11
Squinting at a Sea of Dots: Visualising Australian Readerships Using Statistical Machine Learning*

Julieanne Lamond and Mark Reid

When we presented this chapter at the Resourceful Reading conference, a question was posed to us expressing concern that Australian literary studies faces the prospect of becoming lost in the data. Several papers at the conference argued for a shift in focus away from canonical texts and authors towards an examination of Australian literature as a field, a network, a broader structure. Our interlocutor suggested that such a bird’s eye view results in a meaningless constellation of dots: we can make no sense of them. This question goes to the heart of an issue relevant to employing empirical methods in literary studies. One reason critics have been arguing for a more empirical approach to Australian literary studies is that we have access to new and much broader kinds of data than ever before. Data, however, are of little use in and of themselves. The key question when approaching literary studies with empirical methods is how to move between the generalisations involved in empirical research and the

* The authors would like to thank Tim Dolin for generously providing access to the ACRP database and Jason Ensor for his expertise and time answering technical questions about its organisation.
attention to the particular that characterises literary analysis: in other words, how such data could be made useful to literary analysis? This chapter examines one such approach. Specifically, it uses a collaboration between Australian literary studies and statistical machine learning to suggest how, in practice, empirical modes of research can speak to, enhance, or even help to direct more traditional modes of literary analysis.

As Katherine Bode has argued, the cultural materialist claims that tend to be made in Australian literary studies often imply a bird’s eye view that is not always present in the research that generated them. There is an impulse within the discipline to make broad claims about the nation, the era or literary field based on close literary analysis of a few texts by a few authors and the rationale for basing these arguments on these texts and authors is not always clear. Instead of focusing on a series of canonical authors and texts to make claims about the national literary field, Bode suggests we consider starting from the other end, to use the evidence of what was published and read, where and when, to determine which texts and authors might best warrant close critical attention. Instead of moving from the particular to the general, as has been our wont in the discipline, we should attempt the more methodologically sound approach of moving from the general to the particular: to use empirical evidence to direct our critical attention. We have some tools at our disposal to begin to approach this task: the AustLit database and the Australian Common Reader project are among them.

Large datasets raise new difficulties for literary critics and cultural historians—how do we make sense of these data? We know there must be useful information in this mass of data: how do we extricate it? There are two primary ways in which literary scholars use such databases: by

---

making particular queries of the database or by creating summary statistics for the datasets as a whole. Each of these has its pitfalls. Specific queries are useful for viewing local relationships and small slices of the data but do not give us a sense of their relationship to the whole. Summary statistics such as average readership, books read and published over time offer us some insight but are ‘coarse’ summaries: they jettison a great deal of information in the extent of their generalisations. There are more sophisticated way of viewing a dataset as a whole without throwing away as much information, and this is where interdisciplinary collaboration becomes useful. One of the aims of this chapter is to draw attention to a field of research—statistical machine learning—that dedicates itself to doing what literary critics are not always very good at: drawing inferences from large datasets.

**Collaboration**

It is a cliché these days to say that we are overwhelmed by information. The recent and dramatic reduction in the cost of collecting, storing, copying, transporting and processing large amounts of data has meant individuals and organisations have had to rethink how they make sense of it.² The growing body of research within the field of statistical machine learning, and the increasing interest in it from outside its walls, are responses to this deluge of data.

Tom Mitchell, head of the recently formed Machine Learning Department at Carnegie Mellon University, describes the broader field of machine learning as ‘a natural outgrowth of the intersection of Computer Science and Statistics’.³ It is a large discipline that concerns itself with

---

questions of how to generalise, or ‘learn’, from examples in a computationally efficient manner. This encompasses problems in robotics, speech recognition and computer vision. Statistical machine learning is a subfield of machine learning that emphasises the theory and practice of statistical inference and so is primarily concerned with problems of prediction, modelling and hypothesis testing. Researchers within this discipline invent and study tools that help us to locate needles of knowledge in our ever-growing haystacks of information. Here we would like to provide a case history of the particular collaboration we used to explore the possible usefulness of statistical machine learning for Australian literary studies.

This project had its genesis in attempts to use a particular book history dataset, the Australian Common Reader project (ACRP), to think about how to make useful generalisations about Australian readerships in the late nineteenth century. The ACRP is a database collating loan records from six Australian libraries, kept intermittently for the period 1861 to 1912, alongside some biographical information about the borrowers. For our purposes, the database contains two main axes of information: about books—which libraries held them, how often were they borrowed, and by whom? And about borrowers—what did they borrow and when?

