Week 2, Class 2 reflection

Readings:

- Manovich, Lev. "Database as Symbolic Form." *Convergence: The International Journal of Research into New Media Technologies*, vol. 5, no. 2, June 1999, pp. 80–99.
- Pomerantz, Jeffrey. "Introduction." *Metadata*, The MIT Press, 2015, pp. 1–18.
- Gebru, Timnit, et al. "Datasheets for Datasets." *Communications of the ACM*, vol. 64, no. 12, Dec. 2021, pp. 86–92.
- **Optional**, complimentary reading for further interest:
    - Hoffman, Gretchen. "How Are Cookbooks Classified in Libraries? An Examination of LCSH and LCC." *Proceedings from North American Symposium on Knowledge Organization*, vol. 4, no. 1, 2013, pp. 100–11.

Alison Fortenberry

11:03 PM I understand the distinction Manovich is drawing between cinema ordered as a narrative and cinema presenting as a database, but I'm wondering if the role of database formation plays a role in the narrative structure and if cinema can ever be a true representation of a database. Picking and creating media to go into the database requires intention about what will be photographed/videoed, how the shot will be set up, what kinds of items will be captured, etc. I don't mean to imply that databases aren't databases when someone creates and populates them, but I feel like the idea of narrative construction begins before pieces of media are organized. Can cinema ever truly be separated from narrative if the creator of a film is involved in the data collection process? Gebru et al. propose that "every dataset be accompanied by a datasheet that documents its motivation, composition, collection process, recommended uses, and so on" (Gebru et al. 1). I think this further speaks to the idea that there is some layer of intentionality in creating a dataset, and strengthens the idea for me that cinema cannot be completely separated from narrative, even when it is not specifically ordered, if the film's creator has played a role in the construction of the dataset. Data is not randomly compiled in this instance, it is, to some degree, curated. On a separate note, the breakdown of what metadata is and its different strains in the Pomerantz article was really clearly written, and helped me understand a concept that's felt a little abstract and inaccessible for me in the past.

Melissa Woo

10:27 AM I find Manovich's argument about the emergence of the database as a primary cultural form in the computer age very compelling, particularly his point that new media objects do not follow a linear narrative but exist as collections of individual items without a specific sequence. I took a freshman sem on systems, where we talked a lot about the difference between linear and systematic approaches to modelling and storing information, and particularly about what gets lost or reduced when a complex system is linearized. This, particularly as it applies to datasets, makes me wonder how we should best perceive and organize information, perhaps challenging the traditional linear modes of expression. Furthermore, his assertion that the world can be viewed as an unstructured collection of data underscores the profound impact of the database as a symbolic form on our understanding of reality. Gebru's emphasis on datasheets for datasets aligns with the growing need for transparent and well-documented data sources in technical fields, reflecting the broader implications of data representation and interpretation discussed by Manovich. In my "Ethics of Computing" class last semester, we actually spent a class developing a datasheet for a prominent

database that didn't have one yet – it was actually really hard to find all of the information that was buried deep in methodology and appendix sections. It really highlights how hidden important info can be. Metadata, as well as nonlinear data and datasheets, made me think about things we take for granted and that fly below the radar. If we take the time to investigate each of these things, I wonder what other questions or discoveries we will be able to make.

Clay Glovier

10:28 AM I found Pomerantz's article on metadata fascinating. It was interesting to learn about how the government can weaponize our cell phones to track our locations and create charts matching us to those we choose to call and communicate with. I think that this is why many Americans supported Edward Snowden even though he gave many of our national security secrets to Russia and to the rest of the world. I thought it was also interesting how Pomerantz discussed the purpose of library catalogs, and how they save libraries from falling into chaos. Without knowing where each book is, it would be almost impossible to find texts for research. I also enjoyed reading Gebru's article about datasheets and datasets. I think it is essential for users of machine learning technologies to catalog how they collected the data used to train AI models, to avoid bias and to enable others to check the work to make sure the data was provided accurately. Manovich's article was also enlightening; it was interesting how he noted many things we use often are databases, such as video games. While we may use these technologies daily, I don't think that most gamers think about the fact that they are engaging with a dataset and becoming part of the dataset, acting like a program by responding to stimuli in a set way. An example would be killing enemies in the game on sight, without second thought like a program would attack a virus upon detection. In sum, these articles were all really interesting and I enjoyed engaging with each.

