How can Data Science and Humanities inform each other?

"*Two polar groups: at one pole we have the literary intellectuals, at the other scientists, and as the most representative, the physical scientists. Between the two a gulf of mutual incomprehension.*"

C.P. Snow, 1959

The Two Cultures describes how "the intellectual life of the whole of western society" had become split into "two cultures"- science and humanities. This division, according to Snow, impeded both fields in solving the world's problems. More than sixty years later, does an even wider gulf still remain? More and more crossroads are formed as data science develops across the breadth of humanity as a subject. Data science is an interdisciplinary practice that utilises algorithms and models to record, store and analyse abstract data (from humanity fields); from it extract structured patterns and allow informed insights to be drawn. It creates new methods of analysis and creation in humanities by setting up dataset infrastructures, both logistically and creatively. This creates a new age of algorithm-aided arts, whether it be performing music or visual experiences. Conversely, humanities justify the application of data science, also improving inputs into such methods. Helping transform the implementation of data before and after its analysis. As such, data science interweaves itself with humanities, equally informing each other.

How data science can inform humanities:

Computational analysis of humanity data:

As part of data science, quantifiable evidence is gathered from human sources. Machine learning and artificial intelligence is applied to recognize critical information such as patterns and trends. These automatic procedures provide statistical models from often unstructured, fragmentary, ambiguous, contradictory, multilingual and heterogenous datasets. This provides a new viewpoint and subjective conclusions for humanities scholars, informing their approach on processing methods of data. Thus,

innovative discoveries and interpretations arise from otherwise separate and unreadable data points. Computational analysis is also more effective compared to traditional human-based procedures, it being less prone to error and more efficient.

The digitalisation of analysis allows for ease of cross-field collaborations, yielding new insights into pursued research questions. Lexical semantic change using computational analysis (McGillvray et al., 2019) is an emerging field. The application of diachronic textual data, neural and developing contextual embeddings[1] are used to create models and detect semantic changes. This was launched by the digitalisation of early English text collections in COHA[2] and Google N-grams. A multitude of English corpora (from newspapers to Twitter messages) has been used and harmonized with effective computational techniques. How the meaning of words changed over time could be identified in groundbreakingly streamlined models. This could even be extended in related disciplines such as lexicography or other historical classical languages. At the centre of this linguistic analysis are computational models (specifically distributional hypothesis[3]) from data science.

Dataset for digitalised texts:

Data science techniques are used to archive and collect cultural heritage data and metadata[4]. These include computational statistical analysis, search and retrieval, topic modelling and data visualization. These methods allow the processing of vastly larger data volumes than any human research group can handle. Hence, GLAM [5] can rapidly expand their digital research library. These new collections of historical and literary artefacts are more publicly available on the internet than previously possible with prints. Apart from text, visual media from pottery to paintings; and nuanced sonic resources from melodies to lost languages can be stored in these repositories. Increasing the ability for

[1] Embeddings is to convert non-numerical data (in this case context or specific words) into vectors or numeric values
[2] Corpus of Historical American English
[3] Distributional hypothesis states the relationship between the semantic similarity and distribution similarity, thus semantically similar words appear in similar linguistic contexts
[4] Data that describes other data
[5] Galleries, libraries, archives and museums

humanists to combine data sets, enabling a quick diffusion of new methods, tools and ideas cross disciplinary and cultural boundaries.

The National Archives' web library (200 TB of data) uses data science methods to underscore trust in the conversion process in the databank. Deep learning along with block chain have been implemented to convert over 1000 file types, spanning over several decades, into digital objects in Project Archangel[6]. Through content-aware hashing, file tampering and corruption was prevented. Another project was Deep Discoveries[7], in which archivists used visual search to look for patterns/visual motifs across different collections of over three million papers and catalogues. This would be impossible to describe normally in a visual collection. Additionally, this project also uses similar techniques in handwritten text recognition using the platform Transkribus. The collection of about a million wills over four centuries were transcribed and turned into data. This searchable documentation was full of people, places, connections, property and wealth, providing fascinating insight into the social history of the country.