There are limitations to making generalisations from this kind of data. Whether a collection of data is suitable to support a generalisation made by a human or machine depends on its size, quality, how well it represents the subject of the generalisation and the validity of any assumptions made about it. As Tukey puts it, ‘The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.’

readerships, the data found in the ACRP database have some of these limitations. It is not comprehensive; it does not include all Australian libraries and the records for the periods covered are intermittent. It is a collection of records about borrowing, not reading. Although we deliberately conflate borrowers with readers in this chapter as a convenient shorthand, it is clear from the pattern of individual borrowers that some were likely to be borrowing for others. This is evidenced by records of individuals borrowing across a range of genres including, for example, children’s books. However, the number of borrowers of a work seems a reasonable approximation to its readership. There is also a limit to how far we can generalise from such data, unless we are clear that we are generalising about particular, local, reading communities, and patterns that exist across these communities.

Julieanne began using the database to ask a narrow set of questions. She was interested in what the ACRP might reveal about the extent of Australian readerships for two particular authors, Rosa Praed and Steele Rudd. This involved making simple queries of the database and resulted in straightforward, quantifiable results: numbers of people who borrowed each of Praed and Rudd’s books. From these data, she could ask whether these authors shared a readership. This involved some laborious searching and recording of data to cross-reference the borrowers of each but resulted in the interesting finding that these very different authors did, in fact, share a readership as defined in relation to these data. In doing so, she was taking up Tim Dolin’s suggestion that such databases could be used to compile ‘Amazon-style’ lists of what works particular people borrowed in common: using his example, ‘of the x readers who borrowed Jane Eyre, y also borrowed The Mill on the Floss, Wuthering Heights and so on’. Dolin continues: ‘The outcome of this list is something more than a micro-canon, those works given special status by a particular reading community. What it shows up is a series of distinctive patterns and corre-
spondences among literary works, patterns and correspondences that come into existence for that reading community at that time.\(^5\)

Dolin’s use of the database is one means of going to the data to drive novel, contextualised close readings of particular texts by considering not only who read them and when, but what other works were read in common with them. In this methodology novels are analysed in relation to how they are read ‘within the immediate horizon of other works’.\(^6\) These patterns of readership provide the basis for ‘locally situated re-readings’ of novels: close analysis that is prompted by but not limited to the results of such quantitative data searches. For example, Dolin uses the ACRP data to consider a cluster of books ‘within the horizon’ of Dickens’s *Great Expectations* in colonial Adelaide in 1861 and 1862. This results in a re-reading of the novel in light of the circumstances of colonial Adelaide, and with an eye to resonances with the other novels that were popular with those who read *Great Expectations* in Adelaide at the time, including Eliot’s *Silas Marner* and Bulwer-Lytton’s *What Will He Do With It?*

This approach raises the possibility of using such databases to ask questions about reading communities: people who read similar books in similar geographical or temporal contexts. Dolin’s method involves asking narrow questions of the data, driven by interest in particular texts. We began to think about how we could use the ACRP to think about reading communities more broadly: that is, how to look at the form and structure of reading communities across the whole database, or across particular libraries within the database, without bounding our queries with a narrow focus on particular texts or authors. Could we discern communities of readers or patterns of readership in these library records? Is


\(^6\) Ibid.
there another way, other than querying particular books or particular borrowers?

When Mark heard about the ACRP database and the types of questions Julieanne was asking of it, he saw an opportunity to apply methods from his research area to help refine and answer them. When faced with a new dataset that has been collected by a third party, standard statistical practice is to first perform what Tukey calls ‘Exploratory Data Analysis’. This typically involves creating various summaries of the data such as counts, averages, histograms and other graphs in order to extract a high-level overview. Of these, graphical summaries tend to provide the most insight. This is because, as Ware argues, ‘the human visual system is a pattern seeker of enormous power and subtlety. The eye and the visual cortex of the brain form a massively parallel processor that provides the highest-bandwidth channel into human cognitive centers. At higher levels of processing, perception and cognition are closely interrelated, which is the reason why the words “understanding” and “seeing” are synonymous.’

As Julieanne was most interested in questions of readership, an appropriate form of exploratory data analysis was to create a visual summary of the ACRP database that presented a map of the works within it so that proximity was indicative of shared readership. By doing so, visually apparent clusters of works in this map would immediately suggest micro-canons worthy of further examination.