Anya Kalogerakos

10:55 AM The readings for Thursday were all centered around metadata, with a good general definition being given by Pomerantz in Metadata–that metadata is a sort of identifier that locates data within what would otherwise be a chaotic collection of data. Manovich discusses how walking through these collections of data is a new media experience. On one hand, there is no narrative connected to the seemingly random data within databases, but on the other hand, if you travel in a specific and meaningful way, some level of narrative can be constructed through the database. However, as Hoffman and Gebru discuss in "How Are Cookbooks Classified in Libraries?" and "Datasheets for Datasets" there are certain pathways through databases that can cause harmful effects such as societal bias and a lack of inclusion. I found the last two articles the most interesting, as the topic of bias in data is so important in today's technical disciplines and machine learning is being used more and more to inform important decisions and systems. I appreciated hearing about a potential tool to combat data bias–the datasheet–and I think it would be very interesting to read one of these datasheets in action to see if it causes me to think about what potential harms there could be in using the data within the database. However, the datasheet as a solution to algorithmic bias does not quite feel complete, and I am interested in the other methods researchers can employ to ensure bias does not creep into their data and expand further into machine learning training.

James Sowerby

10:59 AM I thought Gebru's article was extremely clearly written and a good example of measures being taken to safeguard against harmful tendencies of large language models. As someone who is relatively unfamiliar with the field and do not know about the ins and outs of sourcing data or training models, this makes manageable and impactful steps clearer. I feel like this article is important for a lay audience like me

in demystifying these tools—which have very important uses already. One of my annotations—and this is something that I am still wondering about—is how this has progressed in the time since the article was published in 2018. Gebru mentioned how some big tech companies had seemed very receptive to it then, but what is the state of the practice now? What other safety measures are being taken? My brother, who works at a tech company, did some work on this topic once and I found it very interesting, but didn't really understand it in any depth. I also really enjoyed Pomerantz's article on metadata. It, too, was extremely clear and I felt like I got a good grasp of the term through the many examples they gave. It got me thinking a little bit about the kind of metadata I work with in my own life. I mentioned a few examples like my iPhone pictures or an online translator I use, but I started thinking since I read it last night about the metadata that Princeton has about me. Obviously it knows a great deal about me, but a lot of this has been given to them by me in my application or academic sign in forms. I wonder what they collect in the day-to-day about my actions. They know what buildings I prox into and when—that's metadata, I think, because they don't know what I'm doing there. Anyway, there are tons more examples and furthermore I'm sure I don't even know the half of what they observe about me. Just something to think about.

Pia Bhatia

1:28 PM The main issue with my reading of the texts for this class concerns the definitions of some key terms. In the Manovich text, certain new media objects “are collections of individual items, where every item has the same significance as any other,” which seems extremely similar to the definition of data itself – what distinguishes these two terms if anything? I also want to challenge his ideas of how trends in narratives have changed (in the novel, films, etc.). Manovich states that new media plays more or even totally rejects the linearity that used to be more common in these art forms. Maybe I’m being too nitpicky, but I think that these forms are inherently limited to being ‘collections’ – no film, even as it follows a very singular and streamlined arc, is capable of capturing exactly every single instance that has caused the plot to unfold. I also wouldn’t say that the advent of data has created a ‘new media’ that feels cumulative – art movements like Surrealism have played with collaging footage and interweaving seemingly disjointed pieces in a similar fashion. That said, I think I would need to interact with more art that fits Manovich’s definition of new media in order to understand their singularity better.

Colin Brown

7:07 PM First, the Gebru et al. was a very interesting article to read in tandem with currently being in an intro to machine learning class. One of the very first lessons the professors emphasized was that machine learning models can only be applied to scenarios that are identical to the scenarios from which the data came. In other words, a ML programmer has to know an incredible amount of information about their data if they want to make a valid and accurate model; this made the authors’ motivation make a lot of sense to me. A possible concern with universal datasheets is that the kinds of data we have is just so vast; I’m sure it would take a large amount of trial and error, collaboration, and debate to make these datasheets work for everything from personal medical history to YouTube videos. It is also a neat connection, then to see that this information about data is exactly the focus of the Pomerantz paper - metadata! I liked how he made a point that libraries were among the first to use metadata for organizing information; after all, libraries were our first “internets” of data, so it figures that they would pioneer how to effectively make books accessible. This article also made me recall our class discussion about browsing. In retrospect, it seems that much of what guides our browsing is metadata, we just weren’t labeling it as such yet.

