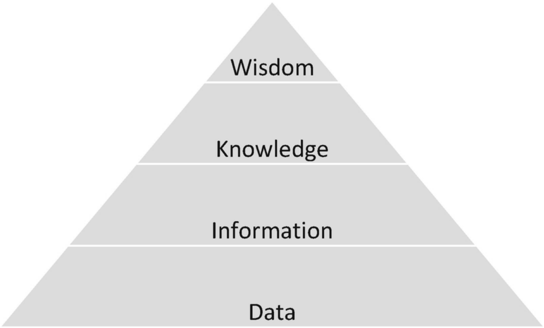
Making sense of your knowledge

“The goal is to turn data into information, and information into insight.”
– By Carly Fiorina, ex CEO of Hewlett-Packard



Lisa Golla
au713978@post.au.dk

An essay for the course
Curating data
Magdalena Regina Tyzlik-Carver
Winter term 2022

Cognitive science
University Aarhus
Denmark, Aarhus
December 2022

Making sense of your knowledge

I. Introduction

Currently we live in an era which is characterized by concise digital features. There are a lot of different approaches trying to capture the concept and the idea of the spirit given by the age. A few for example call it the "network era (...) characterized by vast and dynamic social connectivity, facilitated by widespread digital and electronic media" [1], whereas others rely more on the feature of Big data and name it therefore "era of Big Data [...] enabled by the widespread availability of electronic storage media, specifically mainframe computers, servers and server farms, and storage area networks." [5] and once again another view is to call the current era the "digital era" [3] in which "digital technologies play a prominent role in shaping up and regulating the behaviors, performances, standards, etc., of societies, communities, organizations, and individuals. " [4]. What is the inference out of this knowledge? What have all of these explanations in common? In the first place, it is suitable to say that each of these explanations try to describe the same phenomenon from a different perspective. The network era for instance, focuses mainly on a social media perspective, highlighting social connectivity and the widespread use of media. The named vast and dynamic can also be found in the described Big data era, whereas the digital era emphasizes especially the regulation and manipulation of behaviors. We can conclude that the current era has complex and constantly changing components which makes it hard to grasp and creates a need for multifaceted portrayals. Specific key elements of the current era, whatever you want to call it, are *data, information, knowledge, big data, the process of data curation*. Especially in order to make sense out of these notions, they will be classified into a specific context, and also to be defined properly in section II.

This paper is an overview of how I personally define given concepts that were learned in the seminar Curating Data, as well as 2 critical questions concerning the link of curation and data are posed. During the course of this paper, I will stress the importance of these questions and also I will connect my thoughts to the learning and assignments of the seminar where useful. As was already indicated, in the current era we are in, data are produced to vast amounts, and are everywhere. Critical movements and provocations arose which are dealing with the so called *data life cycle* which is the process starting with the production and ending with archiving and storing. During this specific process concerns dealing with ethics, equity and biases emerge, since data became an object of economy, research and even governance. Who has access to the data and thereof power? How is the data organized, collected and modeled and who is performing that? The idea is to emphasize and explain where in this explained image the process of curating data is located and what issues it is facing.

The 2 critical questions I want to pose in section IV are directly connected to that. Namely, I want to address the question of performance since there is not only human-based curation of data, but also *artificial intelligence (AI) based curation*. I want to ask, What is the difference between human- and AI-based Curation ? While comparing the two types of curation, I will also approach certain advantages and disadvantages of both. Hence, the follow up question in place is, Is AI-based Curation the new human-based curation and can be seen as a replacement? When discussing these questions, the intertwined nature of those different curation types will become obvious and also the topic of mixing those two will be touched. In the last part of this paper in section V I want to stress the importance of those critical questions and embed them into the context of the multilayered era we live in. Consequently, some questions for further research will arise and be posed.

II. Definitions

The current era we are in is coining the notion we have of information, knowledge and data. There exists a so-called triangle of knowledge [7] describing how data can become knowledge. Data refers to a kind of representation of observations, objects, or other entities that are employed as evidence of phenomena [6]. Therefore, data can be seen as collections conveying certain information, for example representing a quantity, quality, or specific facts that are providing a basis for meaning that may be interpreted. A datum, the singular form, is according to that a single value in such a collection. In conclusion, one can say that in order to pursue knowledge one has to extract and interpret the information the data is conveying. Due to that, one can say that data presupposes an interpretive base [5], or in other words "data are never neutral" [8]. Next to that, data has an aggregative nature so to say, since they are collected into assortments [8] and are tied to a specific context [2]. Big data however can be described as "multiple and diverse" [8], whereas the features of searching, aggregating and cross-referencing through the large data sets are crucial. The phenomenon of big data has no ability to be about something anymore, since the origins and interpretations become multiple [2].

