Hi everyone! Let me start off by saying that this draft isn’t quite where I wanted it to be for you. I’m having a particularly rough time tying together the argument in this section and smoothing the progression of ideas. Let me try to summarize it in full so you have some context going in:

This section sets up the conversation about how digital methods evacuate the complexity of their data. It contextualizes this convo with a brief history of technological development, which reveals the assumptions and values encoded into new technologies. It then takes a critical look at some examples of distant reading (Underwood, Da) to see how they similarly perpetuate the researcher’s assumptions to create a “reproducible” form of criticism. This practice is contrasted with DH practitioners like Johanna Drucker who deliberately skew metrics and equivocate the results from analysis, drawing attention to the ways that data is always constructed. Moving forward, I build off Drucker to propose a better way of doing distant reading, one which attends to the idiosyncrasies of embodiment and touch, and works in tandem with my section on Queer Form.

My main concern with this section is the plausibility of the argument itself and the way it comes across to readers who may not be familiar with this history.

* Do you buy the narrative I present here?
* Additionally, are there moments where I am overdoing it (I go into a bit of potentially irrelevant history), need to cut down, or add more context?
* Finally, what do you make of my comparison between Underwood and Drucker? Do you think I’m being unfair to Underwood?

PS: sorry for the spelling/formatting mistakes---I’m not using a spell checker heh.

\*\*\*\*\*

Before I turn to current examples of distant reading, it is useful to contextualize the technological development of digital tools in the latter 20th century. This contextualization reveals how the intentions guiding technological development contradicts common understandings about new information technology being progressive, democratic, or "free". In fact, many of these tools were created with conservative intentions and perpetuate cultural assumptions that elide complexity and difference. In what follows, I will briefly trace the development of networking and software technologies from the 1960s through the end of the 20th century, then turn to the "surveillance" technology of the last two decades, highlighting how more recent technology maintains some of the crucial assumptions from the last century. This contextulization, however brief, will help to situate the ways that digital humanists today approach the use of digital tools in their research methodologies.

First, the development of the internet, which is a global network of interconnected computers, is often credited for "democratizing" access to information. Early networks like Usenet, developed in 1979, popularized online message boards, file-sharing, and eventually e-mail. Built from ideals of open exchange and user agency in "an effort to break down modes of exclusion," the network was developed by people who wanted to communicate horizontally, practice improvisation and "hacking." (Rosenweig, "Wizards, Bureaucrats…, 1549). Moving to 1989, Tim Berners-Lee, a computer scientist at the European Organization for Nuclear Research (CERN), proposed the development of a distributed information system that would eventually become the World Wide Web.[[1]](#footnote-1) While working at CERN, Berners-Lee identified personnel access to the latest information across the center as a major problem for the organization’s workflow, lamenting that "Information is constantly being lost… often, the information has been recorded, it just cannot be found." Berners Lee saw information and people, the connection between human bodies with bodies of text, as the problem. He proposed a new resource for orienting researchers that was accessible, flexible, and emendable, initiating work on the Hypertext Transfer Protocol (HTTP) that would eventually become integral to creating the World Wide Web.

These positive narratives about Usenet and the World Wide Web dominate the history of internet. Less acknowledged is how networking technologies largely support a structure of control over its users. To begin with, the internet's early development was funded by two Department of Defense projects, the RAND corporation (then a Cold War think-tank), and ARPANET (the Advanced Research Projects Agency Network), which later became the Defense Advanced Research Projects Agency (DARPA). These US military stakeholders wanted to preserve command and control in the case of a catastrophic nuclear event,[[2]](#footnote-2) and reasoned that a distributed network would create national communication contingency.[[3]](#footnote-3) The RAND Corporation fist theorized the distributed network, and ARPANET formalized the new technology of packet switching, which is a method of grouping data into small packets that can be later reassembled at the final destination. In order to send information along the network, data has to be appended with protocols, or codes like HTTP, which impose structures on data to make connections possible. As Alexander Galloway points out, whether users know it or not, they "accept… universal standardization in order to facilitate the freer and more democratic medium" (147).[[4]](#footnote-4) The trade-off between access and standardization, freedom and control, is often invisible to the end user, who isn't aware of the packets that are constantly passing through their computer. Wendy Chun uses the image of a window to illustrate the two way direction of information traffic, how using the internet is also always being used by it. She warns: “If you believe that your communications are private, it is because software corporations, as they relentlessly code and circulate you, tell you that you are behind, and not in front of, the window” (22).[[5]](#footnote-5)