_____________________________________________

Maps via Dimensional Reduction

At the time of writing, the database we had access to contained 99,692 loans of 7078 different works from six libraries by one of the 2643 people. To make this more manageable, we focused on popular works that were borrowed at least twenty times and only considered loan records for these. This distilled the database down to what we will call the *reader-ship table*, with each row representing one of 1616 works and each column representing one of 2474 borrowers. Each cell in the readership table contains either a 1 or a 0, with a one indicating the work corresponding to the cell’s row was lent to the borrower corresponding to its column.

The central insight that led to our method of visualisation is that this table of works and their borrowers summarises most of the information in the ACRP database pertaining to readerships. Our goal was to display the information in this table in a way that allows for both a general overview of all works in the database that suggested their shared readerships at a glance, and provides the ability to drill down to examine the details of specific works.

The machine learning technique we used to construct our map of the ACRP works is known as *dimensional reduction*. These are procedures that transform high-dimensional data into a low-dimensional representation while preserving salient information in the data. In the case of the ACRP data, each work has a high-dimensional representation as the sequence of 2474 1s and 0s appearing in its row in the readership table. For the purpose of visualising readerships, the salient information in this

---

9 These are loans for which there is a valid library, work, borrower and loan identifier in the database. This means the numbers here are slightly lower than the number of entries (including incomplete ones) in the database.

10 This number was chosen more or less arbitrarily with the aim to make data processing easier.
high-dimensional representation is the similarity of two works’ readerships. This similarity is quantified as the number of common borrowers they have—that is, the number of columns for which both works’ rows contain a 1. Mathematically, this is known as the inner product between the rows. To construct a map of the works that could be displayed on a screen, it was necessary to reduce this high-dimensional representation to a two-dimensional one that best preserved these inner products for all possible pairs of works.

The algorithm that performs the dimensional reduction is the recent t-distributed Stochastic Neighbour Embedding (hereafter t-SNE) algorithm developed by Laurens van der Maaten and Geoffrey Hinton.\textsuperscript{11} This algorithm is particularly well suited for the visualisation of high-dimensional datasets. To interpret its results, the details of t-SNE are not as important here as a high-level understanding of what the algorithm does to the ACRP data. Consider the case of a simpler readership table where there are only three rows for the works A, B and C. Suppose each has 100 readers, and that A and B share thirty readers while A and C share ten. A good dimensional reduction will display the works A, B and C so that A and B are closer together than A and C. The actual location of each work on the screen is not as important as its relative position. Suppose we also know B and C share twenty readers. Then B and C should be displayed closer together than A and C but not as close together as A and B.

As we consider larger numbers of works and their overlapping readerships, the number of constraints on the relative distances between works grows rapidly, and determining an arrangement on the screen that respects all of these constraints becomes increasingly difficult. Dimensional reduction algorithms such as t-SNE use sophisticated

techniques to find good arrangements of the works that approximately satisfy a given set of constraints. It is necessary to emphasise that any visualisation produced by these techniques is only approximate and by no means the only one possible. In other words, different algorithms will layout the books in different ways and the particular layout used in our visualisation tool is but one possible way of visualising these overlapping readership relationships.

**Using the Visualisation Tool**

The arrangement of works produced by the t-SNE algorithm is nothing more than a collection of screen coordinates—one for each work in the database. In order to more easily interpret this output, Mark developed some software that allows a user to interact with this map of the database. The software he used is freely available over the web and works with most modern operating systems (Windows, Apple OS X, Linux). It can be accessed through a web browser at mark.reid.name/code/acrp/.

Figure 1 shows a high-level overview of the ACRP database. Each circle represents a literary work. The size and colour of each circle is intended to give a sense of the size of the readership for the corresponding work: larger, darker circles for works with larger readerships and smaller, lighter circles for those with fewer readers. As described above, two circles in close proximity is indicative of a high proportion of shared readers.

Hovering the mouse cursor over a circle will reveal the title of the corresponding work with the size of its readership in parentheses. Clicking on a circle—for example, (A) in Figure 1—will reveal details of the work in the location (B), including its author’s surname and date of publication. Grey lines are also displayed, connecting the selected work to others that have a shared proportion of readers greater than the threshold controlled by the ‘Similarity’ slider (F). In the example in Figure 1 the threshold is 0.25, meaning grey lines are only drawn between
the selected work and other works with more than 25 percent of their total readership in common.

