Participations
Journal of Audience & Reception Studies

Volume 16, Issue 1
May 2019

The limits of Big Data for analyzing reading

Simon Peter Rowberry,
University of Stirling, Scotland

Abstract:
Companies including Jellybooks and Amazon have introduced analytics to collect, analyze and monetize the user's reading experience. Ebook apps and hardware collect implicit data about reading including progress and speed as well as encouraging readers to share more data through social networks. These practices generate large data sets with millions, if not billions of data points. For example, a copy of the King James Bible on the Kindle features over two million shared highlights. The allure of big data suggests that these metrics can be used at scale to gain a better understanding of how readers interact with books. While data collection practices continue to evolve, it is unclear how the metrics relate to the act of reading. For example, Kindle software tracks which words a reader looks up, but cannot distinguish between accidental look-ups, or otherwise link the act to the reader's comprehension. In this article, I analyze patent filings and ebook software source code to assess the disconnect between data collection practices and the act of reading. The metrics capture data associated with software use rather than reading and therefore offer a poor approximation of the reading experience and must be corroborated by further data.

Keywords: Reader Analytics; Amazon; Kindle; Ebooks; Big Data; Critical Code Studies; Patents

Introduction

The rise of ebooks since the Kindle's release in 2007 has drawn increased scrutiny towards digital publishing surveillance practices. Mark Davis argues "the urgent question is how to *popularise* a progressive critique of digital network media. In the case of e-books, such a project would seek to open up possibilities beyond a corporatised public culture in which books are reduced to little more than big-data surveillance tools."[1] In this article, I argue that the ebook is a poor "big data surveillance tool" due to the material and technical constraints of digital publishing platforms. Quantifiable metrics of consumption (opening and closing an application, time spent on a page, progression) fail to capture the complete reading process from viewing characters to decoding and interpreting the text's meaning, even when tackling a straightforward text such as a road sign.[2] Viewing is easier to model computationally than the latter two stages, while revealing less about individual readers.

The popular imagination of big data in publishing relies on a blind faith in numbers rather than exploring algorithmic fissures. For example, bestseller lists rank books according to total sales, masking editorial interventions including the removal of select book genres or predicting independent book sales through sampling a secret selection of stores. Despite these hidden complexities, bestsellers are assumed to be relatively straightforward to determine. Terje Colbjørnsen's work on E.L. James's *Fifty Shades of Grey* and Marie-Pierre Poult's case study of Zadie Smith's *White Teeth* demonstrate the complex relationship between literary or popular taste, the marketplace, and emergent reading platforms including social media.[3]

In this article, I explore Amazon's use of big data in the Kindle ecosystem to assess the viability of in-app metrics as a proxy for tracking the reading process. My analysis draws upon patent analysis, Critical Code Studies (CCS), and corpus linguistics. Studies of algorithmic and data cultures often focus on technology companies such as Facebook and Google whose proprietary algorithms are in flux and are hidden from public view. These limitations can be mitigated through analyzing publicly available data from patents and source code. Patents demonstrate intent and signal perceived importance rather than offering concrete evidence of techniques currently in use. Technology companies use patents to outline techniques and methods they believe to be valuable enough to make public to protect. A patent is not just an offensive and defensive mechanism for large technology companies suing one another, it also rewards employee innovation, increasing retention in a fiercely competitive job market.[4] Patents provide insight into the processes that media technology companies were considering at the time the patent was filed. I also draw upon critical code studies, which "does not festishize the code […] but engages the code as an artifact contextualized in a material and social history."[5] Amazon's corporate secrecy does not extend to its use of open standards such as HTML or Linux as the foundation of its hardware, revealing intention behind design and algorithmic choices. Finally, I use corpus linguistics techniques to identify keywords in Amazon's patent filings to reveal the company's focus in reading analytics when documenting Research and Development work.

Discourses of Big Data

The term "big data" is a misnomer, as innovations in the field are less concerned with the size of the dataset than innovative analysis at scale. In an early article defining the field, Adam Jacobs argues that "the pathologies of big data are primarily those of *analysis*."[6] For example, Bowker's *Books in Print* or OCLC's WorldCat catalog are datasets with millions of entries that pre-date the term "big data," but without sophisticated analysis they cannot provide evidence about larger trends in book sales or patron reading habits. Nielsen BookScan data, which records book sales in the United Kingdom, might reveal a rise in the average sales value of hardback books over a 20-year period, but without the context of publishing trends and the economic changes over two decades, the data only provides a partial narrative.[7] A comparative analysis might also introduce new insights: "the largest *cardinalities* of most datasets – specifically, the number of distinct entities about which observations are made – are small compared with the total number of observations."[8] A book's potential audience outnumbers the volume of unique bibliographic objects, and readers produce data every time they open a book. The reading process generates more data than analyzing connections between books.

This article extends the scholarship on Critical Data Studies that challenges the normative representation of big data as a net positive by interrogating the limitations of a data-driven approach.[9] In particular, I critique the limits of reading "*infomediaries*: organizations that monitor, mine and mediate the use of digital cultural products […] as well as audience responses to those products via social and new media technologies."[10] The switch to digital reading platforms enables surveillance practices, which led the Electronic Frontier Foundation (EFF) to release a list of data collection processes of major ebook platforms. The report noted a range of data collection activities from third-party sharing to more benign practices including mechanisms to correct inaccurate data.[11] Such reports do not question the transformative nature of big data but highlight over-reaching data collection practices that may not benefit the consumer. Instead, I follow Sarah Pink *et al*'s approach of "focusing ethnographic attention towards demystifying how digital data is constituted *in situ* [to] better understand data futures, [and] account more fully for the incomplete, contingent and fractured character of digital data."[12] This article presents an analysis of Amazon's Kindle data *in situ* to demonstrate how reading metrics do not map the reading experience and how these metrics cannot document reading.

Data surveillance within publishing is often hidden but several anecdotal reports reveal the potential for readership data to influence decisions in the book trade. For example, Coliloquy, a now-defunct publisher of "active content" interactive fiction for the Kindle, modelled "the 'ideal' male hero" and created books of the "average length preferred by readers, with characters who are readers' preferred ideal types."[13] Coliloquy tracked readers' engagement with character types to tailor content and increase retention. Such personalization requires a critical mass of demographic and consumption data to successfully model user preferences and experiments such as the Netflix Prize demonstrate advanced algorithms do not necessarily lead to better results.[14] Coliloquy went through several restructurings before its acquisition by Pronoun, which shut down during 2017.

Outside of large technology companies, the publishing industry has failed to cope with the demands of big data analysis. As a result, the book trade lags beyond other creative industries in data-driven analysis of consumption. The rise of social media sites with audience-facing metrics including Facebook, Instagram, and Twitter has led to the rapid adoption of "engagement" metrics (likes, shares, follows) as a proxy for popularity in assessing acquisitions for music labels.[15] The rise of "SoundCloud rap," named after a group of likeminded underground rappers' preference for SoundCloud distribution prior to official releases in the late 2010s, demonstrates how user engagement can drive trends in mainstream music.[16] Beyond social media, start-ups such as Spotify and Netflix have embodied the potential of data-driven recommendations, acquisitions and commissioning to disrupt media industries.[17]

Publishing studies scholars have been more pro-active in analyzing digital reading on both a small and large-scale including case studies of GoodReads, fan sites and, Amazon reviews and recommendations.[18] Unfortunately, the most visible trend within the scholarship focuses on a binary of print versus digital reading that relies on participants' perception of comprehension rather than observing the reading process.[19] Perception studies run the same risks as purely data-driven analysis in reducing reading to preferences. Jodie Archer and Matthew Jockers's *The Bestseller Code*, published by Archer's former employer, Penguin Random House, is the most public intervention into the potential of big data for publishing within the industry. Archer and Jockers argue that it is possible to identify the patterns and themes across bestselling novels including "closeness" through computational analysis. The algorithm identifies "closeness" as the core theme of *Fifty Shades of Grey* demonstrating a convergence towards homogenous themes rather than contextualizing E.L. James's success.

