**[Slide 1: Title]**

In May of 1905, a review of *The Troll Garden*, Willa Cather’s first published collection of short stories, appeared in *The* *New York Times*. The title of the review is “Promising Stories” and the length is three paragraphs, or what seems to be a rather short notice when we compare it to reviews of *The Troll Garden* from *Harper’s Weekly*, *Bookman,* or *Atlantic Monthly*, all of which are included in *Willa Cather*: *The Contemporary Revi*ews.[[1]](#footnote-1)

**[Slide 2: Troll Garden]**

But what if we compare this review of *The Troll Garden* to other reviews in *The New York Times*, or other reviews in *The New York Times* in 1905, or other reviews in *The New York Times* in May of 1905 to characterize its length? *The New York Times* archive API makes these comparisons relatively easy, as one can quite easily construct a Python script to retrieve review metadata, including review word counts, by month and year. And here is what we find:

**[Slide 3: NYT-1]**

These word lengths do not have a “normal distribution” or the iconic bell curve we might remember from a statistics class. Instead, they range from 29 to 4,492 words long. A review cannot be shorter than 0 words, the median review length is 287 words, and the largest review length is much larger than most of the reviews. We therefore see what is called ‘a distribution converging to log-normal.” Using a log function, we can convert the values to a normal distribution, and then analyze each observation’s relationship to the norms of the data.

**[Slide 4: NYT-2]**

A quick way to compare a value to a group of values is to calculate its z-score. You may remember from statistics class that an observation’s raw difference from the mean can be deceptive. Being an inch taller than the average could be a little, or it could be a lot, depending on how many observations have been made, and how varied the recorded heights are. A z-score measures an observation’s difference from the mean scaled to the standard deviation. For example, the review of Cather’s collection has a z-score of -.3059, which is to say that it is less than one third of one standard deviation away from the norm. A statistically accurate description would characterize the review as being within the typical length range for *The New York Times* in April 1905.

Perhaps someone in the audience here today is already wondering what the point this introduction has been. It’s not an unfair question. With this anecdote, I wanted to raise two related topics that I will come back to as I speak. The first is what tends to happen when we make heuristic processes to evaluate ostensibly straightforward things like size, frequency, density, or duration instead of making measurements. Second, measuring any number of things is easier when we (1) have data that represents something measurable and (2) use a computer to count something what would otherwise take a very long time to count. These two ideas are always somewhere in the background or foreground when we talk about digital humanities because so much of our work brings humanities interpretation into contact computational tools. As Andrew Piper argues in the inaugural issue of *Cultural Analytics*, digital humanities must not merely be “computer science applied to culture” but instead “a wholesale rethinking of *both* of these categories.” In other words, DH can call our attention to limits of hermeneutics and the limits of quantification, and it can have a generative effect on both kinds of inquiry.

In this paper, I focus on book reviews of the turn-of-the century United States as crucial, currently undervalued book historical objects. I frame my analysis with several important stages of the data lifecycle—study design, planning, and analysis; data creation and collection; data processing; data preservation, distribution, and discovery, and data re-use. This model, I will argue, has enormous potential to guide the creation of large-scale, publicly available book review datasets, and to underscore longstanding compatibilities between book history, and large scale humanities computing. My presentation is deeply tied to a set of principles for conscientious work in this vein that I have been developing for myself, and it is also indicative of a direction I would like to see the book history discipline continue to take.

**The Data Lifecycle**

**[Slide 5: lifecycle-all]**

What is the data lifecycle? This interpretive concept or model has become something of a librarian’s slideshow cliché. Lifecycle diagrams are liable to show up as flow charts; brightly colored circles with arrows; A mostly blue geometric diamond shape with one green triangle that says “Dependent Processes”; swirly thingies making up an outer circle that seems to be swirling around an inner circle that’s swirling in the opposite direction; an infographic with the universal icons for a server, a computer, downloading from the cloud and … I think that’s a Word War II fighter plane? And many more. Just past this preponderance of web chatter, however, is potentially crucial perspective on how research data move through an ecosystem of interlinked practitioners. For this reason, I want to emphasize the model as described by the Data Documentation Initiative (DDI).

**[Slide 6: DDI Lifecycle] -** Ball, A., 2012. Review of Data Management Lifecycle Models. University of Bath.

The DDI Lifecycle version 3.0 emphasizes the study concept, data collection, data processing, data archiving and distribution, data discovery, and data analysis. As with most models, repurposing is considered the end of one iteration of the lifecycle and the beginning of new one.