There are several controls for filtering which works are displayed. Search terms can be entered into the ‘Title’ field (C) or the ‘Author’ field (D) restricting the displayed works with a title or author matching the term. The ‘Borrowers’ slider (E) sets a threshold for the minimum number of readers a work must have to be displayed. By increasing this, less borrowed works are hidden from view, allowing the user to quickly identify popular works. Finally, the drop-down list of libraries (G) can be used to restrict the visualisation to only those works that appear in a selected library. Figure 2 shows how this drop-down list can be used to show only those works appearing in the Port Germein Institute.

**Figure 1** A Screenshot of the Visualisation Tool (Labels A–G highlight different aspects of the user interface that are described in the text)
Figure 2 A View of the Books Contained in the Port Germein Institute Library (The drop-down list at the top is used to select which library is displayed)

Areas of the visualisation may be selected and zoomed in on by clicking and dragging the mouse over an area of interest. Figure 3 shows the result of selecting the region shown by the dashed rectangle in Figure 1. This zoomed area reveals a tight cluster of works connected by grey lines indicating a shared readership with Myrtle Reed’s *Master of the Vineyard* of at least 35 percent. For later reference, the labelled circles represent the following works: (A) Waller’s *Flamsted Quarries*, (B) Richardson’s *The Lead of Honour*, (C) Williamson’s *The Motor Maid*, (D) Stratton-Porter’s *The Girl of the Limberlost*, (E) Barclay’s *Through the Postern Gate*, (F) Bindloss’s *Hawtrey’s Deputy*, (G) Cooke’s *The Girl who Lived in the Woods*, (H) Yorke’s *Patricia of Pall Mall*, (I) Barclay’s *The Mistress of Shenstone* and (J) Barclay’s *The Rosary*. 
This tour of the features of the ACRP visualisation is intended to demonstrate how it may be used to survey the ACRP database quickly and identify clusters of works with shared readerships. It is important to realise that visualising data in this way does not guarantee meaningful patterns will emerge, or that all apparent patterns are meaningful. We wish to emphasise that our ACRP visualisation tool is just that, a tool. Once an interesting pattern is identified careful inspection by a knowledgeable expert is still required to determine whether it is meaningful.

**Using the Visualisation for Literary Analysis**

This visualisation provides a methodology for approaching large datasets by moving between the general and the particular. Instead of asking particular questions of the data, or making broad generalisations about it, this tool enables us to spot patterns in the data which might then warrant further exploration. In other words, it enables us to do something like exploratory data analysis: to explore the data in order to come up with more specific questions or hypotheses we can investigate in more detail. Instead of asking, as Dolin does, what borrowers of *Great Expectations* also borrowed in common, the visualisation asks this one question—what other books were likely to be read by readers of this book?—of all the books in the database simultaneously. The aim, then, is to find interesting and unexpected relationships between books, or communities of readers. These relationships can then be queried in more qualitative ways, as Dolin has done with *Great Expectations*, or tested with other methods of quantitative analysis. The point is that our investigations will not be driven by interest in particular readers or texts, but by interest in patterns and relationships between books and readers. This kind of visualisation can enable us to begin using data to develop research questions, and perhaps begin to look at relationships between non-canonical books and authors in particular Australian reading contexts.

By allowing us to see how a book relates to other books in a reading community, this visualisation can be used to approximate the shape of a
local readership, and to enable further questions, such as: are there patterns in the kinds of books people borrowed in common? Are there clusters according to genre, nationality of author, gender of borrower? How do these clusters pattern themselves temporally? Are there waves of readership of particular books or genres, or trailblazing readers? What is the relationship between the borrowing of periodicals and the books reviewed in them? Statistical machine learning enables us to ask such questions of the library data as a whole, rather than in relation to individual books and authors, and to look at more complex relationships between different aspects of the data.