Major developments in technology also perpetuate racial assumptions. Moving from networking technologies to software development, Tara McPherson explores the parallels between the Operating Systems and race relations, to show how the development of computer software betrays hegemonic assumptions about whiteness and elisions of difference.[[6]](#footnote-6) She focuses on the key moment of 1960s United States, when Operating Systems, which is the foundational software that supports a computer's programs and basic functioning, developed alongside civil rights discourses. Her research focuses on how "the organization of information and capital" in OS development resonates in the struggles for racial justice: "Many of these shifts were enacted in the name of liberalism, aimed at distancing the overt racism of the past even as they contained and cordoned off progressive radicalism" (30). McPherson deconstructs the UNIX operating system which includes a hierarchical file system, a command line interpreter (the Terminal on Mac or Command Prompt on Windows), and a variety of software programs that are designed to work in tandem. McPherson points out that UNIX-based Operating Systems (like Mac and Linux) are distinguished by the ways that they partition and simplify complex processes into discrete components, similar to the ways that identity politics cordones off parts of the (social and technological) system into distinct units. While this cordoning was productive for the promotion of civil rights, it also, according to McPherson, "curtailed and short-circuited more radical forms of political praxis, reducing struggle to fairly discrete parameters" (30).

Crystallizing the intersection between Operating Systems and race relations, McPherson asserts that "Certain modes of racial visibility and knowing coincide or dovetail with specific ways of organizing data" (24). McPherson emphasizes the "rules" of UNIX philosophy, which lay out how UNIX's development prioritized the organization and simplification of data processing:

Rule of Simplicity: Design for simplicity; add complexity only where you must. Rule of Parsimony: Write a big program only when it is clear by demonstration that nothing else will do. Rule of Transparency: Design for visibility to make inspection and debugging easier… Rule of Representation: Fold knowledge into data so program logic can be stupid and robust. 26

The rules of "Simplicity" and "Parsimony" ensure that programs will be composed of small, interlocking parts that can be easily updated and transported to newer versions. The rule of "Transparency" flattens nuance and ambiguity, making program components as legible as possible. The rule of "Representation," particularly the suggestion to "Fold knowledge into data" reduces the complexity of raw data, so that it can be easily input into multiple processes. According to McPherson, all of these rules work together to shore up the central design theory of "modularity,"[[7]](#footnote-7) which stipulates that components are self-contained and interoperable, so they can be independently created, modified, and replaced without affecting the whole system.

The role of control in creating the internet and the emphasis on data reduction in developing operating stystems leave their legacies on 21st century digital technology, where race becomes collapsed into data. Echoing McPherson, Ruha Benjamin asserts that technology reproduces social inequities under the guise of objectivity and progressivism.[[8]](#footnote-8) Turning to technology, Benjamin explores how innovations in Artificial Intelligence and algorithmic computing extend racist paradigms into ever new tools, particularly in data gathering and surveillance. The creators of these new technologies mark, track, and quantify blackness, for example, in databases for healthcare or financial services that associate "black names" with criminality (Benjamin 5). With each update, technology is continually promoted as efficient and progressive in a way that masks how it exploits data about its subjects. Benjamin explains, "we are told that how tech sees “difference” is a more objective reflection of reality than if a mere human produced the same results… bias enters through the backdoor of design optimization in which the humans who create the algorithms are hidden from view" (5-6). As she points out, "the road to inequity is paved with technical fixes" (7). Like the creators of UNIX, the creators of such tools and algorithms operate under assumptions of white universality that inevitably marks blackness as "other."