Raphaela Gold

8:36 PM I really enjoyed thinking about the questions Manovich raised about the relationship between narrative and data and the modern cultural shift from emphasis on narrative to data. However, I wasn't sure about his categorization of narrative and data as merely corollary to one another. To me, they're deeply interlinked, and we can often pull a narrative, or even many narratives, from a single dataset. It reminded me a lot of the image from the Rosenberg reading of data as a constellation, always plural rather than singular. While I agree that data is not a narrative in and of itself, I think it's possible for data to serve as a prerequisite for narrative. Are data and narrative really so different from one another? And is it true that modern society skews toward data and algorithm and more than our predecessors? The Pomerantz reading raised some questions for me as well. I really enjoyed reading about the Edward Snowden case and the shock of metadata tracking through phones, because now it seems so obvious that our data and information are monitored. We are accustomed to our internet activity being tracked, and a choice to be on social media is in many ways a choice to give up privacy. I think Pomerantz captured this when he described metadata as "infrastructural", like an electrical grid or a highway system. This comparison of metadata to infrastructure made me think about how society is structured differently for different individuals and groups of people, and the infrastructure of metadata must be as well. How does metadata play out among different demographics and groups of individuals? For some, the tracking of metadata may be harmless, while for others I could definitely see it being very dangerous and potentially jeopardizing jobs and livelihoods. As for the datasheets for data sets reading, I noticed some new terminology about how we use data. Up until now, we've been mainly focusing on what data is and how it came to be. This article, however, began to address how we use data – the logical next step after deciding what qualifies as data. I got the sense from the article that once we have data, it's up to us to figure out what narrative we're going to tell with it, or what conclusions, if any, can be drawn from it.

Pippa LaMacchia

9:37 PM Manovich's article highlighted some interesting questions I had been considering as we've begun to make our way through this material. It seems to truly place the "humanities" part of this course into perspective because of the focus on a narrative versus a database. The connection between cultural creativity and a newly digitized world was fascinating as Manovich writes, "In general creating a work in new media can be understood as the construction of an interface to a database." As someone who studies narrative and storytelling I don't quite know how I feel about the technological and algorithmic transformations Manovich describes. On one hand there is a deep thread of hope because we now have the ability to catalog and organize all the information that is impossible for one person to absorb. On the other hand, there is a sense of reality and beauty that comes from that impossibility and its digitization (organization into databases) is counterproductive in a sense. I was also struck by the ability to reject traditional linear narratives as datasets exist simultaneously and the linear choice is only one option of many. The further elaboration of metadata in Pomerantz's article further elaborates this concept as literary and digital worlds collide. I cannot help but ask whether or not it's good that metadata is capable of creating simpler interfaces for complex systems.

Andrew

11:21 PM First of all, I think that Manovich's article is written in a very convoluted manner that does not provide a clear narrative through his debate between databases versus narratives, which is ironic. Near the end of the article, he says that "cinema already exists precisely at the intersection between database and narrative," yet how film any different literature or art? Books use a database of every possible word in a certain language (which is finite) yet have a narrative. Art and photography have no sequential order yet can

have a perceived narrative within one frame. Now, when it comes to to online databases in websites, the individual stops becoming the passive reader or observer/watcher but now becomes the 'film editor' circling through all the possible database pages with hyperlinks, thus creating a temporal, sequential 'narrative' through the chaos. I don't fully understand the distinction that Manovich is trying to make. Everything he mentions seems to be at the intersection between databases and narrative, the only difference is individual agency. Also, I very much agreed with Gebru's article as I believe, in the chaos of information and data in the modern day and the fact data has lost its connection to validity, it is extremely important to set parameters and boundaries to allow data to become facts and synonymous with truths. Perhaps legal requirements should be taken into account when dealing with the future of data.

Talia Goldman

12:55 AM The readings, particularly the Pomerantz and Manovich articles, rely heavily on metaphor to describe metadata and databases, which I thought were helpful in communicating how entrenched we are in metadata, databases, and algorithms, even outside of a strictly digital sphere. As an art history major, I was interested in Manovich's use of CD-ROM "virtual museum tours" as an example of ways to approach a database–here, the objects in the museum's collection–beyond a chronological narrative, with metadata-focused approaches, such as organizing by medium or artist, being just as valuable. I would love to further explore the museum as an extended metaphor for datasheets (comparable to object provenance research), metadata (such as medium and artist), and the database (the whole collection). In a way, the internet is a museum of data that can be endlessly curated and is ever-expanding, as Manovich notes. However, Manovich also says that an ever-expanding internet constitutes a collection and is unable to encompass a narrative in its constant change. I am not sure that I agree with this point--even as museums collect new art and reinterpret its collections, these changes become part of its history rather than making it immune to a narrative. I also enjoyed Gerbu et. al's "Datasheets for Datasets," which considered how to navigates ethical considerations related to Pomerantz and Manovich's discussion of metadata and databases. I felt that the questions suggested for inclusion in datasheets presented concrete ways to handle the potential ethical issues in the creation and use of a dataset, grounding DH in the value of transparency in a world where the true influence of data in our day-to-day is not always clear to us, as Pomerantz explains. Overall, I found these readings quite interesting, helping me understand the relevance of the logic of metadata and databases outside the digital world along with the ways DHers may try to regulate this new media. (edited)