Algorithmic creativity:

Although the source of creativity cannot be explained, it can be approximated by notions of "inspiration" and "intuition". Algorithmic creativity is an attempt to formulate this using pattern recognition, language understanding and relating inherited pieces of culture and experiences. Thus, this creates programs for artists to use, where data science acts as a creative agent for autonomous creative tasks. Some artists use data science as material to generate formal and aesthetic outcomes by modifying machine learning models' datasets and parameters to simulate and replicate creativity. Therefore, arts (in humanities) are one of the most diverse fields where data science is used: storytelling, music, poetry, painting, dancing, art preservation are all rapidly developing fields. In music, data science techniques can be used throughout the chain: from composition to performance to production and to distribution.

[6] Collomosse 2019
[7] Angelova, Mackenzie 2020

In composition, David Cope's EMI Project (1987, 1990) emulates the styles of various composers (Bach, Brahms, Debussy, Cope himself, et al.) by looking at their works and discovering recurring patterns. Such patterns ("signatures") were put into a compositional rule analyser to discover the contrasts between respective composers. The analyser included features of voice leading, repeated notes and harmonic progression and used them to create statistical models of the analysed works. Such was the method of breaking down the style and thus "creativity" of composers, then creating original compositions. In performance, Johnson (1992) developed a system which takes inputs of expert performers and outputs an ideal tempo and articulation to be played during Bach's fugues from "The Well-Tempered Clavier". Bach's music is usually a focus of artificial creativity due to its Baroque systematic structures as well as lack of rubato in performance. Such was the work of Robert Thomas (2019) and Ebcioglu's CHORAL (1993).

In visual arts, Harold Cohen's AARON (1995) is able to produce paintings from understanding the concepts of different objects. From these data points sets up parameters to produce unique paintings of those set objects. Simon Colton's Painting Fool (2015) takes a similar approach, simulating styles of digital artists found across online source materials. Colton is an example which focuses more on simulating an abstract human "creativity". These works of data science have been accepted and displayed in London's Tate Modern and San Francisco's Museum of Modern Art, alongside those of human artists.

How humanities can inform data science:

Improving quality of data:

"Garbage in, garbage out". The quality of data directly correlates with quality of the algorithm. Thus, this qualitative endeavour relates to humanities. Human data have methodical complexities, inherent biases and contextual/historical natures. Improving datasets and tackling ethical and systematic questions can improve data science practices/applications. Humanities set up a premise, giving a

better understanding of specific data and its methodological framework. Aligned to this are human issues of equality and diversity: inclusion of religions, races, beliefs, ages, gender reassignments, sexes, sexual orientations, partnerships and disabilities. Avoiding issues of digital colonialism and perpetration of existing inequalities into research data instigates positive changes in relevant data science projects.

Such is the work of D'Ignazio and Klein (2020) or Risam (2018), who seek to focus on feminism in data science. They plan to "rethink binaries and hierarchies, embrace pluralism and challenge gender binary and systems of counting/classification of data that perpetuate oppression". Allowing for representation of minoritized groups in algorithmic models, synthesizing multiple (indigenous) perspectives. Removing flawed assumptions improves the depiction of reality in databases with more complete data profiles. They strive to reduce (government) data bias under influence of the "matrix of domination[8]" which marshals real-world issues like access to resources and childcare caused by classism.

Issues with implementation:

As data science methods become more powerful and valuable, more applications will be discovered across fields. Ethical and methodical issues arise; thus, constraints and standards should be constructed to ensure a secure implementation of data science. Ethically, emerging data sources such as social media posts, subscription lists, credit ratings and location tracking increase surveillance, invade privacy and perpetuate immoral exploitation. Methodically, benchmarks should be built to assess investigation methods and examine the types of errors/biases to which they are prone. These computational results should be falsifiable, allowing the wider community to judge their robustness and replicability. Such is the work for future humanists.