I now turn to computational methods, seeing how they bear out some of the legacies from the above technological histories. Practitioners of "distant reading," a critical method at the intersection of Literary Studies and Data Science, use quantitative analysis to study works of literature. This process involves deploying computer programs to clean, categorize, and count elements in textual data, and is often followed by interpretive analysis, where the critic engages the results of quantification from a humanities lense. More often than not, distant reading is combined with close reading methods, as crtics will use the results of quantitative analysis to identify key moments from the text that merit closer attention.[[9]](#footnote-9)

According to its practitioners, distant reading is most useful for the ways it allows connections to emerge among vast amounts of textual data. Critics who do this work often emphasize the problem of literary scale and human attention, because distant reading allows them to handle the thousands of books in literary history without actually reading these texts. One prominent practitioner of Computational Literary Studies (CLS), Ted Underwood,[[10]](#footnote-10) harnesses the power of quantification and machine learning to glimpse what he calls the "distant horizon" of literary trends across centuries. His argument convincingly begins with the observation that human capacities of sight, attention, and memory preclude them from grasping the larger patterns of literary history across time. Distant reading, where "distance" means abstraction, or the simplification of textual data into computable objects such as publication dates and genres, allows critics to see connections amid the swarm of overflowing information.

Among distant reading practitioners, Underwood's approach is unique in that he models the ways that human assumptions can affect the results of analysis. Underwood is careful to point out the subjective nature of his method, which he calls "perspectival modelling," by turning it into an object of study. He uses machine learning, or programs "trained" by certain data sets, to create models that can then make predictions on other datasets. He explains that, "Since learning algorithms rely on examples rather than fixed definitions, they can be used to model the tacit assumptions shared by particular communities of production or reception" ("Machine Learning and Human Perspective" 93). One of his projects examines gender roles in novels from the 18th century to the 21st century by using a machine-learning model to "guess" the sex of a fictional character based on the words associated with that character. Underwood explains how the test is configured:

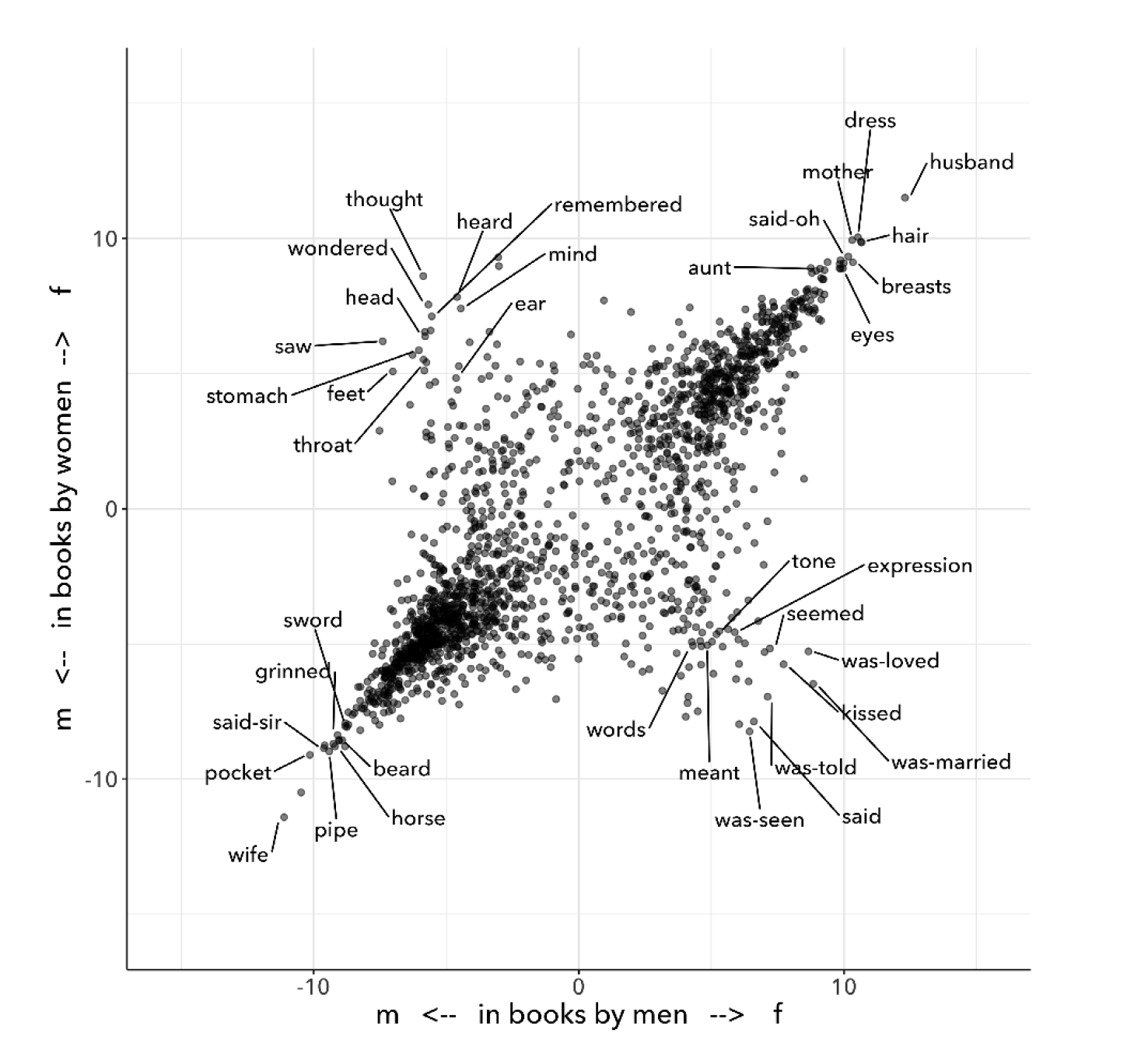
We represent each character by the adjectives that modify them, verbs they govern and so on–excluding only words that explicitly name a gendered role like *boyhood* or *wife*. Then, we present characters, labeled with grammatical gender, to a learning algorithm. The algorithm will learn what it means to be 'masculine' or 'feminine' purely by observing what men and women actually do in stories. The model produced by the algorithm can make predictions about other characters, previously unseen. *Distant Horizons* 115

