manifest advancements in data analysis, text mining, and machine learning applications have confirmed trends as well as added depth to the institutional and pedagogical contexts that help or hinder successful implementation.

This Study Builds on The Other Study:

- For Newman (2023) on interdisciplinary collaboration, this research brings some concrete processes.
- Berry (2022) – critical digital humanities: this research makes a bridge between critical and practice-based activities.
- Schwandt (2020) on digital methods: focus is on pedagogy and infrastructure of this study.

Figure 2. Bridging Traditional and Computational Humanities

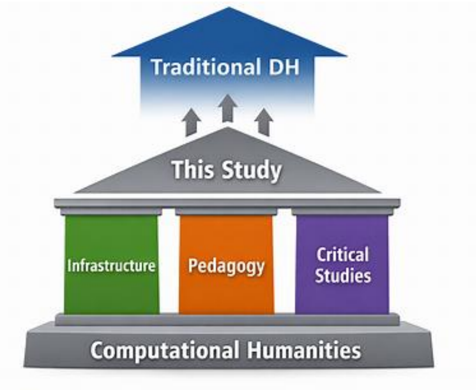


Figure 2. *Bridging Traditional and Computational Humanities*

What this Study Adds:

Integrity between infrastructure, methods, and pedagogy (many of the studies are based within one domain only)

- Concrete collaboration processes besides the generic one of "interdisciplinarity"
- Clearly consider data uniqueness about epistemological matters and methodology
- Focus on Pedagogical aspects: training as key to sustainable integration.
- Ethics 'in' technical practice, not 'tacked-on' away from computational practice.

Changes from Previous Studies:

Unlike more studies devoted to applications or disciplines, this study covers a synthesis of the humanities. It places a stronger emphasis on pedagogy than usual computational methods research, where training and education are important in sustaining a data-based approach. One special contribution is the realisation of the connection between technology and a lack of technology without ethical rules.

b. Principles for Ethical and Effective Integration

Based on the results, I propose five rules for introducing data science in humanities research.

Table 6: Principles for Ethical and Effective Integration

Principle	Description	Operational Guidance
1. Data Justice	Ensure marginalised communities are represented and their rights protected.	Audit datasets for representational gaps; involve communities in curation decisions.
2. Methodological Pluralism	Combine computational and traditional methods.	Use computational analysis as one method among many; prioritise research questions over methods.
3. Interpretive Transparency	Document all analytical decisions and assumptions	Maintain research logs; publish code and workflows; discuss limitations
4. Sustained Collaboration	Build structures for genuine interdisciplinary partnership	Allocate dedicated time for mutual learning; develop equitable governance models
5. Critical Pedagogy	Train researchers in both technical skills and critical frameworks	Integrate ethics throughout curriculum; use "small data" for contextual learning

c. Future Research Directions

The results of this study suggest specific areas for further investigation that require academic attention.

Perhaps more importantly, the critical data realm in the humanities, an area where emerging critical theory from the sociological and information science fields will have to be adapted to produce a critical approach that is distinctively humanistic and responsive to the unique features of cultural heritage materials and humanistic interpretation. The interaction between humanistic ways of reading and making meaning with data critique must be explored: What is data justice, particularly in cultural heritage contexts, where questions of representation, ownership, and colonial legacies come to the fore?

The geographic and institutional variation in the current state of research is highly inadequate, with most research taking place in Western European and North American settings, where there is strong funding and institutional support. Further study is needed that explores what is happening (or not) about how to implement computational humanities approaches in Global South settings in various kinds of institutional contexts, such as research universities, community archives and museums, and comparative analyses of different types of national policy frameworks and cultural contexts.

Digital humanities infrastructure has become a crisis point in terms of its long-term sustainability because more exciting projects lack long-term funding beyond initial grants. Clearly, we need to ask questions about sustainable models, the set-up of institutional support, collaborative governance structures in which multiple partners collaborate in ways that share responsibility, and social and organisational conditions (usually only considered from a technical perspective) which are crucial to the staying power of digital resources.

