When we think about the digital humanities, we often go straight for breadth. The temptation that we have been sold by "distant reading" is that we can add ever-more texts to a corpus and that we might, thus, cheat death's ability to cut short our reading. For consider that, in 2015, according to Bowker data, almost three million new books were printed in English alone, of which two hundred and twenty thousand were novels. A good estimate for the number of days in a human lifespan is 26,000 (approximately 71 years), using the World Health Organization's figures as of 2015, so one would need to read an average of ten novels per day, every day from age ten onwards, to have read all English fiction published in 2015. The promise of distant reading is a way around this: even as we cannot hope to read all of this material, we might be able to study it computationally and statistically.

This wide, distant reading, with large corpora, has led to some anxiety. For instance, Matthew Wilkens sees digital methods – albeit referring to specific types of geographic information systems – as existing in tension with textual attention. If we deploy these methods, he claims, "we'll almost certainly become worse close readers". What does it mean, though, to be a good, bad, better, or worse close reader?[1] As Peter Middleton notes, the phrase "close reading" refers to "a heterogeneous and largely unorganized set of practices and assumptions".[2] Indeed, just as "different versions of distant reading" are not really a "singular project", in Andrew Goldstone's words, there is no singular method that constitutes close reading.[3] To many in the field of literary studies, though, this question of what we mean by "close reading" might seem so obvious as to need no answer. We are used, in the present moment, to paying close attention to the language of writers and to using the fruits of this practice to make arguments. As Jonathan Culler puts it, "examining closely the language of a literary work or a section of it, has been something we take for granted, as a sine qua non of literary study".[4] This was not always so. Indeed, in Jessica Pressman's recent assessment, mirrored by others, close reading only "became a central activity of literary criticism" in the "modernist" period.[5] That said, although the discipline of "English language and literature" is relatively young, being founded in 1828 at University College London, it can feel surprising, from our contemporary vantage point, that it took until the modernist period for close reading to develop.[6]

The spatial relationship between the metaphors of closeness and deepness, of proximity and profundity, in reading practices has never been entirely clear but it has certainly been the subject of debate. As Nancy Armstrong and Warren Montag note, even the canonical figures of the digital field "won't let us construe the distance implied by distant reading in opposition to the closeness and polysemy of literary language".[7] Yet, *depth* is also a type of scale. The underwater Mariana Trench is huge, despite the fact that it is deep. What might it mean, then, to think distantly (in terms of reading) at depth? Are there ways in which computational techniques might act as a digital *microscope* that makes huge and multiple what was previously small and singular?

---

1 Matthew Wilkens, 'Canons, Close Reading, and the Evolution of Method', in *Debates in the Digital Humanities*, ed. by Matthew K. Gold (Minneapolis, MN: University of Minnesota Press, 2012), pp. 249–58 (p. 256) <http://dhdebates.gc.cuny.edu/debates/part/5>.

2 Peter Middleton, *Distant Reading: Performance, Readership, and Consumption in Contemporary Poetry*, Modern and Contemporary Poetics (Tuscaloosa, AL: University of Alabama Press, 2005), p. 5.

3 Andrew Goldstone, 'The *Doxa* of Reading', *PMLA*, 132.3 (2017), 636–642 (p. 641).

4 Jonathan Culler, 'The Closeness of Close Reading', *ADE Bulletin*, 2010, 20–25 (p. 20) <https://doi.org/10.1632/ade.149.20>.

5 Jessica Pressman, *Digital Modernism: Making It New in New Media*, Modernist Literature & Culture, 21 (New York, NY: Oxford University Press, 2014), p. 11.

6 Ted Underwood, *Why Literary Periods Mattered: Historical Contrast and the Prestige of English Studies* (Stanford, CA: Stanford University Press, 2013), p. 81; see also Franklin E. Court, *Institutionalizing English Literature: Culture and Politics of Literary Study, 1750-1900* (Stanford, CA: Stanford University Press, 1992); and Gerald Graff, *Professing Literature: An Institutional History* (Chicago, IL: University of Chicago Press, 1989).

7 Nancy Armstrong and Warren Montag, '"The Figure in the Carpet"', *PMLA*, 132.3 (2017), 613–19 (p. 617).

The project that I have been working on for some time is called *Close Reading with Computers*. I am turning to the ways in which computational methods might be used to study single texts and to reintegrate the data findings of these explorations with more traditional literary-critical practices. For, as Lisa Gitelman puts it in the title of her edited collection: there is no such thing as "raw" data; all data-driven processes require hermeneutics. I would also note that I am hardly the only scholar working in such a space of textual analysis.

While critics claim that the straw man argument of "counting words" can tell us nothing of literature – see the recent opinion piece by Timothy Brennan – we have long been accustomed to thinking with numbers. For instance, as Erik, on this panel, pointed out to me, Thomas Schaub claimed in 1981 that "the word 'bloom' is one of the most oft-repeated words in" Thomas Pynchon's *The Crying of Lot 49* (although he's actually wrong). Dartmouth College offered a course entitled "Literary Analysis by Computer" as far back as 1969.[8] Further, as Nicholas Dames has noted, Vernon Lee proposed a "statistical experiment" – a quantitative analysis – on literature in her 1923 *The Handling of Words*, itself prompted by a letter to *The Times* from Emil Reich, several years earlier.[9] Quantitative approaches to fiction – up close – are really not that new.