The duo have since launched a consultancy, Archer Jockers, that uses the algorithm to assess the quality of an author's manuscript according to "narrative time," "thematic analysis," and "major themes" before publication, allowing authors to meet market expectations.[20] Archer and Jockers assert that the content of the manuscript alone will dictate its success within the literary marketplace (an intentional constriction: the algorithm does not work with nonfiction). In response to Archer and Jockers's positivist approach to textual analysis as a tool for identifying bestsellers, Claire Squires argues that ignoring authors and the socio-material conditions of publishing can re-enforce gatekeeping functions that have led to a homogenous literary culture at the expense of voices from BAME, LGBT+ and other marginalized communities.[21]

Further to the issues with gatekeeping and the taste of acquisitions editors, Archer and Jockers' approach removes the agency of readers. For example, publishers responded to the Donald Trump presidency with books riffing on his mannerisms, political gossip and a range of novelty publications but only a couple of core titles have enjoyed the majority of sales. Michael Wolff's *Fire and Fury* and J. D. Vance's *Hillbilly Elegy* sold more compared to the longer tail.[22] *Fire and Fury* sold 160,000 copies in Great Britain as of March 2018 compared to the accumulative 67,000 sales for the 127 titles featuring the president's name in the title including Trump's *The Art of the Deal*, Watt T. Dickens' *Trump's Christmas Carol*, and M.G. Anthony's *The Trump Book of Insults: An Adult Coloring Book*.[23] A book's theme is insufficient to promote a book, and other factors for success including marketing, word-of-mouth, and cross-media adaptations need to be accounted for. Trump's tweets decrying the publication helped promote Wolff's book. Big data in publishing cannot rest on the laurels of analyzing sales figures, but must instead triangulate various data points to understand what is read and how. A formula built upon both content and context allows insight into the reading process.

JellyBooks offers a blueprint for context-aware data analysis through providing publishers with tracked reading habits for ebooks.[24] The company provide users with Advanced Reader Copies of ebooks with the expectation that readers will opt-in to reporting their consumption back to JellyBooks. Groups of between 200 and 600 participants report metrics including when they open the book and when they abandoned reading.[25] Publishers can then use the data to finalize the marketing budget for a book that most readers abandoned within the first chapters or ask the authors to make adjustments.[26] JellyBooks is expanding beyond trade publishing with academic presses including MIT Press using the service to supplement traditional peer review to reflect the growing necessity for crossover appeal in scholarly communications.[27] JellyBooks' data collection methods are effective since they focus on individual titles for specific targeted information. Bespoke analysis at a smaller, consensual scale provides more useful evidence than the promise of big data analysis tackling a larger body of books.

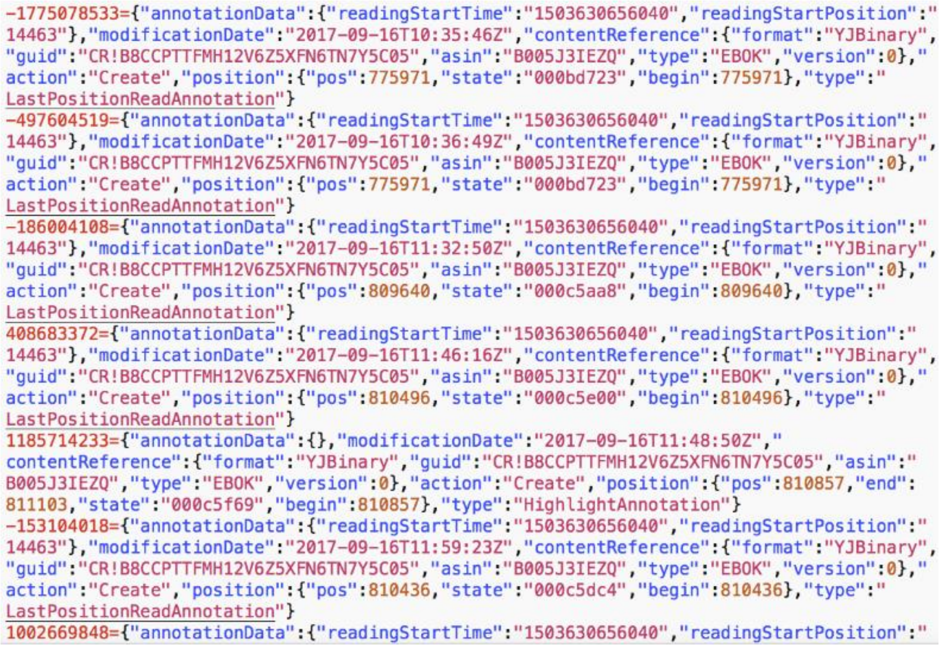
Reading Metrics

The merger of the World Wide Web Consortium (W3C) and International Digital Publishing Forum (IDPF), the authorities maintaining the HTML and EPUB standards respectively, shows how ebook data collection habits mirror those on the web. While the vocabulary for ebook surveillance has yet to develop, the concept of the "cookie" is firmly entrenched in users' understanding of the Web. The Internet Engineering Task Force (IETF) definition of a cookie proposes that "the state management mechanism allows clients and servers that wish to exchange state information to place HTTP requests and responses within a larger context, which we term a 'session.'"[28] The concept of a "state" refers to the user agent, the technical term for browser or computational interface with the World Wide Web, rather than the user. A "state" tracks what information the user agent is providing the server. User agents are the equivalent to ebook's reading system, and both refer to the intermediary between user and service. Since cookies track user agents, there is a gap between user intent and cookie generation. For example, *The Guardian* homepage maintains active connections to various *Guardian* servers as well as Google advertising services, Google Analytics, Facebook, and YouTube. Each of these connections generate cookies and evidence of user agent activity even if the user did not intend to view the advertisements or use Facebook. Ebooks cannot rely on cookies as the format exists within a self-contained closed ecosystem. Conversely, readers often download ebooks from sources such as Amazon and Apple, who store demographic data volunteered by users. Platforms collect ebook metrics through cloud computer services such as Amazon's WhisperSync.

Reading systems fail to capture intent in the data collection process. For example, Amazon monitor timestamps for opening and closing ebooks and infer information about reading through these patterns. Kindle software automatically opens the last book a user read, so accidentally opening the application will register as a reading session. A reader may also idle on a page while multitasking without looking at the page. Standard metrics therefore capture data about the user agent rather than the reader, who cannot be profiled without more invasive data collection. In sum, unless data collected from the user agent can be correlated with some external evidence of reading, the resulting data map software usage rather than reading. Matthew Kirschenbaum warns of the distorting logic of “programmatic computational environments.”[29] This can be seen in date-stamps, which measure time in milliseconds that can be compared with other instances even if this level of accuracy is a feature of Unix time rather than reader intent.

Amazon’s development of reading analytics stems from the WhisperSync cloud backup service as well as the desire from users to access their highlights and annotations on an off-device site. In response to the volume of data, Kindle engineers formed a specific Reading, Mining, and Analytics (RMA) team in 2009 to consider how to best use the data. RMA explored the use of metrics including relatively straightforward elements such as “total access time,” user location, and whether the user was recommended the item. These measures were supplemented by more advanced calculations including “elapsed time since last access” and “data derived from other sensor input” to determine whether an ebook was abandoned.[30] The most complex metric, however, is “access velocity by time and position in content item,” which correlates metrics to create a more complex understanding of a reader’s habits. If a reader slows down their consumption of a text considerably, this likely indicates diminishing interest in the title, although as with other metrics, this ignores any external context.

Before sending data to WhisperSync, Kindle hardware holds a local list of user actions that indicate Amazon’s interests. The database of actions on the local device includes information about the words users have highlighted, which is sent back to Amazon for analysis and press, including a “list of the most looked up words on Kindle books.”[31] **Figure 1** shows details from my “annotation log,” which includes where a reading session finished with a “LastPositionReadAnnotation.” This is contrasted with the “readingStartTime,” which notes when the book was first opened to map reading chronologically. The terminology of “type” offers the greatest insight into Amazon’s approach to reading since both “last position read” and “highlight” are suffixed by “annotation” indicating that these trackers are being broadly referred to as annotations rather than “bookmark” or “location.” “HighlightAnnotation” is also a non-literal descriptor as any word selected is recognized as a highlight rather than those with user intent.