When we deal with the life cycle of data, it is necessary to highlight the process of data curation. This process is about creating, organizing and maintaining data in a way that they are usable and accessible for users. In other words, curation is a form of management of data to make it useful [1]. Especially, when putting curation into the context of the digital era, curation is a key mechanism of sociality. In terms of social media there is an abundance of information that you in the first place could consume or also produce. Therefore, regarding the social media context, curating is a process of selection which can be seen as an “art” [1] of the curator who is manipulating the data in a way that a specific outcome is reached. This process of manipulating can also be described as a “curational strategy” [1]. Data goes through processes of collecting, where the data is generated and digitized, where for example pictures are taken. Additionally, data also can be classified in a way that specific data units are selected and categorized together. Besides data can be displayed, can be visualized and published on social media platforms based on curational strategies. But also archiving is a specific process data can live through, in the context of social media this means organizing, for example, a collection of pictures and storing them at specific storages. An important difference to make is to distinguish productive from consumptive curation [1]. Productive curation means producing, sharing, and documenting. In a sense, productive curation can be described as the “intersection of self-presentation and privacy maintenance” [1], since it is an active decision based on strategies on what to post and to present to others. What is the user generating, or collecting and how is he displaying, organizing or sharing that? The user can manage his own identity by data curation. Consumptive curation however focuses more on the classifying process where the user has to select information that raised his or her attention. In doing that, the user can create a specific lens through which his world is displayed. All in all, what is the relation between data and curation? The relation is based on the usage of data. Data conveys information which can be interpreted and can be produced to knowledge. In the context of social media where data curation can be used to manage an identity or produce a personal lens of the world, the relation can be described in a way that curation is used as a tool to craft a specific art which underlies social practices and cultural patterns. In this context, data can be seen as material which can be formed into art and curation is the process of actively performing so to say. Curation is the process of transforming data into valuable information which is crafted in a specific subjective, biased way.

III. Thoughts connected to the course

During the course I reflected on the process of data curation. Namely, the assignment number 1 in which we had to create our own dataset was an appropriate task to observe the process of digitization and also of collecting, generating. When creating your own dataset you have to think of several decisions, for example what is the purpose of the dataset? What object do I want to digitize, or transform into data? How do I want to realize that? In a way this assignment helped me develop a critical framework when it comes to data or datasets. Especially, the papers [5] and [2] supported me in developing a critical perspective on the bias, or interpretive base of data, to reflect on the fact that a subject made specific decisions on purpose. Besides, in task number one also classifying and displaying played a key role. One had to classify, categorize and select certain objects that might match or make sense together in a dataset, to fulfill a certain purpose or to convey certain information. Next to that the displaying was also a crucial part since for the documentation one had to visualize the given dataset in a useful way. Moreover, in the assignment number 2 where we built our own taxonomy the process of archiving was trained. As a group, we discussed a lot about how to arrange the dataset according to our main goals or information we wanted to convey. Also organizing a taxonomy can be demanding in terms of coming up with a useful structure to obey that is easily understandable and rich of information and connections. The task of visualizing the taxonomy was a bit challenging, because our taxonomy was quite complex. Showing all the connections was somehow impossible to combine in one visualization. Instead, we decided to do several visualizations representing single facets of the taxonomy. This task really made obvious to me how important visualization is and what impact it can have. You can draw attention to certain features of your object to visualize and decide therefore what is the essence of your object to display. On the other hand, as a viewer of a visualization, it is really important to be aware that probably only a part of the whole object is captured by that and may underlie certain biases or purposes.

Finally, what comes to my mind concerning the course is that the main focuses of the seminar were pretty well embedded into the contents and incitements. For example, the user is a producer and user of data, as well as shaped by it, was a focus that could be followed quite well along the course. The paper [1] stressed the importance of social media in the digital era whereas it also addressed the user as a curator, productive and consumptive. In Particular, it was shown well how data collection or curation of digital data affects social relations and institutions. It was shown that curation is a form of cultural patterns and social practices and is therefore tied very close to the processing of data. The way one user is curating his data may influence if you are interested in that person on social media or not, or if you consume his content for example. Data curation in terms of sharing what was collected can be a part of social practices such as maintaining friendships for example. However, it is really important to keep in mind that social media has an economic background. Curated data can be interpreted as some kind of value. Data means power, the companies who have a lot of data have more material to make predictions on with the help of statistical algorithms, machine learning. They can sell more data to advertisers and get money for that. Big tech companies may have full access to the data whereas universities for example have to pay to get specific data, however not every university may afford it. What is created is some sort of inequality [2] and with it comes some class-based structures. A conclusion out of this is that the group of people who are able to analyze all the data are the most privileged, even more than the people who collected and created that data.