The following examples of how the tool could be used by literary critics or cultural historians are introduced to raise questions rather than answer them. But this is the point—databases are not just a means of answering questions, but of posing them, or finding new questions to ask of the novels and their contexts. We might start with the visualisation of the data as a whole and look for patterns and clusters, then drill down to look at these in more detail. The pattern that is most immediately apparent in looking at the data as a whole (Figure 1) is the large clumping or clusters the books form. The first question to be asked of these is, do they relate to individual libraries? By selecting individual libraries on the visualisation we can quickly confirm that this is, to some extent, the case. Unsurprisingly, readership communities were bounded by the availability of books at particular libraries. More notably, holdings at the libraries in the database are relatively discrete. This finding takes us back to the data, further investigation of which reveals that nearly half the works in the database were found exclusively in a single library. About 30 percent had copies in two libraries, 14 percent were found in three libraries, 5 percent in four libraries, and less than 1.5 percent in five and less than 1 percent of works (five works in total) in all six libraries. This unequal spread could lead the book historian to ask questions about patterns and processes of distribution to such libraries in the period covered by the data.
We can look at other clusters in more detail by zooming in on them, as described above. This method reveals some other expected clustering, for example, by author: it is unsurprising that readers who borrow works of one author are likely to also borrow other works by that author. For example, there is a cluster with a large degree of similarity and high number of borrowers of two of Anthony Trollope's novels, *Three Clerks* and *The Warden*. These texts also had a high degree of similarity with Eliot's *Silas Marner* (also not altogether surprising as these were popular and well-read texts). More interesting are clusters of books which, on first sight, appear to have little in common with one another: all three of these novels are clustered with *Paul Fane*, the only novel by American editor and poet Nathaniel Parker Willis. Why might these four novels have shared a readership? This is a question not easily answerable only by
reference to the data: it sends us back to the texts themselves, and to their contexts.

Many of the clusters apparent in a brief glance at the visualisation involve texts not usually subject to literary analysis. However, the extent of their shared readership across borrowers in these data renders them interesting. One of the most readily apparent clusters which does not relate to a single author centres around Myrtle Reed’s *Master of the Vineyard*, a domestic romance published in America in 1910 (Figure 3). In the Rosedale Library data, this book shares a high degree of similarity with an eclectic collection of novels: Harold Bindloss’s *Hawtrey’s Deputy* (1911) and *The Protector* (1919); adventure novels set in the Canadian Northwest; bestselling English romance novel *The Rosary* (1909) by Florence L. Barclay; *The Motor Maid*, a British motoring novel by Alice Muriel Williamson (1910); and *The Girl of the Limberlost* by American naturalist Gene Stratton-Porter (1909). The only common feature immediately apparent here is that these novels were all published between 1909 and 1911, confirming a high readership of contemporary work in these data as noted by Dolin.\(^\text{12}\) But such clusters also raise the question of what else might account for these novels having a readership in common? Is it something to do with the texts, the circumstances of their publication or circulation, or the readers’ circumstances? Again, answering such questions takes us back not only to the data but to the texts themselves and their particular contexts.

**Conclusions**

While such a visualisation tool lends itself to the shift in emphasis argued by proponents of empiricism away from a focus on canonical authors, it also offers a way around the simplistic opposite of this approach: that is, focusing on the most popular authors or texts. It offers another criterion

\(^{12}\) Ibid.
for assessing the significance of a text that is not inherent either in the text itself or in the extent of its readership: its relationship to other texts and readers. Interdisciplinary collaboration of this sort can do more than provide methodological tools; it can suggest alternative conceptual frameworks for posing problems within our discipline. In this case, the emphasis in statistical machine learning on defining similarity has been particularly useful. Literary analysis tends to group texts according to categories such as period, author, genre, subject matter, nationality and form. This kind of data analysis can get us thinking about grouping books (thinking of them as similar) according to their readerships in more complex ways than simply looking at their relative popularity. We can look at relationships between local and temporal readerships, and readerships across time, as well as readerships grouped by class or gender, physical or temporal proximity.

If we are to make use of the data now available in relation to Australian literature and cultural history, we need to find ways to make them meaningful to our work. In order to do so, we may well need to venture outside our own discipline to seek help. This collaboration suggests that working with machine learning and other fields of statistical analysis can help not only to find answers to questions we might pose of the data, but to formulate questions and, further, to reconceptualise the kinds of questions we are able to ask of the data in the first place. It also suggests that co-opting scientists and scientific methods to our cause does not limit us to quantitative analysis alone. So although this visualisation does represent the Australian literary field, or at least one corner of it, as a sea of dots, it certainly does not reduce the field to the data, or result in meaningless generalisation. Rather, it provides a way for literary scholars to find patterns in the mass of data and use these patterns to return to the realm of the particular in which we are most comfortable: to the use of detailed literary and contextual analysis to seek explanations and find meaning in the patterns.