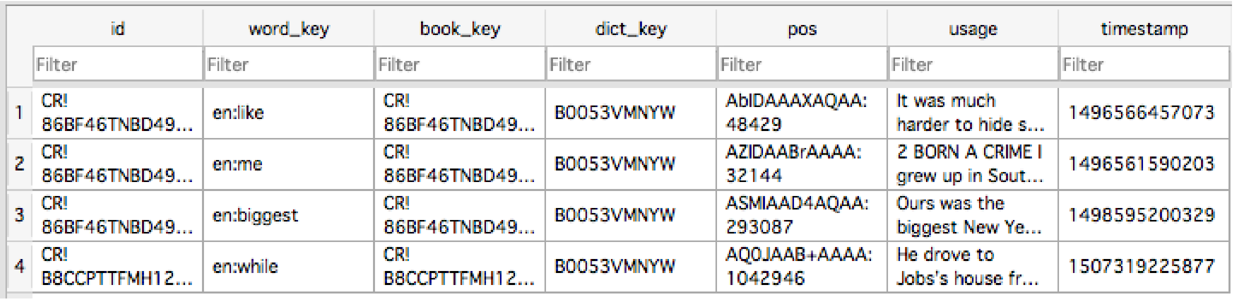


```
-1775078533={"annotationData":{"readingStartTime":"150363065604O","readingStartPosition":"
14463"},"modificationDate":"2017-09-16T10:35:46Z","contentReference":{"format":"YJBinary",
"guid":"CR!B8CCPTTFMH12V6Z5XFN6TN7Y5C05","asin":"B005J3IEZQ","type":"EBOK","version":0},"
action":"Create","position":{"pos":775971,"state":"000bd723","begin":775971},"type":"
LastPositionReadAnnotation"}
-497604519={"annotationData":{"readingStartTime":"1503630656040","readingStartPosition":"
14463"},"modificationDate":"2017-09-16T10:36:49Z","contentReference":{"format":"YJBinary",
"guid":"CR!B8CCPTTFMH12V6Z5XFN6TN7Y5C05","asin":"B005J3IEZQ","type":"EBOK","version":0},"
action":"Create","position":{"pos":775971,"state":"000bd723","begin":775971},"type":"
LastPositionReadAnnotation"}
-186004108={"annotationData":{"readingStartTime":"1503630656040","readingStartPosition":"
14463"},"modificationDate":"2017-09-16T11:32:50Z","contentReference":{"format":"YJBinary",
"guid":"CR!B8CCPTTFMH12V6Z5XFN6TN7Y5C05","asin":"B005J3IEZQ","type":"EBOK","version":0},"
action":"Create","position":{"pos":809640,"state":"000c5aa8","begin":809640},"type":"
LastPositionReadAnnotation"}
408683372={"annotationData":{"readingStartTime":"1503630656040","readingStartPosition":"
14463"},"modificationDate":"2017-09-16T11:46:16Z","contentReference":{"format":"YJBinary",
"guid":"CR!B8CCPTTFMH12V6Z5XFN6TN7Y5C05","asin":"B005J3IEZQ","type":"EBOK","version":0},"
action":"Create","position":{"pos":810496,"state":"000c5e00","begin":810496},"type":"
LastPositionReadAnnotation"}
1185714233={"annotationData":{},"modificationDate":"2017-09-16T11:48:50Z","
contentReference":{"format":"YJBinary","guid":"CR!B8CCPTTFMH12V6Z5XFN6TN7Y5C05","asin":"
B005J3IEZQ","type":"EBOK","version":0},"action":"Create","position":{"pos":810857,"end":
811103,"state":"000c5f69","begin":810857},"type":"HighlightAnnotation"}
-153104018={"annotationData":{"readingStartTime":"1503630656040","readingStartPosition":"
14463"},"modificationDate":"2017-09-16T11:59:23Z","contentReference":{"format":"YJBinary",
"guid":"CR!B8CCPTTFMH12V6Z5XFN6TN7Y5C05","asin":"B005J3IEZQ","type":"EBOK","version":0},"
action":"Create","position":{"pos":810436,"state":"000c5dc4","begin":810436},"type":"
LastPositionReadAnnotation"}
1002669848={"annotationData":{"readingStartTime":"1503630656040","readingStartPosition":"
```

Figure 1: Detail from a user's annotation log from a Kindle 8

The reading metrics available through the "annotation log" only capture one dimension of the user's experience within the reading system. Each Kindle device contains several other databases that monitor usage patterns and report back to Amazon including a database of all dictionary look-ups (**Figure 2**). The vocab.db database records every word highlighted regardless of purpose and stores this in a database linked directly to a dictionary via "dict_key" and a timestamp to track reading progress. The data does not track if the reader is unaware of the word's definition, since a dictionary entry appears by default when a user presses on a word. Each look-up entry is collated with reference to a specific copy of a book, its location within the book, and the length of the word in the dictionary. Amazon uses this data to profile users' reading level, but the metric is prone to capture user error rather than unknown words. For example, **Figure 2** features four accidentally clicked entries of basic English words.

While the locally hosted databases reveal the data collection practices, any relevant information is transmitted via WhisperSync for remote analysis, so there is no on-board source code to reveal how Amazon process the data. The United States Patent and Trademark Office's (USPTO) patent archives provide more comprehensive coverage of Amazon's potential for analyzing reader data. I conducted a search of the USPTO's Patent Full-Text and Image Database (PatFT) in May 2018 for patents with the assignee name of "Amazon Technologies" containing the terms "digital content," Amazon's catch-all name for ebooks, and "read" in any section of the text. PatFT only publishes granted patents and with the extended time for processing patents or any incomplete or withdrawn patents, the search cannot be comprehensive. The search produced 519 results. After filtering out irrelevant applications on topics as diverse as drones, data storage and display technology, there were 123 patents granted relating to the interaction between readers and the Kindle software.



	id	word_key	book_key	dict_key	pos	usage	timestamp
	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	CR! 86BF46TNBD49...	en:like	CR! 86BF46TNBD49...	B0053VMNYW	AbIDAAAXAQAA: 48429	It was much harder to hide s...	1496566457073
2	CR! 86BF46TNBD49...	en:me	CR! 86BF46TNBD49...	B0053VMNYW	AZIDAABrAAAA: 32144	2 BORN A CRIME I grew up in Sout...	1496561590203
3	CR! 86BF46TNBD49...	en:biggest	CR! 86BF46TNBD49...	B0053VMNYW	ASMIAAD4AQAA: 293087	Ours was the biggest New Ye...	1498595200329
4	CR! B8CCPTTFMH12...	en:while	CR! B8CCPTTFMH12...	B0053VMNYW	AQ0JAAB+AAAA: 1042946	He drove to Jobs's house fr...	1507319225877

Figure 2: Detail from the Kindle 8's "vocab.db," the database file that contains definitions users have search for.

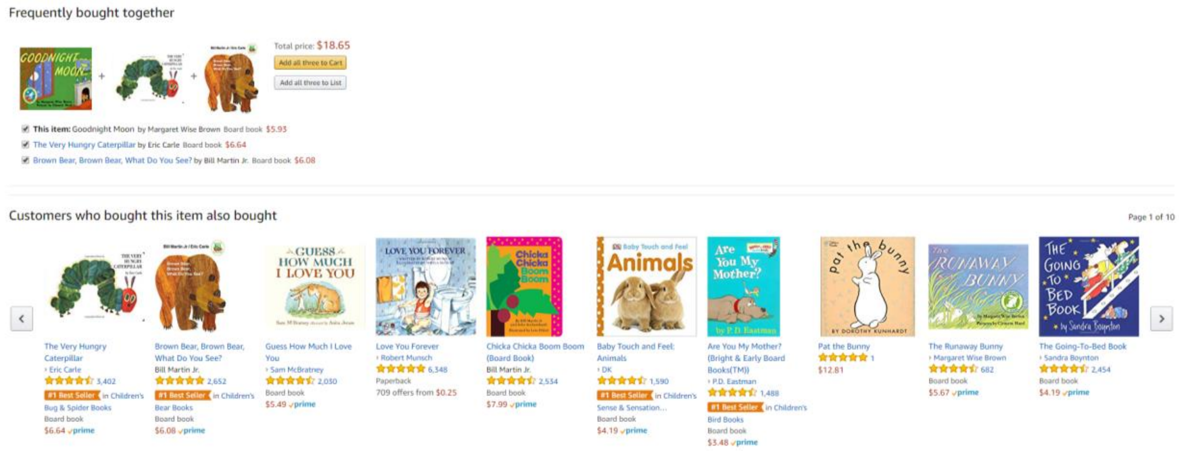
Word	Rank	Frequency	Keyness
highlight	7	2545	12460.743
reading	9	2632	9317.710
supplemental	10	2373	9055.825
mailed	13	2443	7608.088
reader	19	2785	6665.148
annotation	22	1440	5921.087
progress	26	1351	5531.162
pages	30	3925	4982.245
bookmark	37	832	3461.814
chapter	38	790	3298.090
text	39	2762	3159.078
portion	40	3515	3041.611
synchronization	41	960	2911.301
abandonment	47	478	2590.592
location	48	3358	2586.353
club	49	537	2451.809
discussion	51	1017	2421.091
library	54	997	2234.243
portions	58	1451	1858.062
sample	65	777	1722.841