It remains questionable if the notion of producing and curating data will change over time, if users from social media platforms will accept the exploitation of the big tech companies passively, or if there will be a certain revolution of data curation and its value. But also concerning the dimension of law it is creating a need to handle the pre given exploitative structures of capitalist companies. Eventually, also the last focus of the part is quite present, the part of data futures. What are alternatives, what can be realized differently? Something that comes to my mind was the "Datasheets for Datasets" workshop from the seminar. I think it was pretty well presented that the idea of offering some metadata about the datasets is helpful to overcome some biases and to make it more accessible for which purposes a given dataset is suitable or not. This workshop also has shown the dimension of critics in the digital era, how difficult it is to realize a change in this field, because the internet offers such an overload of information and it is hard to draw attention to such an abundance. Another challenge to face is also that there is not one group of people creating datasets, but a mass of people from different countries, with different working experience and background and aims. You cannot contact a group and inform them about a consistent procedure, the connections are highly complex. This showed to me that the digital era still has a lot of open questions and also critics. There are improvement suggestions, there are critical papers concerning the ethical dimensions, or biases or equity. However, there are no real "rules" yet so to say. The digital era seems to me a bit like the wilderness and it is one of the challenges to face in this era how we address this, how we can create a certain structure, rules. Is it the responsibility of the government? Of the Big tech companies having Big data available as well as power?

IV. Human-based and AI-based Curation

Connected to the process of curation it is worth noting that there is human and AI-based curation. Especially motivating for me for this topic were the papers [9] and [10]. In the papers it is highlighted what kind of an impact AI-based curation of pictures has on society, culture and how the ethical dimension is raised in terms of inequality, discrimination and bias. Machine learning based algorithms in the specific field of deep learning are performing super and unsupervised classification of data. Moreover, visual essence is discernible by using statistical methods to look for formal patterns across a collection of labeled images. However, some of the training images from WordNet as shown in [9] are built on the basis of shaky assumptions. The humans who build those dataset for training are transferring some society coined discrimination and bias into the data. The ethical dimension arises in terms of responsibility. When such datasets contain discriminative features, this is serious since "automated interpretation of images is a social and political issue" [9]. If humans google certain discriminative terms specific pictures could show and amplifies prejudice and discrimination in our society. If for example one searches for looser and an overweight person would be matched by that word the societal association of this discrimination is embedded in the AI system and has an impact of the users. Especially children who just get to know what specific terms mean may be coined by such discriminative features that become evident through AI-based systems. There is a need that in future this ethical dimension is handled. Again, this addresses the problem of responsibility. Is it the government who should take legal actions? Is it the responsibility of the one's coding the machine learning algorithms and creating the datasets? Is it the responsibility of the individual user to refuse and report such occurrences?

The idea of human-based curation was explained in section II in detail. AI-based curation by contrast, involves employing an AI-based solution tool to perform a curator's work. In this case, an AI tool completes data curation, thus making it more efficient and faster to access. [10]. AI tools make it easy to work with tons of data. When employing an AI tool, a curator doesn't have to worry about the data's volume and complexity. But, who is curating? The specific curation process is coined and shaped by statistical methods neural networks are based on, or in other words "statistics is immanently a science of classification" [10]. Classification is another word for generalization. Putting certain elements together to create patterns and structures. Additionally, the question is not only who is curating but, also what data is the AI-based system curating? As we saw in the excavating AI paper, often this bias or discrimination exists due to datasets based on shaky assumptions which are performed by humans who label the datasets when it comes to supervised learning mechanisms. There is so to say a mixture of this process, the pre given, human made dataset and the AI-based classification. Also for "seemingly unsupervised models" [10] it is important to notice that they become supervised due to classification work, for example when setting the number of topics, or cleaning data with a human-based understanding of what are meaningful clusters. Also are those clusters interpreted manually [10]. In the paper "Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media" they showed how unsupervised learning classifies text and highlighted that in Natural language processing there is still a high level of human control. Again, also in the unsupervised learning case, a lot of additional human work is done. A question arises, can we call it AI-based curation then? When parts of the collecting and displaying part are done by humans? This is hard to answer, but personally I would suggest that there is still a huge difference between those approaches. Specifically, the so-called blackbox of the evaluation systems from deep learning approaches prove to be less powerful due to the fact that one does not know based on what feature the neural net is categorizing. I would say that in a world with big data and increasing amounts of data, there is a need for automated classification systems. And also I would state that the fully AI-based curating is more in theory than in practice because as was discussed, machine learning approaches still require a lot of human interpretation and support, from datasets, to coding them, to creating clusters manually. However, in favor of comparing those two different ways of curation I will call them AI-based and human curation in view of the fact that the naming is still questionable in terms of representing the essence of the curation. Generally this poses the question whether it would be reasonable to distinguish more in detail, to be more specific where the boundaries in between ai and human-based actions are, that we need more research and more clear boundaries to compare those actions.