So what am I actually doing? I want to give a few examples of the type of work that I am pursuing at present, mostly centring around David Mitchell's *Cloud Atlas*.

[EXPLAIN *CLOUD ATLAS*]

[SLIDE]

- I want, first, to speak of the possibilities for textual genetics, textual scholarship or criticism more broadly and visualization in this space
- Describe working in contemporary fiction
- Describe working on *Cloud Atlas* and plot
- Describe "diff" methodologies and visualizations

[SLIDE]

- However, we are entering a difficult dark age of documentation for scholarship
- Manuscript culture is on the way out, despite Don DeLillo and Jennifer Egan etc.
- Talk about unpublished works and selectivity of literary market
- Talk about ability to preserve work based on economic selectivity

[SLIDE]

- I also want to talk about texts as statistics; another horizon opened up by digital possibilities
- Distant reading as telescope, close as microsope
- Talk about trigram PoS visualisation

[SLIDE]

---

8 Annette Vee, '"Literary Analysis by Computer" Offered at Dartmouth, Winter 1969, Working with Paradise Lost. #1960sComputing Pic.Twitter.Com/DPrnY23cpU', *@anetv*, 2017 <https://twitter.com/anetv/status/919219418189660160/photo/1> [accessed 18 October 2017].

9 Nicholas Dames, *The Physiology of the Novel: Reading, Neural Science, and the Form of Victorian Fiction* (Oxford: Oxford University Press, 2007), p. 188; Vernon Lee, *The Handling of Words* (London: The Bodley Head, 1923) <https://gutenberg.ca/ebooks/lee-handling/lee-handling-00-h.html#ch_VI> [accessed 2 November 2017].

In *Cloud Atlas*, the first chapter purports to have been written between 1850 and 1910. The first section of this chapter consists of 13,246 words. I wanted to know: what does it mean to "write as though you are writing in the 19th century from a twenty-first century perspective"? Is it about mimetic accuracy of the language that is used?

To begin answering this, I wrote a computer program that looked up etymological first-use dates of terms within the novel in Dictionary.com and the OED. It wasn't perfect, but its operation can be summarised as: "return as many as possible, but not necessarily all, words in a text that have etymological first-usage dates after 1910".

The upshot of this is that I found three terms that would have been inaccessible either to Mitchell's historic author or the intra-diegetic editor: spillage, from ~1934; latino, from ~1946; and lazy-eye, from ~1960. In the case of spillage, the text is here recounting the debate between the Moriori elders as to whether "the spillage of Maori blood" will "also destroy one's *mana*". Interestingly, the Online Etymology Dictionary disputes this entry, claiming it for the nineteenth century.[10] Mitchell could have avoided this slip through reverting to the verb form, "spilling". On the other hand, latino is definitely a twentieth-century construction: "'Passionate Latinos,' observed Henry, bidding me a second good-night". While this term did not actually come to prominence until after the Second World War, the use, here, of a racial epithet has an important different effect for the construction of a stylistic imaginary of the nineteenth century, to which I will turn shortly. Finally, Mitchell gives us a "parlour […] inhabited by a monstrous hog's head (afflicted with droop-jaw and lazy-eye, killed by the twins on their sixteenth birthday". The sources that I consulted give this slang term for amblyopia as appearing in the middle of the twentieth century.

While this was a very good attempt at linguistic mimesis, it clearly also busts the logic that it has to be accurate to "sound right". On the other hand, I also took a 2004 magazine corpus, called COCA, and examined words that appear in *Cloud Atlas* that are not present in contemporary magazines. The most striking finding here was that terms of racist abuse – or at least outmoded terms with colonial overtones – Blackamoor, blackfella, darkies, harridan, womenfolk, bedlamite, mulatto, quadroon, and mixedblood – all occur significantly more frequently in *Cloud Atlas* than in contemporary popular discourse.

[SLIDE]

- Talk about difficulties
- Talk about born-digital texts with DRM
- Copyright exemptions for researchers
- EU Directive 2001/29/EC
- Re-keying or feed scanning

[SLIDE]

- Final: natural language generation
- Describe character-based recurrent neural network
- Examples:

10 'Spillage', *Online Etymology Dictionary*, 2017 <https://www.etymonline.com/word/spillage> [accessed 2 December 2017].

    - ‘The problem’, as the network aptly phrased it, ‘is that the poem is a construction of the self as a strategy of self-consciousness and context’
    - They ‘provide the fraud of the epistemological practices of knowledge’
    - ‘I shall find our intellectual values, by rewriting their very ties’
    - ‘John Spottisley, ‘The privatized climax’. (1929), p. 4, emphasis in original’
    - ‘see David Pillar, New Bibliography, ed. Donald Davis (London: Lawrence & Wishart, 1979)’
- Natural language generation already used in sports writing