Table 1: Selected book-related terms from a keyword analysis of Amazon's "digital content"/"read" patents against a larger corpus of 4,222 patents filed by Amazon

Table 1 shows the words Amazon associate with reading ebooks. The words are ranked according to keywords, or "a word which appears in a text or corpus statistically significantly more frequently than would be expected by chance when compared to a corpus which is larger or of equal size."[32] For example, in both British and American English, "blimey" is a rare word, but it is more likely to appear in the British English corpus, and therefore would be a keyword in British English. Table 1 highlights the features Amazon emphasize in its ebook patents as compared to the rest of its filings. Several of these terms focus on the technical capabilities of the Kindle software to display portions of a work through the sample system or synchronize data. This analysis highlights Amazon's emphasis on *social reading* through "highlight," "annotation," "club" and "discussion" as well as *retention* in the form of "progress," "abandonment" and "location." The third trend in Amazon's reader-oriented patent filings focuses on *recommendations*, but this does not appear in the preceding analysis since recommendations permeate Amazon's data modelling processes and are not unique to books. There is a strong body of scholarship on Amazon's social reading infrastructure through Popular Highlights and Goodreads, so I will instead focus on recommendations and retention.[33] Through narrowing the focus on patents related to these topics, it is possible to assess how Amazon view the data analysis process.

Recommendations

Amazon's recommendation engine is the most visible instance of the company profiling consumption. Recommendations are adjacent to tracking reading as a proxy for taste: readers who bought the same title are likely to share interests with other books. The company launched its "item-based collaborative filtering" recommendation engine in 1998, which enabled the development of a complex dataset over two decades with caches of browsing and purchasing data from millions of customers.[34] The algorithm sorts recommendations according to how previous customers interacted with the product pages. For example, the Amazon.com product page for Wise Brown and Hurd's *Goodnight Moon* features "Frequently bought together" and "Customers who bought this item also bought" (**Figure 3**), prominently displayed near the top of the page with "Customers who viewed this item also viewed" in the footer, demonstrating the lack of importance of the latter category in Amazon's marketing strategy. There are further filters to avoid unfortunate mixes. In the case of "Frequently bought together," there are parameters in the HTML to check if the items are available from the same seller as Amazon want to bundle the purchases typically from its store. Likewise, recommendations based upon others' longer-term consumption primarily include other books the users may want to purchase, while only occasionally offering suggestions from other parts of Amazon. In the case of *Goodnight Moon*, there are six recommendations without ISBNs, which include plush toys related to *Goodnight Moon*. Unless an item from another product category has a direct connection such as existing in the same franchise, it will not be recommended regardless of the volume of joint purchases. The recommendation system therefore creates a filter bubble, where users are recommended titles with the highest likelihood of joint purchase rather than expanding the user's horizons.



Frequently bought together

Total price: $18.65

☑ **This item:** Goodnight Moon by Margaret Wise Brown Board book $5.93

☑ The Very Hungry Caterpillar by Eric Carle Board book $6.64

☑ Brown Bear, Brown Bear, What Do You See? by Bill Martin Jr. Board book $6.08

Customers who bought this item also bought

Page 1 of 10

The Very Hungry Caterpillar
› Eric Carle
3,402
#1 Best Seller in Children's Bug & Spider Books
Board book
$6.64 prime

Brown Bear, Brown Bear, What Do You See?
Bill Martin Jr.
2,652
#1 Best Seller in Children's Bear Books
Board book
$6.08 prime

Guess How Much I Love You
› Sam McBratney
2,030
Board book
$5.49 prime

Love You Forever
› Robert Munsch
6,348
Paperback
709 offers from $0.25

Chicka Chicka Boom Boom (Board Book)
Bill Martin Jr.
2,534
Board book
$7.99 prime

Baby Touch and Feel: Animals
› DK
1,590
#1 Best Seller in Children's Sense & Sensation…
Board book
$4.19 prime

Are You My Mother? (Bright & Early Board Books(TM))
› P.D. Eastman
1,488
#1 Best Seller in Children's Bird Books
Board book
$3.48 prime

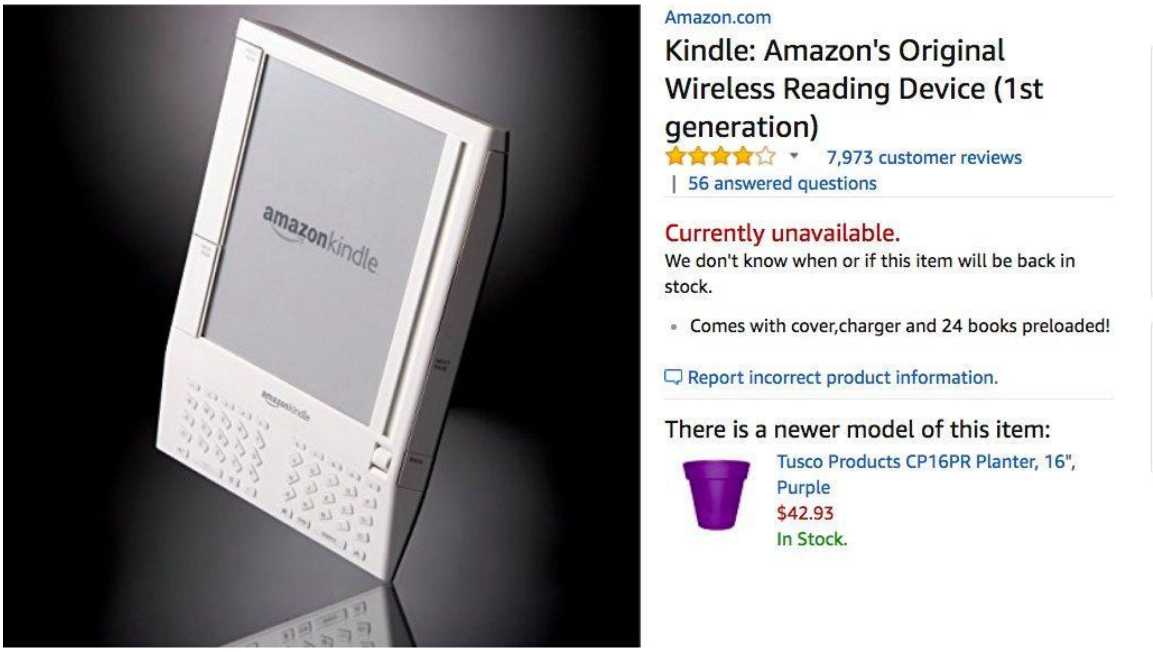
Pat the Bunny
1
$12.81

The Runaway Bunny
› Margaret Wise Brown
682
Board book
$5.67 prime

The Going-To-Bed Book
› Sandra Boynton
2,454
Board book
$4.19 prime

Figure 3: Recommendations on the *Goodnight Moon* product page (13 March 2018; with sponsored recommendations removed)[35]

It is inevitable that transparent, user-generated recommendations are exploited for unscrupulous marketing purposes such as a recommendation to buy a purple planter as a newer model of the first-generation Kindle (**Figure 4**). Out-of-stock items are more susceptible to malicious recommendation attacks, since there are fewer sales. While Amazon face recommendation problems with the breadth of its catalogue and the changing purchasing profile of customers, recommendations within books are easier to account for, since recommendations are largely limited to more books. Big data approaches are often confounded by niche categories. In **Figure 3**, the recommendations are all canonical works of children's literature, where many parents would purchase *Goodnight Moon, The Hungry Little Caterpillar*, and *Brown Bear, Brown Bear What Do You See?* While this leads to a narrow perspective on children's publishing, it is not as arbitrary as the opposite approach, where recommendations are built on limited data. For example, James Secord's *Visions of Science: Books and Readers at the Dawn of the Victorian Age* is matched with Celeste Ng's *Little Fires Everywhere* and Naomi Alderman's *The Power* in "customers who bought this item also bought," despite the former being a book history monograph while the latter two titles are works of literary fiction. The recommendation algorithm made the most statistically significant connections based upon a limited volume of data, leading to profiling a few individuals' tastes rather than a strong link. Automated recommendations have become less prominent with recent changes to the web store as sponsored products appear in search results and on product pages in place of organic recommendations. Amazon engineers are also exploring the introduction of what they term "opaque recommendations" in a further move from the transparent algorithms.[36] Nonetheless, the long-standing openness of recommendations on Amazon offered an unparalleled insight into consumption patterns that was made possible by the scale of the company's holdings and customer base.