In simplest terms, the program studies some given adjectives associated with a male or female character in order to make predictions about other characters' genders. Inevitably, the resulting output is always determined by this initial input. Underwood carefully asserts that these models reveal, not the truth of literary histroy, but the approaches and choices made by those who create the models: "Machine learning algorithms are actually bad at being objective and rather good at absorbing human perspectives implicit in the evidence used to train them" ("Machine Learning and Human Perspective" 92). This particular model reveals that that, over time, gender roles in novels become more flexible while the actual number of female characters declines (*Distant Horizons* 114). The graph shows a steady overlapping of words traditionally associated with women, such as "heart," with words typically assoicated with men, like "passion," toward the middle of the 20th century. One of the many explanations for this result, Underwood reasons, is that the practice of writing became more commonly pursued as a male occupation in the middle of the 20th century than it was previously (*Distant Horizons* 137). This fact, coupled with the tendency of men to write more about men than women, suggests why less women writing would led to a decline in female characters. This explains how Underwood's seemingly paradoxical conclusion, that gender roles become more flexible while the actual prevalence of women dissapates from fiction, might be possible.

However, the results of Underwood's "perspectival modeling" can only be as good as the questions he asks. From a critical gender perspective, Underwood's approach imposes the very structure that he is attempting to deconstruct. In other project, he where he similarly measures the "transformations" of gender across time periods, he explains that simplification is necessary ("Machine Learnig and Human Perspective" 93):

I recognize that gender theorists will be frustrated by the binary structure of the diagram. To be sure, this binary has folded back on itself, in order to acknowledge that social systems look different from different positions in the system. But the diagram does still reduce the complex reality of gender identification to two public roles: men and women. I needed a simple picture, frankly, in order to explain how a quantitative model can be said to represent a perspective. "Machine Learning" 98