Emanuelle Sippy

2:03 AM The Manovich reading, in particular, raises important questions about how we are inclined to set up a binary relationship between "narrative" and "data." His example about the oral histories of holocaust survivors is instructive of this. Their stories are both data and narrative. Moreover, Manovich uses film to complicate this binary, thinking about the syntagmatic and paradigmatic. I think this can be applied to other narrative forms as well, including photography, journalism, creative nonfiction, memoir, and poetry. Manovich's point that the developments of linearity in film take place in the context of the post-industrial period (99) echo Pomerantz's assertion that what is designated as metadata and what is designated as requisite information/processes is, of course, contextual (10). I am interested in looking at other examples of how context determines or plays an outsized influence on what we deem "metadata" and what we seem "narrative." As not only scholarship and art making but also everyday life continually shifts in the digital world, it is worth considering how historical and contemporary lifeways push back on drawing clear lines between metadata and basic information, as well as narrative and data. When Manovich suggests that we ought to be more aware of metadata in our lives, akin to a highway system, that we view as infrastructure

that requires maintenance (3), I was wondering about concrete ways educators might promote this kind of awareness in educational settings; it seems that "Datasheets for Datasets" is this—a tangible way of promoting critical thinking not only about datasets alone but how they came to be.

Layla Williams

2:59 AM At the beginning of this course we discussed that one of the motivations for studying the digital humanities could be to present information in a new way so that people outside of the field could understand the new information that is being brought forward. And I think that the introduction to the metadata chapter by Pomerantz was a great example of this (in a semi-related way). For example, I have heard of the phrase “metadata” a lot in my life (usually in negative contexts or in spy/heist films), but I have never fully understood what it was. In my mind, I imagined it being an abundance of green ones and zeros, in the style of The Matrix. However, Pomerantz’s metaphor, which describes metadata as a map, was a helpful way to explain the information because it was a tool I had experience with already. In addition, Pomerantz also made the metadata seem approachable for someone who may not be as experienced with coding and the digital humanities, which, again, achieves the goal of presenting information in a more accessible way. I hope I am able to carry this experience into my data biography assignment as I keep in mind the uses of data sets and metadata and the audience that I might be presenting the information to.

Helen Gao

3:21 AM In “Database as a Symbolic Form”, I found the discussion of games vs narratives to be quite interesting, as it was focused on the user/reader experience. However, sometimes it felt like the author was trying too hard to contrast databases and narratives, and it seemed like the 'human' side of databases was not fully discussed. In contrast, “Datasheets for Datasets” really investigated the 'human' side of databases: more specifically, the possibility of human bias in database creation. This was a very important topic, and the list of questions seemed quite comprehensive. While I'm somewhat doubtful that this will become the industry standard anytime soon, since it's probably not very profitable for companies to dedicate time and money to making their datasets conform to the expectations set by this article, it definitely seems like a step in the right direction to helping make datasets more transparent (and hopefully improve the results of ML models). I also liked that this reading was relatively modern in comparison to the other articles we've read. Finally, the Pomerantz introduction was very accessible, and I liked that the author included analogies to non-digital formats in order to explain the metadata. The explanation of the evolution of how we store information was also quite interesting, and sort of reminded me of our previous readings/discussions about how new technologies are constantly replacing old ones.

Ethan Haque

8:43 AM My favorite article of the ones we read was Gebru's article. Just from that single article alone it's easy to tell that she has a deep understanding of both computer science and tech ethics. I worked for a lab for about two and half years studying stigma and bias from a psychological and technological perspective, and many of the ideas she touched on resonated with me. Having a sort of nutrition fact card for datasets will certainly increase the quality of the datasets in the wild. Those without these datasheets will be scrutinized more closely and datasets with low quality datasheets will show a lack of care for the dataset as a whole. Datasets with high quality datasheets will have an extra level of polish that show a deep care for the data and its uses. Some of the questions she poses, like questions about consent and privacy, are often left out of papers whose entire purpose is to describe a new dataset for a particular task. She demonstrates a vast amount of domain knowledge that I don't think the other readings achieve. The Pomerantz article

discusses the question of what is metadata and I don't think captures some of the more fundamental aspects of what it is in today's world. So many things could be considered metadata yet the underlying data they are describing is sometimes separable and sometimes not separable for its metadata. The separability is also highly dependent on the application. If you can't separate a piece of data from it's metadata, isn't that metadata just part of the data itself? It seems weird to try and classify and describe all these different kinds of metadata when most of it is contextual. The Manovich article was the most dubious of them all for me. It felt like he knew about the existence and use cases of a lot of technologies, but didn't really know how a database works or how to build and maintain a large database. Like his goal was less about being correct and more about just saying something interesting.