Amazon.com

Kindle: Amazon's Original Wireless Reading Device (1st generation)

★★★★☆ ▾ 7,973 customer reviews | 56 answered questions

Currently unavailable.
We don't know when or if this item will be back in stock.

- Comes with cover,charger and 24 books preloaded!

Report incorrect product information.

There is a newer model of this item:

Tusco Products CP16PR Planter, 16", Purple
$42.93
In Stock.

Figure 4: Errant algorithmic recommendations

Retention

Retention and abandonment are important metrics for platforms looking to close the gap between reading and consumption. Kobo's big data white paper and JellyBooks' business model both integrate completion rate into their calculations, and the Kindle Reading, Mining and Analytics team filed a patent focusing on abandonment, a concept that needs to map intent to complement data about when the book was last accessed.[37] Any retention algorithm must account for temporary and permanent abandonment: a reader may set aside a book momentarily with the intention to return to it, the reader might be required to read a few chapters for a class and have completed the necessary sections of the text, readers may finish a book but ignore the extensive backmatter. It is difficult to measure abandonment from the reading system alone unless clear patterns develop.

Francisco Kane Jr, Tom Killalea and Llewyn Mason's "Recommendations based on progress data" outlines a recommendation algorithm that accounts for progress within a book and abandonment, recommending books based on similarities in where reading stalled.[38] It is futile to speculate on what, if any, elements of the patent have been implemented, but the patent indicates which metrics the company felt important enough to invest in protecting. For example, in a patent describing methods of recommendation based upon abandonment, Kane Jr, Killalea, and Mason identify three primary usage categories: not opened for the first time, unfinished, and finished. The core innovation noted in the patent comes from separating items "in progress" and "abandoned."[39] Amazon claim to be able to determine when a reader has stopped reading a book momentarily or permanently. The algorithmic measurement was not publicly implemented, but the same language appeared in self-reporting data on the Kindle Popular Highlights website and more recently via GoodReads, where users can mark a book through “reading,” “hope to read,” “read,” and “stopped reading.”[40]

Book series have been transformed by emergent data practices in the publishing industry and depend on retention to build a loyal audience. A Kobo white paper highlights several methods for correlating the percentage of readers completing a title with sales to assess the engagement of a fandom. The nature of series ensures that publishers should expect the first book in a series to outsell the rest, but once a series’ sales plateau, monitoring drops in engagement can allow publishers and authors to change content to meet the new needs of the audience.[41] The economics of low-cost subscription services such as Kindle Unlimited have encouraged authors to consider the analytics of their book series. Users can subscribe to the service for $9.99 per month to gain access to a catalogue of over one million titles. While the scheme offers the occasional high-profile series such as Susanne Collin’s *Hunger Games* or J.K. Rowling’s *Harry Potter*, most available titles are Kindle Direct Publishing ebooks designed for the scheme. Kindle Unlimited is monetized through the “Kindle Edition Normalized Page Count (KENPC)” that “measure[s] the number of pages customers read in [a] book, starting with the Start Reading Location (SRL) to the end of [the] book” up to a maximum of 3,000 pages.[42] Each author is assigned a number of pages read, and a total kitty, valued at $19.9 million for December 2017 is divided per author, leading to a rate of roughly half a cent per page. The drive towards pay-per-page has led some authors to create lengthy books with tricks to encourage readers to skip to the end, but a more nuanced data-driven approach has also emerged. Authors realized that brand loyalty and fandom could be leveraged and enterprising authors such as Bella Forrest have built business models around Kindle Unlimited. Forrest’s primary series, *A Shade of Vampire*, features a new book each month. Forrest has produced 60 books on a monthly schedule as of June 2018. These titles are divided into “seasons,” with the books in the seventh by 2018. The season introduces new characters and plots according to interest in the books. The data re-enforced environment of Kindle Direct Publishing has become a lucrative career for those willing to play the game. Jeff Bezos’s 2018 letter to shareholders noted that 1,000 authors earned more than $100,000 from Kindle Direct Publishing in 2017.[43]

Amazon’s data analysis of abandonment extends to determining if complexity contributed to the reader’s choice. The complexity of the book could be measured “from a Flesch-Kincaid Readability score or statistics based on statistically improbable phrases [or words that appear only in that particular book and might therefore represent an imaginary language or something far outside normal diction].”[44] Rudolf Flesch and J. Peter Kincaid developed the Grade Level test in the 1970s to judge the complexity of text according to the number of words and the length of syllables and sentences to determine what US school grade level the text would be appropriate for:

$$0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right) - 15.59$$

Amazon's patent filing suggests that the company is confident in the measure's ability to determine reading level when contrasted to demographics. This confidence belies external data that may influence a user's reading ability. For example, readers are unlikely to consume all their books on a Kindle, and may choose to only read fiction with a lower Flesch-Kincaid score than their average reading, triggering a false flag when they abandon something perceived to be more difficult for them. Flesch-Kincaid provides a rough measure of the complexity of text for the author but is less useful for gauging a reader's comprehension level.

Metrics that infer behavior about software usage rather than reader interpretation offer better data sources for Amazon's computational analysis. For example, dedicated Kindle e-readers feature a variety of sensors including GPS and accelerometers (introduced with the launch of the Touch in 2011), which are not exempt from Amazon's reader tracking systems: "accelerometer data may be included to determine when the user was in motion during consumption of content."[45] The sensors are used to track small movements in orientation that map over to various models including reading on a plane, train, or bus. Amazon's rich data collection was matched with an ambitious profiling system to enhance their views and purchases network. The patent outlines the notifications users would receive according to their previous abandonment habits:

> Congratulations on finishing chapter 13 of "Linux Kernel." Similar users jumped ahead to chapter 17 before reading chapter 14… Thank you for your interest in "Derive Your Own Linear Equations. People like you who abandoned "Derive Your Own Integrals" usually abandoned "Derive Your Own Linear Operations."[46]

The patent acknowledges how reading is shaped by context through the example of set reading, where users may jump around a text according to a syllabus. These metrics are imperfect, however, as they cannot consider reader agency and determine if a book has been temporarily or permanently discarded. For example, measures include the removal of a book from a device, which might be the result of storage conditions as much as disliking the specific title. The final example notification flattens abandonment to a metric ignoring the context. The unique habits of individual readers cannot be modelled from the data collected, and any inferences are likely to be generalizations. The nuance of the patent text's contextualization and the suggested metrics reveal tensions in the quantification of reading and its richer context.

Amazon's theoretical use of reading-based data extends towards altering the book's content. In a patent filed in 2010, Daniel Rausch outlines a method for adapting content for users according to their reading level.[47] The algorithm alters sentences to account for the reader's Flesch-Kincaid score. In an example from the patent, the customization runs from "The dog ran" as the most basic statement to "After a ball was thrown, the enormous golden Labrador hurried through the grass to retrieve the ball and stopped abruptly," (**Figure 5**) introducing a range of extra clauses for extra information. Such a project misunderstands the aesthetic value of sentences as the meaning shifts according to each of the various levels with "reading level 1" just including a summary of the fuller narrative developments of "reading level 6." The patent does not outline how Amazon would discern a user's reading level beyond self-identification, but several of the company's data collection processes point towards its ability to guess reading level. For example, the aforementioned vocabulary databases offer insights into the words a user looked up in a dictionary, indicating terms they struggle with. Amazon also collects detailed demographic information including education level in its broader monitoring activity. All these data points can coalesce into the customization of an ebook for the reading level of a text. Amazon did not implement the technology outside of Word Wise, a tool that identifies difficult words and offers brief interlinear definitions.[48] Users only have the option to turn the feature on or off, with no indication of their reading level, which could be deduced from Amazon's Flesch-Kincaid data.