Underwood admits that he needs a "simple" model in order to bring into relation the dynamics of gender (See Fig. 2).[[11]](#footnote-11) However, he underestimates the extent to which his initial assumptions determine the final result. Although he considers the possibility that he finds a structural tension between gender "because [he] explores gender, for the most part, as a binary opposition" (/Distant Horizons 140), he neglects to consider how the collapsing of gender into a single graph perpetuates the structural categories of male/female in a way that is neglects the assumptions behind such a category.[[12]](#footnote-12) Moreover, the issue is not just with the assumptions at the outset which reproduces the result, but with the guiding question of the entire project, which is not about deconstructing gender, but about reifying it. To begin with, why should humanists seek to automate the conscription of gender norms within these terms? Asking a machine to replicate the conscription of gender for the purpose of seeing how male and female roles in novels change over time only creates a model of gender that is "simple" enough to be computed by the system. How does simplifying the concept of gender contribute to our study of it? The results of using the machine can only be as good as the questions we ask.

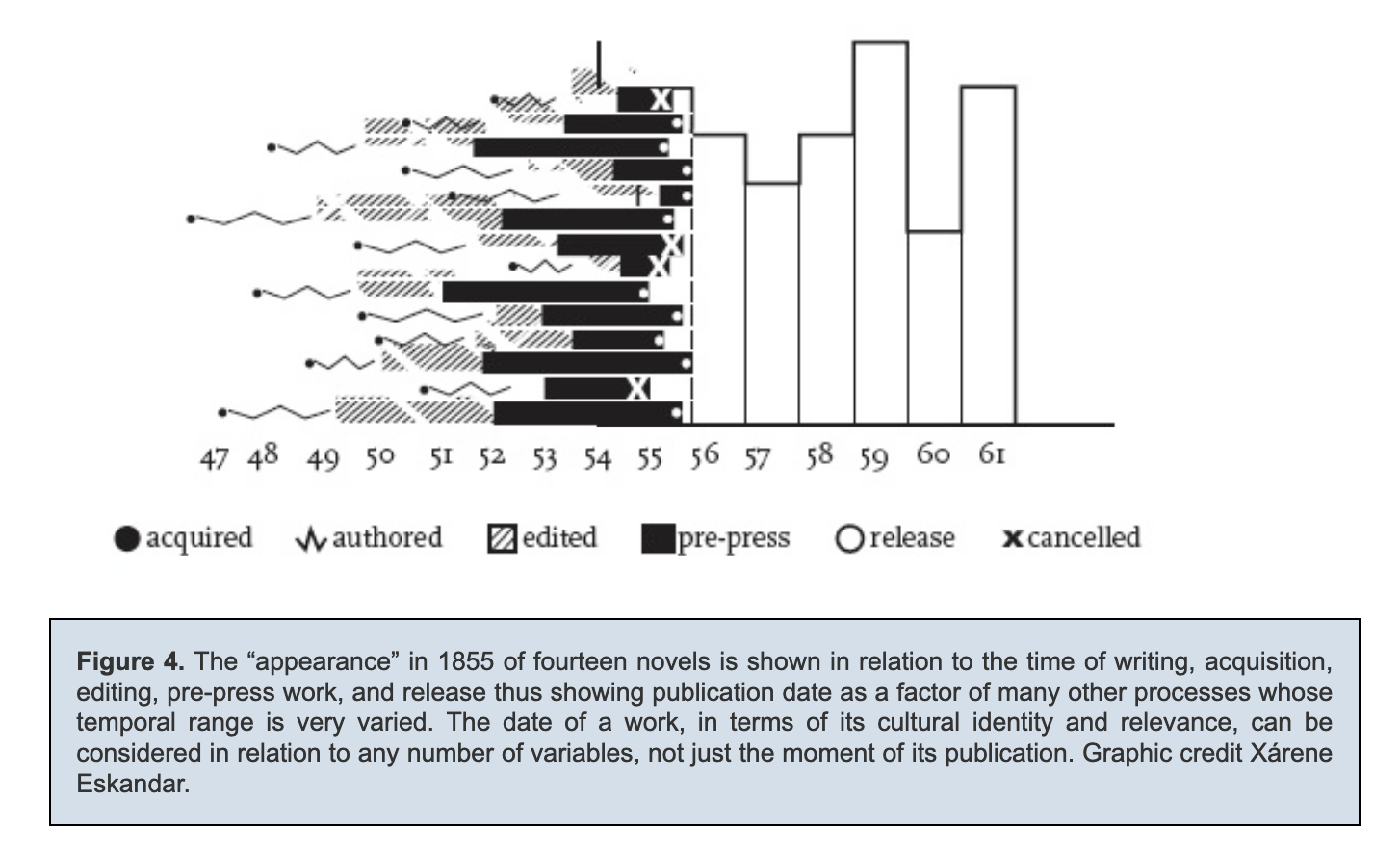


Critiquing scholars like Underwood, Nan Z. Da argues that quantitative methods are ill-suited for literary criticism. She accusses Underwood and other distant reading practitioners for trading "speed for accuracy, and coverage for nuance" (620). Of her many gripes with quantitative methods, which include "technical problems, logical fallacies," and a "fundamental mismatch betwen the statistical tools that are used and the objects to which they are applied" (601), she emphasizes the lack of reproducible results, the idea that one researcher's process can be reproduced by another with identical output, which is essential to statistical methodologies. She demonstrates with an experiment of Topic Modelling, which is the processing of large texts in order to generate a number of "topics" within the corpus. Researchers often use Topic Modelling as a way of speed-reading a massive corpus to get a sense of what it is about without having to actually look at the text. Da attempts to verify the results of a Topic Modelling experiment by replicating the process on her own machine, a replication that fails. She concludes that, "if the method were effective, someone with comparable training should be able to use the same parameters to get basically the same results" (628-629).[[13]](#footnote-13) For Da, reproducibility of method is a benchmark for reviewing and assessing the efficacy of quantification.