**Phases I and VII: Study Concept and Data Analysis**

I have paired the study concept stage (sometimes called the planning stage) of the lifecycle with the data analysis stage because so much of data creation, conversion, and modeling pertains to the kind of queries one would like to execute. Generally, datasets are generated by a specific study, or as a purchasable commodity that anticipates a known demand associated with an area of study or a set of questions. Most among us, I suspect, would prefer to think first about our research questions and then about the data strategies and methods that would fit those questions. In a “humanities computing” context, however, we must pair this kind of discussion with a general understanding of the kind of data we would need to use certain methods or answer certain questions.

For many, reviews are everyday documents so familiar as to seem self-evident, pure expressions of recommendation, condemnation, or something between the two. The review is too seldom treated as a highly rehearsed and performative genre with deep rhetorical and material codes of information transfer. Book historians and book historically sensitive literature scholars, meanwhile, have worked extensively with book reviews and have done much to dispel the idea that the content of periodicals can be taken as neutral information. Gerard Genette theorizes “the review” as a paratextual or intertextual document; Pierre Bourdieu dwells extensively on the prestige function of the critic. Nina Baym’s *Novels, Readers, and Reviewers: Responses to Fiction in Antebellum America* (1984) consults “more than two thousand novel reviews” that make “some attempt at description and evaluation” (14). Joan Shelley Rubin’s *The Making of Middlebrow Culture* (2000) engages forcefully with how reviewers participating in construction the idea of middlebrow culture. Numerous book historians have looked to book reviews for signs of how an author, text, or genre was received. More recently, Ashley Champagne has turned to Goodreads reviews for evidence of how readers participate in literary canon making, and Allison Hegel has done computational analysis on a range of contemporary review types in search of patterns that characterize how genre informs reader expectations and judgments of quality. For the moment, I’m lumping so-called professional reviews with their nonprofessional counterparts, but many of the people I have just named parse these categories—and the liminal states between them—in important ways.

**[Slide 7: *web accessible* and computable are not the same thing]**

To get to a point where pursuing these questions computationally is possible, we must understand that *web accessible* and computable are not the same thing. The preponderance of digital page images and OCR have given us the false impression that everything is available and at our fingertips, but this is far from true. Even an excellent resource like the reader experience database can be browsed and searched much more easily than it could be computed over. Champagne and Hegel have turned to Goodreads precisely because its content can be webscraped. If we have each review as its own block of machine readable text and some basic metadata about each review, we can already employ some of the most well established distant reading techniques associated with digital humanities, such as sentiment analysis, topic modeling, term collocations, part of speech tagging, and named entity recognition (e.g., computationally recognizing references to people and places). For many of these methods, we only a spreadsheet of all terms and term counts in a document. However, most reviews are not yet in this state.

**[Slide 8: TF Table]**

Certain queries on literature, such as an analysis of various characters’ patterns of speech in a novel, require extensive hand encoding. Analogously, for reviews, there are many queries that require the book that is being reviewed to be tagged. There are natural language processing or computer vision methods to try and determine which book a book review is reviewing, but a book historical standard would at minimum demand hand correction to avoid lumping different texts with matching titles, or to group different titles with a common underlying text.

Many bibliographers or book historians would also stress the importance of being able to facet by edition, issue, state, or impression, for a review is often specific to a one such manifestation.

**[Slide 9: Amazon]**

Take, for example, an Amazon.com product review warning that an on-demand edition of *The King in Yellow* has “the font and layout of a term paper,” or a Goodreads review—

**[Slide 10: Goodreads]**

—which comments extensively on the consequences of choosing the first edition Doubleday text over the revised and expanded Pennsylvania edition text when preparing the Blackstone Audiobook version of *Sister Carrie*. Such faceting might not yield major results on a large-scale level, as we may discover that most reviewers believe themselves to reviewing and responding to some abstracted or Platonic version of whatever text they are reading, but we might at least earn something about when this widely held myth of a disembodied text is most likely to break down in the face of materialization.

**[Slide 11: Bookman Brevities]**

Going even further, many print reviews discuss numerous books in one monthly or weekly column, so we want, at minimum, to isolate multi-book reviews into their own category or better still, tag every block of text to indicate the book being discussed. This is not a straightforward matter at all when we consider sentences that compare two books or speak generally about all books or generalize about the author’s tendencies.

**[Slide 12: recent women short story]**

I would be quite excited to explore further how these kinds of review subgenres affect the rhetoric of reviews, the forcefulness of review judgments, or the thoroughness of reviewers’ examples to support their claims. Overall, thinking about how reviews are likely to be studied should help us determine what kind of structure book review data must have to be desirable for reuse or later expansion.