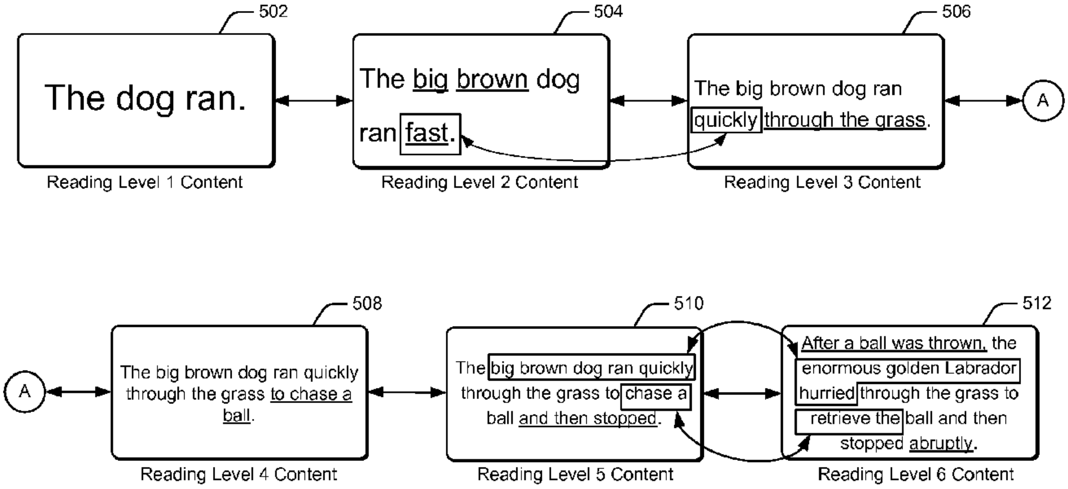

502
The dog ran.
Reading Level 1 Content
504
The big brown dog ran fast.
Reading Level 2 Content
506
The big brown dog ran quickly through the grass.
Reading Level 3 Content
A
A
508
The big brown dog ran quickly through the grass to chase a ball.
Reading Level 4 Content
510
The big brown dog ran quickly through the grass to chase a ball and then stopped.
Reading Level 5 Content
512
After a ball was thrown, the enormous golden Labrador hurried through the grass to retrieve the ball and then stopped abruptly.
Reading Level 6 Content

Figure 5: Detail of Amazon's proposed adjustments for text according to a reader's Flesch-Kincaid score[49]

Conclusion

Amazon's bombast in patent filings does not mirror the eventual implementation of big data approaches to readership. For example, Amazon's recommendation engine is a derivative of twenty-year-old algorithms rather than technology developed during the "big data revolution." The gap between rhetoric and implementation demonstrates a large disconnect between the potential for data collection and how this can be implemented in practice. Partially, the gap is the result of Amazon's voracious appetite for patents in a fiercely competitive marketplace. The company was among the fifteen most prolific patent applicants of 2017 alongside the other large technology companies (excluding Facebook) and hardware manufacturers.[50] In total, Amazon were granted 1,960 patents in 2017. Many of these patents are speculative to claim territory and are unlikely to be implementable beyond the discourse of patents. The noise of patent filings and over-reaching data collection obfuscates Amazon's current data collection and analysis practice, but the patents and source code I analyzed suggest some fundamental flaws in the company's approach to big data analysis of reading.

Start-ups including Colloquiy have struggled to model reader tastes but the more pragmatic approaches of companies such as JellyBooks demonstrates the importance of engaging with users when collecting data to ensure intentionality. When a user elects to send data back to JellyBooks, it is likely to be more meaningful than the over-arching surveillance practices of Amazon and other technology companies. Ebook data collection is limited by observing interactions between the reading system and the cloud storage systems rather than revealing how readers interact with ebooks. We are still in the experimental era of in-app metrics for reading, which still only track software rather than users. This is likely to remain the case while ebook applications are largely distinct from the Web, which has greater strengths in profiling and monitoring via cookies.

The merger of the IDPF and W3C in 2017, and the development of Packaged Web Publications, a format designed to bring ebooks in-line with the open Web, suggests that future developments could lead to more intense monitoring that might increase the efficiency of data analysis. Cookies allow publishers to draw data from a user's web browsing habits in a more comprehensive manner, creating a fuller profile of consumption and reading. App-based publishing offers a further alternative, although one that has diminished in scale in recent years, that also afford publishers access to a greater range of data. The strengthening of data protection legislation through the EU's General Data Protection Regulations (GDPR) may curtail the potential for collecting and analyzing data *en masse*, but a hybrid opt-in approach, expanding upon the JellyBooks' model post-publication could fulfill the promise of a big data approach to publishing.

Biographical note:

Dr Simon Peter Rowberry is Lecturer in Digital Media & Publishing at the University of Stirling. His research on digital reading and ebooks has appeared in *Convergence* and *Language and Literature*. His first book, *Four Shades of Grey: The Kindle* is under contract with MIT Press. Contact: Simon.rowberry@stir.ac.uk.

References:

Allington, Daniel. "'Power to the Reader' or 'Degradation of Literary Taste'? Professional Critics and Amazon Customers as Reviewers of The Inheritance of Loss." *Language and Literature* 25, no. 3 (2016): 254–78.

Alter, Alexandra, and Karl Russell. "Moneyball for Book Publishers: A Detailed Look at How We Read." *The New York Times*, March 14, 2016. http://www.nytimes.com/2016/03/15/business/media/moneyball-for-book-publishers-for-a-detailed-look-at-how-we-read.html.

Amazon.com. “Royalties in Kindle Unlimited and Kindle Owners’ Lending Library.” Amazon Kindle Direct Publishing, 2018. https://kdp.amazon.com/en_US/help/topic/G201541130.

Amazon.com Inc. “Annual Report 2017,” 2018. http://phx.corporate-ir.net/External.File?item=UGFyZW50SUQ9NjkyMDIxfENoaWxkSUQ9NDAyOTkyfFR5cGU9MQ==&t=1.

———. “Goodnight Moon Board Book.” Amazon.com, 2018. https://www.amazon.com/gp/product/0694003611/.

———. “Read More Challenging Books.” Amazon.com, March 17, 2015. http://web.archive.org/web/20150317030835/https://www.amazon.com/gp/feature.html?ie=UTF8&docId=1002989731.

Anthony, M.G. *The Trump Book of Insults: An Adult Coloring Book*. London: Post Hill Press, 2016.

Archer, Jodie, and Matthew L. Jockers. “Archer Jockers.” Archer Jockers, 2018. http://www.archerjockers.com/.

Asay, Clark. “The Informational Value of Patents.” *Berkeley Technology Law Journal* 31, no. 1 (2016): 259–324.

Baker, Paul, Andrew Hardie, and Tony McEnery. *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press, 2006.

Barnett, Tully. “Social Reading: The Kindle’s Social Highlighting Function and Emerging Reading Practices.” *Australian Humanities Review* 56 (May 2014). http://www.australianhumanitiesreview.org/archive/Issue-May-2014/barnett.html#bio.

Baron, Naomi S. *Words Onscreen: The Fate of Reading in a Digital World*. Oxford: Oxford University Press, 2015.

Baym, Nancy K. “Data Not Seen: The Uses and Shortcomings of Social Media Metrics.” *First Monday* 18, no. 10 (September 29, 2013). https://doi.org/10.5210/fm.v18i10.4873.

Bergström, Annika, and Lars Höglund. “E-Books: In the Shadow of Print.” *Convergence* OnlineFirst (2018). https://doi.org/10.1177/1354856518808936.

Blatt, Ben. *Nabokov’s Favourite Word Is Mauve: The Literary Quirks and Oddities of Our Most-Loved Authors*. London: Simon & Schuster, 2017.

boyd, danah, and Kate Crawford. “Critical Questions for Big Data.” *Information, Communication & Society* 15, no. 5 (2012): 662–79.

Cameron, Lauren. “Marginalia and Community in the Age of the Kindle: Popular Highlights in The Adventures of Sherlock Holmes.” *Victorian Review* 38, no. 2 (2012): 81–99.

Caramanica, Jon. “The Rowdy World of Rap’s New Underground.” *The New York Times*, June 22, 2017, sec. Arts. https://www.nytimes.com/2017/06/22/arts/music/soundcloud-rap-lil-pump-smokepurrp-xxxtentacion.html.

Cohn, Cindy, and Parker Higgins. “Who’s Tracking Your Reading Habits? An E-Book Buyer’s Guide to Privacy, 2012 Edition.” Electronic Frontier Foundation, November 29, 2012. https://www.eff.org/deeplinks/2012/11/e-reader-privacy-chart-2012-update.

Colbjørnsen, Terje. “The Construction of a Bestseller: Theoretical and Empirical Approaches to the Case of the Fifty Shades Trilogy as an EBook Bestseller.” *Media, Culture & Society* 36, no. 8 (2014): 1100–1117.

Davis, Mark. “E-Books in the Global Information Economy.” *European Journal of Cultural Studies* 18, no. 4–5 (2015): 514–29.

Dickens, Watts T. *Trump’s Christmas Carol*. London: Ebury Press, 2017.

Dilworth, Dianna. “The Most Looked Up Words on Kindle,” 2015. http://adweek.it/1GfOWaa.