However, the contrast between computational and traditional methods should not be thought of as oppositional but enriching to the other, and there is scarce knowledge on how the methods can be productively integrated. In particular, an important question is how computational analysis can be used to inform and enrich approaches to close reading, and how, in turn, traditional approaches to close reading can be integrated into the design of computational processes to help mitigate the flattening of abstraction inherent in algorithms.

Lastly, there are ethical implications of new and more powerful approaches to computation, which warrant ongoing research, including how to assess algorithmic bias and how to mitigate its risks in cultural heritage environments, data privacy and consent in cultural heritage environments with sensitive resources, and field-specific ethical guidelines that extend beyond the blanket principles of AI to the specific challenges of the computational humanities.

F. CONCLUSION

a. Summary of Key Points

This study explored the presence of data science and big data analytics in humanities research, and changes in research practice, integration problems, and models to facilitate effective integration in the humanities.

Key Findings

The analysis in this study shows that data-driven methods provide unprecedented opportunities for pattern discovery and analysis at a large scale, and for the use of interdisciplinary collaboration, a situation that is greatly simplified by the increasing democratisation and ease of use of computational tools that are accessible to humanities researchers. However, some major obstacles must be addressed: humanities data are context-dependent, interpretive, and subjective, and involve conscious choices throughout the research process; and data-driven approaches that provide broad trends have a loss of granularity that depends on critical reflection on what has been gained and lost as a result of computational abstraction. Beyond this, algorithmic bias, representational harm, and data justice concerns further undermine this situation and necessitate explicit ethical frameworks. Analysis of these cases revealed four essential conditions for undertaking integration well: sustained interdisciplinary work that establishes genuine collaborative relationships; explicit attention to humanities data specificity in the creation of datasets and analysis of the data; imbalanced attention to algorithmic patterns and contexts; and complete data literacy as well as ethical guidelines that lead to responsible practice. Indeed, from the standpoint of successful integration, the technical layer of integration is only part of the picture; it also needs to be underpinned by fundamental institutional support, continued pedagogical growth in terms of competence in computational and critical skills, and ongoing critical reflection on the transformation of knowledge practices in the humanities by the integration of computational methods.

b. Final Thoughts

The application of data science and big data analytics to the humanities is a new phenomenon in scholarship history. It is not merely a methodological novelty; it represents a change in the understanding and study of cultural materials. The computational humanities will help advance the humanities in new ways, facilitating interdisciplinary work and extending humanistic thought.

However, further research is needed to examine the challenges and risks of its use if this potential is to be realised. Thus, the problem of interpreting the meaning and complexities of human experience has always been a concern in the humanities. Such concerns should not be neglected in the pursuit of computational methods. Rather, they should reveal the development and practical use of computational methods. These are the critical frameworks that differentiate humanistic enquiry, which is needed to ensure that the data-driven approach does not distort understanding.

Essential Investments:

Training and Collaboration:

Faculty are required to embed the 4 Es of Data Literacy in their courses.

Some institutions are expected to facilitate program-interdisciplinary development and faculty development.

Technical innovation and critical reflection need to be part of research funding.

Core Commitments:

Contextual sensitivity – awareness of cultural materials as such embedded in the language and the culture.

Ability to see cultures as intricate; to see things as nuanced, ambiguous,

An awareness of and a sensitivity to the significance of cultural phenomena

Beyond Academia:

The work of cultural memory studies, public history, and cultural policy is influenced by the frameworks of computational humanists. Humanities data science is no longer an entirely ivory tower project but is also involved in understanding and remembering societies and their histories.

CONCLUSION

Data Science for the humanities should be investigated and developed further using an interdisciplinary approach. Challenges are critical, but whatever exists also has opportunities. A researcher can use technical skills and critical reflection to create a more dynamic and responsible attitude toward the use of computational approaches in the study of human culture.