**Phase II: Data Collection**

To consider acquisition or data creation (or conversion of primary resources to data fields) sufficiently, we must understand the range of norms for book reviews in the context of an already articulated research question (or a survey of common research questions if the data are to be produced purely for public access). For two years, I have worked a team of digital humanities colleagues at the University of Pittsburgh on a project titled “Computational Approaches to Textual Networks.”[[2]](#footnote-2) Our primary goal has been to create workflow models to address challenges of producing machine-readable textual corpora with strong relational attributes.[[3]](#footnote-3) Our efforts have included surveying the availability of digitized book reviews, experimenting with workflows to adapt or “up-coding” existing digital assets, and conducting discrete, small-scale digitization of materials.[[4]](#footnote-4) Our survey of the current landscape of available data suggests a tremendous potential to move from state where book reviews are highly discoverable and searchable, to a state where book reviews are fully computable. To summarize:

* Sites like HathiTrust and Internet Archive have many searchable volumes of periodicals, but these are not structured to partition individual articles into separate units.
* Chronicling America has lots of dirty OCR page scans containing book reviews, but they are mixed in with all other page content

**[Slide 13: SF]**

**[Slide 14: SF-ocr]**

* *The New York Times* has an API where you can retrieve metadata (review content is labeled as such) and pdf files, but not underlying full text. Other newspapers have similar search-ability without computability.
* Proquest and others also have a ‘review’ datatype. You can search and download metadata in small chunks but will get banned for taking too much too fast.
* I have seen historical reviews modeled or entity referenced to the book or books they review, but this is a logical next step, so others might be doing it. (Goodreads and Amazon have their own ways of doing this.)

This brief overview of the field provides a partial sense of what you are likely to find if you go looking for book reviews to compute with. You find a field with browsing access to much more than you could ever read, and computational access to very little. If you begin the long process of making your corpus of book reviews, you will probably find the labor dull and progress slow. My team has brought in graduate and undergraduate student assistants with mixed results. We’ve reached out Proquest and found that *American Periodicals Series* data is allegedly for sale but can’t be furnished without a 9- to 18-month turnaround window. As with all things digital, copyright is in part a genuine concern and more often a convenient excuse for inaction. If an individual scholar generates data in conjunction with an article, that scholar has every short-term incentive to guard their data as they might guard a good idea. Yet I can’t help but feel that, as a result of this mix of obstacles and ambiguities and perfectly reasonable concerns, we are all missing out on something worth doing.

**Phases III and IV: Processing and Preservation**

The processing and preservation stages of the lifecycle focus on “actions or steps performed on data.” Rigorous documentation “to ensure the utility and integrity of the data,” as well as reproducibility, is strongly recommended.[[5]](#footnote-5) Common processing practices include anonymization for things like medical data, merging two categories into one, tagging data with attributes, normalizing spelling, etc. For book reviews, hand encoding or correcting of metadata seems like most important thing to document. If a data creator is going to use words like edition or format, it would be important to know if that person has the bibliographical proficiency to use those terms accurately. On the computational side, it is typical to do things like remove all punctuation from a text before generating a term frequency list, so this kind of practice should be very well documented in anticipation of a less computationally inclined audience.[[6]](#footnote-6)

**[Slide 15: Document]**

Under ideal circumstances, a “readme” or similar file should exist at a predictable endpoint so that users otherwise unfamiliar with the data can learn more about it. The United States geological survey data producers to assume that “transfer may occur via automated or non-automated mechanisms” (USGS). In other words, datasets should be archived in a way that supports a human clicking a link to download a set of files, as well as in a way that would allow a Python script to download all data to a subdirectory.

The preservation stage typically begins with an understanding of when to preserve versus destroy data. In a book history context, we are unlikely to destroy data but this consideration is required in contexts where privacy is a chief concern. Here one typically considers the necessary context for outside audience, how to best document practices, and how to articulate absence and uncertainties. Perhaps most importantly, data makers must contextualize their data for humanities audiences that is very familiar with well-known critiques of how other disciplines speak about or use data.

Perhaps obviously, I do not think we should eschew the term *data* or rename it something less threatening to people who, in Michael Suarez’s terms, may have found “algebra uncongenial and calculus unkind.”[[7]](#footnote-7) I agree that skeptics of scientism are right to be skeptical, and I think scholars like Lisa Gitelman and Virginia Jackson have raised important points about how positivism mischaracterizes the primacy and autonomy of data.[[8]](#footnote-8) Simultaneously, even the most theory-heavy humanistic scholarship is interspersed with what we might call pseudo-quantitative claims or characterizations. To return to my opening example, I suspect that few would have a problem looking at *The* *New York Times* book review like the one I showed you all and calling it short. The point of comparing that review to others like it was to suggest that claims like that one have quantitative consequences that might make one generalization more accurate than another.