Driscoll, Beth, and DeNel Rehberg Sedo. “Faraway, So Close: Seeing the Intimacy in Goodreads Reviews.” *Qualitative Inquiry* Online First (September 26, 2018). https://doi.org/10.1177/1077800418801375.

Eriksson, Maria, Rasmus Fleischer, Anna Johansson, Pelle Snickars, and Patrick Vonderau. *Spotify Teardown: Inside the Black Box of Streaming Music*. Cambridge: MIT Press, 2019.

Finn, Ed. “New Literary Cultures: Mapping the Digital Networks of Toni Morrison.” In *From Codex to Hypertext: Reading at the Turn of the Twenty-First Century*, edited by Anouk Lang, 177–202. Amherst and Boston: University of Massachusetts Press, 2012.

Hallinan, Blake, and Ted Striphas. “Recommended For You: The Netflix Prize and the Production of Algorithmic Culture.” *New Media & Society* 18, no. 1 (2016): 117–37.

Hayler, Matt. *Challenging the Phenomena of Technology: Embodiment, Expertise, and Evolved Knowledge*. Basingstoke: Palgrave Macmillan, 2015.

Iliadis, Andrew, and Federica Russo. “Critical Data Studies: An Introduction.” *Big Data & Society* 3, no. 2 (2016). https://doi.org/10.1177/2053951716674238.

IPO. “Top 300 Organizations Granted U.S. Patents in 2017.” IPO, 2018.

Jacobs, Adam. “The Pathologies of Big Data.” *ACM Queue*, 2009. http://queue.acm.org/detail.cfm?id=1563874.

Jellybooks. “Test Reading Campaign Organised by Jellybooks.” Jellybooks, 2018. https://www.jellybooks.com/campaigns/d46de142ded8f6bf7394d155927d2221.

Kane Jr, Francisco J, Tom Killalea, and Llewyn Mason. Recommendations based on progress data. United States Patent no. 9,153,141, filed June 30, 2009, and issued October 6, 2015.

Kirschenbaum, Matthew G. *Mechanisms: New Media and the Forensic Imagination*. Cambridge: MIT Press, 2012.

Kobo. “Publishing in the Era of Big Data,” 2014. http://cafe.kobo.com/_ir/159/20149/Publishing%20in%20the%20Era%20of%20Big%20Data%20-%20Kobo%20Whitepaper%20Fall%202014.pdf.

Kristol, David M., and Lou Montulli. “RFC2109: HTTP State Management Mechanism.” Network Working Group Request for Comments, 1997. https://tools.ietf.org/html/rfc2109.

Madrigal, Alexis C. “How Netflix Reverse Engineered Hollywood.” *The Atlantic*, January 2, 2014. https://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/.

Marino, Mark C. “Reading *Exquisite_code*: Critical Code Studies of Literature.” In *Comparative Textual Media: Transforming the Humanities in the Postprint Era*, edited by N. Katherine Hayles and Jessica Pressman, 283–309. Minneapolis and London: University of Minnesota Press, 2013.

Martens, Marianne. *Publishers, Readers, and Digital Engagement: Participatory Forums and Young Adult Publishing*. London: Palgrave, 2016.

Matthews, Jolie C. “Professionals and Nonprofessionals on Goodreads: Behavior Standards for Authors, Reviewers, and Readers.” *New Media & Society* 18, no. 10 (2016): 2305–22.

Morris, Jeremy Wade. “Curation by Code: Infomediaries and the Data Mining of Taste.” *European Journal of Cultural Studies* 18, no. 4–5 (2015): 446–63.

Murray, Simone. *The Digital Literary Sphere*. Baltimore: Johns Hopkins University Press, 2018.

Nakamura, Lisa. "'Words with Friends': Socially Networked Reading on Goodreads." *PMLA* 128, no. 1 (2013): 238–43.

Patankar, Rashmi Arun, Jeffrey Matthew Bilger, and Colin Ian Bodell. Providing opaque recommendations. United States Patent No. 10,049,397, filed March 6, 2013, and issued August 14, 2018.

Phillips, Angus. "Have We Passed Peak Book? The Uncoupling of Book Sales from Economic Growth." *Publishing Research Quarterly* 33, no. 3 (2017): 310–27.

Pink, Sarah, Minna Ruckenstein, Robert Willim, and Melisa Duque. "Broken Data: Conceptualising Data in an Emerging World." *Big Data & Society* 5, no. 1 (June 1, 2018). https://doi.org/10.1177/2053951717753228.

Pouly, Marie-Pierre. "Playing Both Sides of the Field: The Anatomy of a 'Quality' Bestseller." *Poetics* 59 (2016): 20–34.

Rausch, Daniel B. Determining Reading Levels of Electronic Books. United States Patent no. 8,744,855, filed August 9, 2010, and issued June 3, 2014.

Rhomberg, Andrew. "Reading Fast and Slow." *Andrew Rhomberg* (blog), April 8, 2018. https://medium.com/@arhomberg/reading-fast-and-slow-d6d329ad7715.

Smith, B., and G. Linden. "Two Decades of Recommender Systems at Amazon.Com." *IEEE Internet Computing* 21, no. 3 (June 2017): 12–18.

Squires, Claire. "Taste and/or Big Data?: Post-Digital Editorial Selection." *Critical Quarterly* 59, no. 3 (2017): 24–38.

Thomas, Bronwen. "Trickster Authors and Tricky Readers on the MZD Forums." In *Mark Z. Danielewski*, edited by Joe Bray and Alison Gibbons, 86–102. Manchester: Manchester University Press, 2011.

Trump, Donald J. *The Art of the Deal*. London: Random House, 2016.

Vance, J.D. *Hillbilly Elegy: A Memoir of a Family and Culture in Crisis*. London: William Collins, 2016.

Werner Vogels. "Werner." Kindle, 2017. https://kindle.amazon.com/profile/Werner/160.

Wolff, Michael. *Fire and Fury: Inside the Trump White House*. London: Little, Brown, 2018.

Notes:

[1] Mark Davis, "E-Books in the Global Information Economy," *European Journal of Cultural Studies* 18, no. 4–5 (2015): 527.

[2] A point of contention at the centre of the print vs. digital reading debate. Hayler offers a thoughtful analysis of the phenomenology of reading on screen in Matt Hayler, *Challenging the Phenomena of Technology: Embodiment, Expertise, and Evolved Knowledge* (Basingstoke: Palgrave Macmillan, 2015).

[3] Terje Colbjørnsen, "The Construction of a Bestseller: Theoretical and Empirical Approaches to the Case of the Fifty Shades Trilogy as an EBook Bestseller," *Media, Culture & Society* 36, no. 8 (2014): 1100–1117; Marie-Pierre Pouly, "Playing Both Sides of the Field: The Anatomy of a 'Quality' Bestseller," *Poetics* 59 (2016): 20–34.

[4] Clark Asay, "The Informational Value of Patents," *Berkeley Technology Law Journal* 31, no. 1 (2016): 259–324.

[5] Mark C. Marino, "Reading *Exquisite_code*: Critical Code Studies of Literature," in *Comparative Textual Media: Transforming the Humanities in the Postprint Era*, ed. N. Katherine Hayles and Jessica Pressman (Minneapolis and London: University of Minnesota Press, 2013), 283.

[6] Adam Jacobs, "The Pathologies of Big Data," *ACM Queue*, 2009, 4, http://queue.acm.org/detail.cfm?id=1563874.

[7] See Angus Phillips, "Have We Passed Peak Book? The Uncoupling of Book Sales from Economic Growth," *Publishing Research Quarterly* 33, no. 3 (2017): 310–27 for an example of triangulating Gross Domestic Product with the value of book sales to argue publishing is under-performing.

[8] Jacobs, "The Pathologies of Big Data," 8.

[9] Andrew Iliadis and Federica Russo, "Critical Data Studies: An Introduction," *Big Data & Society* 3, no. 2 (2016), https://doi.org/10.1177/2053951716674238; danah boyd and Kate Crawford, "Critical Questions for Big Data," *Information, Communication & Society* 15, no. 5 (2012): 662–79.

[10] Jeremy Wade Morris, "Curation by Code: Infomediaries and the Data Mining of Taste," *European Journal of Cultural Studies* 18, no. 4–5 (2015): 447.

[11] Cindy Cohn and Parker Higgins, "Who's Tracking Your Reading Habits? An E-Book Buyer's Guide to Privacy, 2012 Edition," Electronic Frontier Foundation, November 29, 2012, https://www.eff.org/deeplinks/2012/11/e-reader-privacy-chart-2012-update.