The human-based curation has several advantages over AI-based curation. It is able to build context, and to reflect on the given information. The AI-based curation on the other hand simply analyzes on the basis of fixed patterns. The AI-based curation is not able to 'understand' what the given categorizing means and is therefore unable to reflect [11]. Moreover, human curation offers the opportunity of posing new questions whereas AI-based curation simply provides answers, without the ability to question these. AI-based curation serves more of a specific knowledge, whereas human-based curation is able to produce a holistic knowledge, or even wisdom. Moreover, human-based curation is able to detect culture and to make sense of data in terms of the social world. In a sense, human-based curation can recognize a certain impact of the data, and analyze it further, whereas AI-based curation stops at producing output [11].

So, you may ask yourself, Is AI curation the new human curation and can be seen as a replacement? What can be said about that is that, as elaborated human curation has some specific features that cannot be replaced, like recognizing a cultural level of knowledge, or building context. Additionally, it was also shown that AI-based curation still has some human work in it like providing the datasets or deciding manually for specific clusters. I would state that curation cannot be seen as an independent phenomenon, but rather as a useful tool in a world of big data that is still dependent on human decisions and reflection. A human can analyze content, find common themes and display only the relevant information. Even though AI is here to stay, as of now, it cannot replace human curation. Every curator is different as his/her skills and knowledge are different. If in future machine learning systems can accomplish that is questionable but rather unlikely based on the current status or skills of AI-based systems.

V. How are the questions important?

Despite the common mythos that AI and the data it draws on are objectively and scientifically classifying the world, everywhere there is politics, ideology, prejudices, and all of the subjective stuff of history. When we survey the most widely used training sets, we find that this is the rule rather than the exception. It is necessary to spread the awareness of AI-based systems amplifying prejudices and ideologies. Since the current world we live in is dominated by social media and the internet, the individual is highly influenced by the data which is consumed. The posed questions are specifically important because they are pointing to a specific problem of AI-based systems, namely the so-called blackbox and the biased datasets which are combined when performing AI-based curation. The awareness is important that in a big data coined era, such AI-based systems are necessary and that we should find a way to cope with it rather than simply criticizing it. We also need some ideas on how we can integrate those systems in our world without named disadvantages. The questions also point to the dimension of responsibility for the given prejudice and ideologies which are transferred through AI-based systems. Moreover, it was shown that the user of social media when being confronted with AI-based classification should be aware of this. Next to that also for research it is important to define in more detail what exactly does AI-based curation mean? What does mixtures of human and AI-based curation mean and how can those actions be better defined in order to explain the nuances better. Also in terms of a legal dimension this is important, who is responsible for that? Who is the curator, what was curated? Is it the people who created certain datasets or added some clusters manually based on what makes sense? Or is it the programmer of the machine learning algorithm? Is it the social media platform allowing such AI-based classifications? There are more restrictions, laws needed in an 'era of wilderness' so to say.

VI. Pictures title page

Data triangle

References

[1] Jenny L. Davis (2016): Curation: a theoretical treatment, Information, Communication Society, DOI: 10.1080/1369118X.2016.1203972

[2] danah boyd Kate Crawford (2012) CRITICAL QUESTIONS FOR BIG DATA, Information, Communication Society, 15:5, 662-679, DOI: 10.1080/1369118X.2012.678878

[3] Duane Windsor (2020): Ethical Values and Responsibilities of Directors in the Digital Era, DOI: 10.4018/978-1-7998-2011-6.ch005

[4] Jayantha P. Liyanage (2012): Hybrid Intelligence through Business Socialization and Networking: Managing Complexities in the Digital Era, DOI:10.4018/978-1-61350-168-9.ch030

[5] Lisa Gitelman and Virginia Jackson (2013) "Raw data is an oxymoron", Chapter 1. Doi: https://doi.org/10.7551/mitpress/9302.003.0002

[6] C. L. Borgman (2015) Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press. Cambridge MA, DOI:https://doi.org/10.7551/mitpress/9963.001.0001

[7] Ackoff, R. (1989). From data to wisdom. Journal of Applied Systems Analysis, 16, 3-9.

[8] Andrew Iliadis Federica Russo (2016) "Critical data studies: An Introduction". DOI: https://doi.org/10.1177/2053951716674238

[9] Kate Crawford and Trevor Paglen (2020) "Excavating AI. The Politics of Images in Machine Learning Training Sets", DOI:10.1007/s00146-021-01162-8

[10] Anja Bechmann and Geoffrey C Bowker (2019): Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media, DOI: 10.1177/2053951718819569

[11] Radhecka Roy (2018): Human curation in an Ai world. In: https://dokumen.tips/documents/human-curation-in-an-ai-world-a-a-human-curation-ai-worldpdf-storytelling.html?page=1, date: 1.12.2022.