When we make intuitive generalizations, a range of common missteps such as anchoring bias, attentional bias, confirmation bias, and overconfidence are all potentially influencing what seems like a straightforward judgment.[[9]](#footnote-9) Positioning oneself against “positivistic, strictly quantitative, mechanistic, reductive and literal” measures and methods has become increasingly prevalent in the humanities, but the humanities have never abandoned our need to speak of external, observable phenomena and divide those phenomena into legible units.[[10]](#footnote-10) Bibliographers and book historians are often quietly reminding literary studies scholars that careful observation and tabulation lays the foundation for the close reading and literary theory that elevates “The Text” as an independent, often de-historicized and sometimes de-materialized object.

**Conclusion: Phases V, VI, VIII: Distribution, Discovery, and Reuse**

I have chosen to cluster distribution, discovery, and reuse for my conclusion because I see these steps of the data lifecycle as a different way to interrogate data not “as a natural resource but as a cultural one that needs to be generated, protected, and interpreted.”[[11]](#footnote-11) Distribution pertains to sharing one’s datasets in a way that readers and especially other scholars can find and access. It entails making one’s interpretive choices transparent, and demonstrating convincingly that well-known techniques of distortion such as P-hacking have not been employed to reach a false result.

A published article must cite its sources; ideally, a published article should also link to the datasets used as evidence of its conclusions. Reproducibility is the first and most important standard of distribution because a peer should be able to access data in a state that precedes as many interpretive choices as possible, changes seemingly minor choices like removing all punctuation or treating gothic fiction and ghost stories as members of the same category—and observe how dramatically those changes affect the results.

The timeliness of data releases should also be considered tremendously important because a work of scholarship associated with a dataset likely contains the most rigorously constructed interpretation of results for that dataset that will ever exist. Yet we should upload these data with the hope that others will find our ideas fascinating and will want to reproduce our results, to interrogate our findings, to build on our interpretations, to make our work stronger by criticizing it. And on that note, I look forward to hearing your criticisms of this paper.

1. *Willa Cather*: *The Contemporary Reviews,* edited by Margaret Anne O’Connor, which organizes Cather’s reviews by work reviewed rather than grouping them by periodical of origin or placing them in chronological order. [↑](#footnote-ref-1)
2. The faculty members with whom I worked are, Alison Langmead, Benjamin Miller and Annette Vee. Graduate students Daniel Libertz and Lucia LoTempio served as summer 2016 graduate student assistants on the project, and undergraduate Zach Luettgen did an independent study on book reviews as networks in Spring 2017. [↑](#footnote-ref-2)
3. In addition to documenting the idiosyncratic considerations of digitizing bibliographical objects like book reviews, our project has also focused on best practices for data modeling and data curation when working with these materials. [↑](#footnote-ref-3)
4. For example, Proquest’s *American Periodicals Series* and the *HathiTrust* collection. [↑](#footnote-ref-4)
5. “USGS Data Management - Data Lifecycle Overview.” [↑](#footnote-ref-5)
6. For examples of a range of data documentation practices, see Ted Underwood, “The Life Cycles of Genres”; Katherine Bode, “Thousands of Titles Without Authors: Digitized Newspapers, Serial Fiction, and the Challenges of Anonymity”; and Andrew Piper, “Fictionality.” [↑](#footnote-ref-6)
7. Suarez, “Book History from Descriptive Bibliographies,” 199. [↑](#footnote-ref-7)
8. See Gitelman, *Raw Data Is an Oxymoron*. [↑](#footnote-ref-8)
9. Anchoring is the tendency to overvalue a particular trait. Attention bias is the tendency to be affected by recurring thoughts. Confirmation bias describes a tendency where the “selection and evaluation of information that would confirm a focal hypothesis is given priority—or even exclusivity” (Rajsic, Wilson, and Pratt, “Confirmation Bias in Visual Search,” 1353). Overconfidence is a tendency to overvalue one’s own answers to questions, which might in turn prevent someone from taking a measurement or consulting an outside source. [↑](#footnote-ref-9)
10. Drucker, “Humanistic Theory and Digital Scholarship,” 86. [↑](#footnote-ref-10)
11. “Overview,” press materials for “Raw Data Is an Oxymoron.” [↑](#footnote-ref-11)