REFERENCES

1. Barbecho, L. B., Muñoz, S. R., García, E. G.-B., & Toscano, M. (2023). Digital humanities at global scale. Interdisciplinary Science Reviews, 48(3), 446–459. https://doi.org/10.1080/03080188.2023.2193799
2. Bednarowska-Michaiel, Z., & Uprichard, E. (2025). Bringing Interdisciplinary Data Science Education Challenges into the Classroom. Journal of Statistics and Data Science Education, 34(1), 1–29. https://doi.org/10.1080/26939169.2025.2507366
3. Belteki, D., Rees, A. J., & Sichani, A.-M. (2025). Datafication and Cultural Heritage Collections Data Infrastructures: Critical Perspectives on Documentation, Cataloguing and Data-sharing in Cultural Heritage Institutions. Journal of Open Humanities Data, 11, 14. https://doi.org/10.5334/johd.277
4. Belyak, G. N. (2021). On potential development strategies for Digital Humanities. Voprosy Literatury, 4, 70–94. https://doi.org/10.31425/0042-8795-2021-4-70-94
5. Berry, D. M. (2022). Critical Digital Humanities. Bloomsbury Academic. https://doi.org/10.5040/9781350232143.ch-12
6. Casillo, M., Colace, F., Gaeta, R., Lorusso, A., & Pellegrino, M. (2025). Artificial Intelligence in Archaeological Site Conservation: Trends, Challenges, and Future Directions. Journal of Computer Applications in Archaeology, 8(1), 224–241. https://doi.org/10.5334/jcaa.207
7. Cha, J. (2023). Big Data Studies: The Humanities in Uncharted Waters. Korean Studies, 47(1), 274–299. https://doi.org/10.1353/ks.2023.a908625

8. Chen, A., Jesus, R., & Vilarigues, M. (2025). Synergy of Art, Science, and Technology: A Case Study of Augmented Reality and Artificial Intelligence in Enhancing Cultural Heritage Engagement. Journal of Imaging, 11(3), 89. https://doi.org/10.3390/jimaging11030089
9. Edvinsson, S., Mandemakers, K., & Smith, K. R. (2023). Introduction: Major Databases with Historical Longitudinal Population Data: Development, Impact and Results. Historical Life Course Studies, 13, 186–190. https://doi.org/10.51964/hlcs14840
10. Ibrus, I., Schich, M., & Tamm, M. (2021). Cultural Science Meets Cultural Data Analytics. Cultural Science Journal, 13(1), 1–15. https://doi.org/10.2478/csj-2021-0001
11. Li, W., Zhang, P., & Wang, J. (2023). Humanities Scholars' Understanding of Data and the Implications for Humanities Data Curation. Proceedings of the Association for Information Science and Technology, 60(1), 1034–1036. https://doi.org/10.1002/pra2.936
12. Mayernik, M. S. (2023). Data Science as an Interdiscipline: Historical Parallels from Information Science. Data Science Journal, 22. https://doi.org/10.5334/dsj-2023-016
13. Ncube, M. M., & Ngulube, P. (2025). Harnessing Data Analytics for Enhanced Public Programming in Archives and Museums: A Scoping Review. Heritage, 8(5), 163. https://doi.org/10.3390/heritage8050163
14. Newman, J. (2023). Promoting Interdisciplinary Research Collaboration: A Systematic Review, a Critical Literature Review, and a Pathway Forward. Social Epistemology, 38(2), 135–151. https://doi.org/10.1080/02691728.2023.2172694
15. Parti, K., & Szigeti, A. (2021). The Future of Interdisciplinary Research in the Digital Era: Obstacles and Perspectives of Collaboration in Social and Data Sciences - An Empirical Study. Cogent Social Sciences, 7(1). https://doi.org/10.1080/23311886.2021.1970880
16. Schwandt, S. (2020). Digital Methods in the Humanities. Bielefeld University Transcript. https://doi.org/10.14361/9783839454190
17. Sendra, A., Late, E., & Kumpulainen, S. (2024). Understanding Researchers' Perspectives on Work Tasks in Digital Humanities and Computational Social Sciences. Digital Humanities in the Nordic and Baltic Countries Publications, 6(1). https://doi.org/10.5617/dhnbpub.11488
18. Waters, D. J. (2022). The emerging digital infrastructure for research in the humanities. International Journal on Digital Libraries, 24(2), 87–102. https://doi.org/10.1007/s00799-022-00332-3