Despite their vastly different committments, scholars like Underwood align with Da on the value that they place on reproducibility, which is an ultimately conservative investment. Underwood demonstrates how the critic reproduces their assumptions in the questions and data used at the outset in a way that structures the final result. Da's emphasis on the reproducible suggests that, to be useful, quantitative literary criticism ought to resemble something more like statistical analysis: if the method can be verified, can be copied and reproduced, then the interpretive conditions might be universalized.

Underwood and Da overlook the way that quantification can be used to disrupt assumptions or reveal the constructed nature of data. In contrast to Underwood and Da, Johanna Drucker is careful to dispell the illusion of "raw data," which comes already reduced to fit whatever parameters required by analysis. Because data always undergoes a transformation in order to be quantified, its complexity is always reduced. As a result, Drucker argues, quantification techniques such as visualizations in graphs and charts inevitably misrepresent the data they are meant to convey. To illustrate this process, Drucker presents a chart displaying the amount of books published over several years. The chart appears to convey production during this specific time period, but Drucker explains that publication date is an arbitrary metric for capturing production.[[14]](#footnote-14) She brings to the surface all the assumptions made in such a metric, for example, the limitations of "novel" as a genre and the connotations behind "published," which suggests date of appearance, but has no indication of composition, editing, review, distribution. Each piece of data carries with it the result of many interpretive decisions, that carry with them varying degrees of opacity, which are all necessary in order to present complex concepts like book production as a bar on a chart. Drucker explains: "the graphical presentation of supposedly self-evident information… conceals these complexities, and the interpretative factors that bring the numerics into being, under a guise of graphical legibility" (Drucker par. 23).

To resist the reductions of "data," a term that deceptively connotes that which is "given," Drucker proposes thinking of data as "capta," which suggests that which is taken. Drucker's "capta" is deliberately creative, turning graphical expressions into expressive metrics: components used for measurement, like lines or bars on a graph, break, blur, or bleed into one another. Objects are not discrete entities, but interact with the other objects in the visualization. For example, in a bar graph of book publications by year, she warps the graphical metrics, making some of them fuzzy, wider, shorter, in an attempt to show that publication as a metric elides other information such as composition, editing, purchasing, etc.



Emphasizing "capta" is a way of figuring elements that have been reduced, resolved, or ignored in traditional quantitative analysis. Drucker makes evident what is overlooked or assumed when dealing with complex subjects by muddling (rather than simplifying) the relationship between elements.