[12] Sarah Pink et al., "Broken Data: Conceptualising Data in an Emerging World," *Big Data & Society* 5, no. 1 (June 1, 2018): 1, https://doi.org/10.1177/2053951717753228.

[13] Davis, "E-Books in the Global Information Economy," 515.

[14] Blake Hallinan and Ted Striphas, "Recommended For You: The Netflix Prize and the Production of Algorithmic Culture," *New Media & Society* 18, no. 1 (2016): 117–37.

[15] Nancy K. Baym, "Data Not Seen: The Uses and Shortcomings of Social Media Metrics," *First Monday* 18, no. 10 (September 29, 2013), https://doi.org/10.5210/fm.v18i10.4873.

[16] Jon Caramanica, "The Rowdy World of Rap's New Underground," *The New York Times*, June 22, 2017, sec. Arts, https://www.nytimes.com/2017/06/22/arts/music/soundcloud-rap-lil-pump-smokepurrp-xxxtentacion.html.

[17] Maria Eriksson et al., *Spotify Teardown: Inside the Black Box of Streaming Music* (Cambridge: MIT Press, 2019); Alexis C. Madrigal, "How Netflix Reverse Engineered Hollywood," The Atlantic, January 2, 2014, https://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/.

[18] Simone Murray, *The Digital Literary Sphere* (Baltimore: Johns Hopkins University Press, 2018); Beth Driscoll and DeNel Rehberg Sedo, "Faraway, So Close: Seeing the Intimacy in Goodreads Reviews," *Qualitative Inquiry* Online First (September 26, 2018), https://doi.org/10.1177/1077800418801375; Bronwen Thomas, "Trickster Authors and Tricky Readers on the MZD Forums," in *Mark Z. Danielewski*, ed. Joe Bray and Alison Gibbons (Manchester: Manchester University Press, 2011), 86–102; Daniel Allington, "'Power to the Reader' or 'Degradation of Literary Taste'? Professional Critics and Amazon Customers as Reviewers of The Inheritance of Loss," *Language and Literature* 25, no. 3 (2016): 254–78; Marianne Martens, *Publishers, Readers, and Digital Engagement: Participatory Forums and Young Adult Publishing* (London: Palgrave, 2016); Ed Finn, "New Literary Cultures: Mapping the Digital Networks of Toni Morrison," in *From Codex to Hypertext: Reading at the Turn of the Twenty-First Century*, ed. Anouk Lang (Amherst and Boston: University of Massachusetts Press, 2012), 177–202.

[19] See, for example, Naomi S. Baron, *Words Onscreen: The Fate of Reading in a Digital World* (Oxford: Oxford University Press, 2015); Annika Bergström and Lars Höglund, "E-Books: In the Shadow of Print," *Convergence* OnlineFirst (2018), https://doi.org/10.1177/1354856518808936.

[20] Jodie Archer and Matthew L. Jockers, "Archer Jockers," Archer Jockers, 2018, http://www.archerjockers.com/.

[21] Claire Squires, "Taste and/or Big Data?: Post-Digital Editorial Selection," *Critical Quarterly* 59, no. 3 (2017): 24–38.

[22] Michael Wolff, *Fire and Fury: Inside the Trump White House* (London: Little, Brown, 2018); J.D. Vance, *Hillbilly Elegy: A Memoir of a Family and Culture in Crisis* (London: William Collins, 2016).

[23] All sales figures taken from Nielsen BookScan on 12 March 2018 for period between 1 January 2016 and 1 March 2018. Donald J. Trump, *The Art of the Deal* (London: Random House, 2016); Watts T. Dickens, *Trump's Christmas Carol* (London: Ebury Press, 2017); M.G. Anthony, *The Trump Book of Insults: An Adult Coloring Book* (London: Post Hill Press, 2016).

[24] Alexandra Alter and Karl Russell, "Moneyball for Book Publishers: A Detailed Look at How We Read," *The New York Times*, March 14, 2016, http://www.nytimes.com/2016/03/15/business/media/moneyball-for-book-publishers-for-a-detailed-look-at-how-we-read.html.

[25] Andrew Rhomberg, "Reading Fast and Slow," *Andrew Rhomberg* (blog), April 8, 2018, https://medium.com/@arhomberg/reading-fast-and-slow-d6d329ad7715.

[26] Alter and Russell, "Moneyball for Book Publishers."

[27] Jellybooks, "Test Reading Campaign Organised by Jellybooks," Jellybooks, 2018, https://www.jellybooks.com/campaigns/d46de142ded8f6bf7394d155927d2221.

[28] David M. Kristol and Lou Montulli, "RFC2109: HTTP State Management Mechanism," Network Working Group Request for Comments, 1997, https://tools.ietf.org/html/rfc2109.

[29] Matthew G. Kirschenbaum, *Mechanisms: New Media and the Forensic Imagination* (Cambridge: MIT Press, 2012), 133.

[30] Francisco J Kane Jr, Tom Killalea, and Llewyn Mason, Recommendations based on progress data, United States Patent no. 9,153,141, filed June 30, 2009, and issued October 6, 2015, fig. 5.

[31] Dianna Dilworth, "The Most Looked Up Words on Kindle," 2015, http://adweek.it/1GfOWaa.

[32] Paul Baker, Andrew Hardie, and Tony McEnery, *A Glossary of Corpus Linguistics* (Edinburgh: Edinburgh University Press, 2006), 97.

[33] See, for example, Tully Barnett, "Social Reading: The Kindle's Social Highlighting Function and Emerging Reading Practices," *Australian Humanities Review* 56 (May 2014), http://www.australianhumanitiesreview.org/archive/Issue-May-2014/barnett.html#bio; Lauren Cameron, "Marginalia and Community in the Age of the Kindle: Popular Highlights in The Adventures of Sherlock Holmes," *Victorian Review* 38, no. 2 (2012): 81–99; Jolie C. Matthews, "Professionals and Nonprofessionals on Goodreads: Behavior Standards for Authors, Reviewers, and Readers," *New Media & Society* 18, no. 10 (2016): 2305–22; Lisa Nakamura, "'Words with Friends': Socially Networked Reading on Goodreads," *PMLA* 128, no. 1 (2013): 238–43.

[34] B. Smith and G. Linden, "Two Decades of Recommender Systems at Amazon.Com," *IEEE Internet Computing* 21, no. 3 (June 2017): 12–18.

[35] Amazon.com Inc, "Goodnight Moon Board Book," Amazon.com, 2018, https://www.amazon.com/gp/product/0694003611/.

[36] Rashmi Arun Patankar, Jeffrey Matthew Bilger, and Colin Ian Bodell, Providing opaque recommendations, United States Patent No. 10,049,397, filed March 6, 2013, and issued August 14, 2018.

[37] Kane Jr, Killalea, and Mason, Recommendations based on progress data.

[38] Kane Jr, Killalea, and Mason.

[39] Kane Jr, Killalea, and Mason, fig. 1.

[40] For example, Werner Vogels, "Werner," Kindle, 2017, https://kindle.amazon.com/profile/Werner/160.

[41] Kobo, "Publishing in the Era of Big Data," 2014, http://cafe.kobo.com/_ir/159/20149/Publishing%20in%20the%20Era%20of%20Big%20Data%20-%20Kobo%20Whitepaper%20Fall%202014.pdf.

[42] Amazon.com, "Royalties in Kindle Unlimited and Kindle Owners' Lending Library," Amazon Kindle Direct Publishing, 2018, https://kdp.amazon.com/en_US/help/topic/G201541130.

[43] Amazon.com Inc, "Annual Report 2017," 2018, [viii], http://phx.corporate-ir.net/External.File?item=UGFyZW50SUQ9NjkyMDIxfENoaWxkSUQ9NDAyOTkyfFR5cGU9MQ==&t=1.

[44] Kane Jr, Killalea, and Mason, Recommendations based on progress data, col. 9.

[45] Kane Jr, Killalea, and Mason, Recommendations based on progress data, col. 7.

[46] Kane Jr, Killalea, and Mason, fig 19.

[47] Daniel B. Rausch, Determining Reading Levels of Electronic Books, United States Patent no. 8,744,855, filed August 9, 2010, and issued June 3, 2014.

[48] Amazon.com Inc, "Read More Challenging Books," Amazon.com, March 17, 2015, http://web.archive.org/web/20150317030835/https://www.amazon.com/gp/feature.html?ie=UTF8&docId=1002989731.

[49] Rausch, Determining Reading Levels of Electronic Books.

[50] IPO, "Top 300 Organizations Granted U.S. Patents in 2017," IPO, 2018.