Apart from macro implementation issues, suitability of computational tools in the ever-expanding arsenal is a related one. Part of these data processes (data quality, iterations and prototyping)

[8] Unequal social, historical and economic conditions

contain assumptions and generalizations; these have to be considered while developing methods for data-driven research in humanities. Communication is key between humanities and data science groups to avoid diversion of goals. Humanists have to have certain knowledge on these tools, allowing them to clarify data exploration and quantitative evidence extraction in their study methods.

"*Scientists have it within them to know what a future-directed society feels like, for science itself, in its human aspect, is just like that*."

C.P. Snow, 1962

Science, particularly data science, has an intrinsic aspect of humanity. The realization of this truth arises as we overcome the "mutual incomprehension" between science and humanities. A wealth of opportunities evolves from the joint collaborations through bidirectional informing between data science and humanities. With data scientists constructing infrastructure and processing methods that push organizational (digitalized texts) and creative endeavours (algorithmic arts) to new limits; while humanists excel in the detection and confrontation of bias, improving datasets, placing the human society at the centre of data science implementation. As both fields inform each other, the full potential of interdisciplinary research is reached. No more shall the "gulf" remain.

Word count: 1642

(title, references and footnotes excluded)

Bibliography:

Lora Angelova, Liz Fulton, in Deep Discoveries: A new way of exploring and connecting digitised image collections (2020), The National Archieves [online]

McGillivray, Barbara et al., in The challenges and prospects of the intersection of humanities and data science: A White Paper (2020), The Alan Turing Institute [online]

Mark Bell, in Machine Learning in the Archieves (2020), CogX 2020: Artificial Intelligence in the Arts and Humanities [online]

David M. Berry, in What are the digital humanities? (2019) The British Academy [online]

Tu Bui, Daniel Cooper, John Collomosse, Mark Bell, Alex Green, John Sheridan, Jez Higgins, Arindra Das, Jared Keller, Olivier Thereaux, Alan Brown, in ARCHANGEL: Tamper-proofing Video Archives using Temporal Content Hashes on the Blockchain (2019) Cornell University, Computer Vision and Pattern Recognition [online]

Cohen, H., in "The further exploits of Aaron, painter." (1995) Stanford Humanities Review 4(2): 141–15

Ramón López de Mántaras, in Artificial Intelligence and the Arts: Toward Computational Creativity (2017) The Next Step: Exponential Life, Artificial Intelligence Research Institute (IIIA), Bellaterra, Spain

Catherine D'Ignazio, Lauren F. Klein, in Why Data Science Needs Feminism (2020), in Data Feminism, Chapter 1, 2, 6, Massachusetts Institute of Technology

GDDNetwork; in the GDDNetwork Final Report (2020), AHRC Network for a Global Dataset of Digitised Texts [online]

Colton, S. Halskov, J., Ventura, D., Gouldstone, I., Cook, M., and Pérez-Ferrer, B., in "The Painting Fool sees! New projects with the automated painter." (2015) International Conference on Computational Creativity 2015: 189–196

Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, Haim Dubossarsky; in Computational Approaches to Semantic Change (2021), Berlin: Language Science Press [online]

Johnson, M. L. in "An expert system for the articulation of Bach fugue melodies." (1992) In Readings in Computer Generated Music, D. L. Baggi (ed.). Los Alamitos, CA: IEEE Press, 41–51

National Legislation Services, in Big Data for Law (2021), UK Research and Innovation, the National Archieves [online]

Roopika Risam, in Decolonizing the Digital Humanities in Theory and Practice (2018), in The Routledge Companion to Media Studies and Digital Humanities: 78-86 [online]

Rebecca Sealfon, in Is Data Science a Humanities Field? (2020), Medium [online]

C.P. Snow, in The Two Cultures, (1998), Cambridge University Press, University of Cambridge

Nina Tahmasebi and Simon Hengchen, in the Strengths and Pitfalls of Large-Scale Text Mining for Literary Studies (2019), in Uppsala: Svenska Litteratursällskapet, 2019. Vol. 140, p. 198-227