[The next step, which I want to take here, is to show how paying attention to the assumptions (deconstructing) offers a return to embodiment, thinking about how the user engages with the quantifications. This allows us back into the concept of touching–mirrored in the queer form section]

# 

1. Berners-Lee, Tim. “Information Management: A Proposal.” CERN (1989). Tim Be [↑](#footnote-ref-1)
2. For more information about computer technology helped develop the discourse of centralized command and control, see Paul N. Edwards, *The Closed World: Completers and the Politics of Discourse in Cold War America* (Cambridge, Mass., 1996). [↑](#footnote-ref-2)
3. Stephen J. Lukasik, the deputy director of DARPA, explains that the goal of creating new network technologies included: "to meet the needs of military command and control against nuclear threats… and improve military tactical and management decision making. Lukasik, Stephen J. (2011). "Why the Arpanet Was Built". IEEE Annals of the History of Computing. 33 (3): 4–20. Bruce Sterling, "Short History of the Internet," 1993 [↑](#footnote-ref-3)
4. Galloway, Alexander. **Protocol**, 2004. [↑](#footnote-ref-4)
5. Chun, Wendy, *Control and Freedom: Power and Paranoia in the Age of Fiber Optics,* 2006. Print. [↑](#footnote-ref-5)
6. Tara McPherson’s "U.S. Operating Systems at Mid-Century: The Intertwining of Race and UNIX," Race After The Internet, ed. Lisa Nakamura and Peter A. Chow-White. Routledge, 2012. [↑](#footnote-ref-6)
7. Potentially revise and deepen this section by linking to Barad & Haraway on situated knowledges and feminist science: Being modular in itself isn't bad, as long as you are aware of the ways that modularity creates limitations/reductions of data. Modularity needs a critical awareness of its own tools. [↑](#footnote-ref-7)
8. Her work also extends Michelle Alexander's ideas from *The New Jim Crow* (2010), which argues that modern society perpetuates racist violence and segregation by criminalizing race through the war on drugs and mass incarceration. [↑](#footnote-ref-8)
9. Andrew Piper's methodology, which he calls "bifocal" reading, demonstrates how distant and close reading are used together, with distant reading providing the context or framework that guides close reading"“We are no longer using our own judgments as benchmarks… but explicitly constructing the context through which something is seen as significant (and the means through which significance is assessed)…. It interweaves subjectivity with objects” (Piper, Andrew. Enumerations: Data and Literary Study, 2018, 17). [↑](#footnote-ref-9)
10. Underwood, Ted. *Distant Horizons*, 2019.; Underwood, Ted. “Machine Learning and Human Perspective.” PMLA, Vol. 35 No. 1, January 2020, pp. 92-109. [↑](#footnote-ref-10)
11. He measures the "gendering of words used in characterization" ("Machine Learning and Human Perspective" 95), that is, gender portrayed in novels by women and in novels by men. The verticle axis visualizes the representation of words by women, and the horizontal by men, with positive numbers signifying overrepresentation of these terms. So terms on the top right are words that are used often by men and women writers, and terms in the upper left and lower right are ones used most often by women and men, respectively. [↑](#footnote-ref-11)
12. Add a quote here from Laura Mandell on F/M categories? [↑](#footnote-ref-12)
13. Da's emphasis on the “reproducible” in CLS extends Franco Moretti's originating call for a “falsifiable criticism”: both advocate for a methodology that is as reliable and verifiable as the social sciences. According to Moretti: “Testing” literary interpretations be the same process as in scientific disciplines – demanding that interpretations are “coherent, univocal, and complete,” and are tested against “data” that appears to contradict it (*Signs* 21). (another quote: “The day criticism gives up its battle cry ‘it is possible to interpret this element in the following way,’ to replace it with the much more prosaic, ‘the following interpretation is impossible for such and such a reason,’ it will have taken a huge step forward on the road of methodological solidity” (*Signs* 22).) [↑](#footnote-ref-13)
14. Drucker implicitly refers to the first chapter from Franco Moretti's *Graphs, Maps, Trees* (2007), throughout which Moretti graphs novels by their publication date between 1700 and 2000 and draws conclusions about the relationship between genre and generations of readers. [↑](#footnote-ref-14)