This article is available open access under a CC BY-NC-ND 4.0 license thanks to the generous support from a global network of libraries as part of the Knowledge Unlatched Select initiative.

Afterword – Crafting Data, Crafting Worlds Across Disciplines

Katherine R. Amato and Roberta Raffaetà

This special issue draws from the AAA panel 'Entangling data while entangling disciplines: discussing the future of anthropological collaborations with data scientists'. It deals with experiences of anthropologists who have collaborated with data scientists. To render the panel truly interdisciplinary, the organizers of the AAA panel invited a data scientist as discussant, Katie Amato. Below her comments have been elaborated in the form of a dialogue with anthropologist Roberta Raffaetà. This afterword aims to sketch some paths of reflections about what data scientists think of collaborations with anthropologists.

The dialogue starts by discussing the tension between reductionism and complexity in technoscience. It links to the role of authority and credibility in times of post-truth- This develops in a proposal for an interdisciplinary art of data analysis in which the conventional concept of 'context' (with an example about race and racism) is redefined. The dialogue ends with a discussion on reality and scale.

Katie Amato is a biological anthropologist that studies how microbes influence the biology and health of humans and non-human primates. Her research relies heavily on collecting data describing human and non-human primate environments and lifestyles as well as collecting biological samples such as feces. From these biological samples, her lab generates data describing the composition and function of the associated microbial community, often relying on DNA sequencing and other molecular techniques. These genetic and molecular data must be quality-controlled, organized, and associated with the appropriate environmental, lifestyle, and health data, etc.…

RR: Thanks Katie for been available to experiment with us in this interdisciplinary dialogue between anthropology and data science. During your discussion of the papers presented at the AAA panel 'Entangling data while entangling disciplines: discussing the future of anthropological collaborations with data scientists', you efficaciously captured the beauty but also the ambiguity of data, these being "at the same time very complex, and letting us kind of grasp the complexity of the world, but also oversimplifying it". Many of the papers included in this special issue have emphasized the usefulness but also the simplifications created by data work. Could you please elaborate on that?

KA: Many anthropologists are concerned that data scientists are oversimplifying the processes and relationships that they study, and there often seems to be an assumption that data scientists either do not know or do not care that this is happening. But as someone from this community, I would argue that most data scientists know and care that this is happening. They know that they're not describing every intricacy by creating this data and that the patterns and relationships they are identifying are simplified. However, real-world scenarios and questions are big and difficult to understand, and so we need to look at small parts to try to begin to build understanding. When data scientists simplify, I would argue that they are doing their best to identify what are likely to be important parts of a problem and to generate data that can inform their relationships with each other. This idea came up in every area of study that was covered by the contributors to this special issue. How do we classify what types of farm environments are we looking at? How are clinicians making decisions? How do we describe how musicians playing together? I don't think it's ever lost on the people using data that it's not perfect. They know they are creating simplified models to try to explain complex phenomena.

Nevertheless, I do think some simplifications are more accurate than others, and the extent to which data scientists can recognize this can often be limited by both professional and personal backgrounds.

Anthropology in Action, 30, no. 3 (Winter 2023): 59–65 © The Author(s)
ISSN 0967-201X (Print) ISSN 1752-2285 (Online)
doi:10.3167/aia.2023.300307

As a result, I think it is extremely important for us to identify those limitations to help further advance understanding. An easy place to start with this is by considering selection bias. Who is getting included in the research? Are data scientists aware of who made those decisions and why? What biases are researchers coming in with in terms of the questions they are asking and the methods they are using? I think anthropology becomes extremely important in this context. It can help us take a step back and recognize how and why some of these decisions are being made. Depending on their training, data scientists are not always taking a moment to step back and think about these things. They may know the data are oversimplified, but they may not recognize a key source of bias or a confounding variable because they simply aren't used to considering a problem from that perspective.

RR: In which way do you think it may be beneficial for data scientists to identifying gaps? Could you give us an example?

KA: I think that as data scientists begin to identify gaps and limitations in their models, our understanding of relationships and processes being modelled will become more accurate, even as they continue to be simplifications. Biomedical research that includes race as a variable is a great example. Many data scientists will try to avoid oversimplifying their data or ignoring a potentially confounding variable by recording the race/ethnicity of their participants. There has been a lot of discussion around this practice and its potential pitfalls lately (Benn Torres 2020, Ioannidis, Powe and Yancy 2021, Yudell et al. 2016). First of all, it is often unclear what is meant by race/ethnicity because it isn't defined clearly by the researchers. Secondly, the fact that researchers are recording race implies that they expect it to influence their results, but how and why they expect it to influence their results is rarely explained. Are there underlying genetic differences? Are there lifestyle differences? Geographic confounds? Unfortunately, biomedical studies that find an effect of race often imply that there may be some underlying genetic explanation for biological differences, but we know that race has no genetic basis. Biological differences between races are instead due to social influences on health such as structural racism shaping people's environments. However, since people's experiences with racism can vary even among people that self-identify with the same race, what we should be measuring and recording is not race but racism. In the process of trying to measure racism, it is likely that there will still be an oversimplification of people's individual experiences of racism. Nevertheless, this variable is still likely to be more powerful for research than 'race'. If biomedical studies started to use racism as a variable instead of race, I suspect there would be much clearer results in a wide range of studies.

RR: Thanks Katie, so interesting. What you say, however, also makes me think that 'selection bias' may include not only which people to include in a study design but how scientists choose features in terms of time and space. From my experience of working with microbiome scientists, often these choices mostly rely on previous literature on the topic, customary practices or beliefs and less on into-the wild and historical biosocial observations and analysis of the current features of a specific place or situation. For example, Kuthyar and Reese (2021) have recently dealt with this issue pointing to the need to go beyond standard categorizations as either industrialized or traditional and consider that intermediate lifestyles, such as those rep resented in intermediate populations such as those living in urban slums in developing megacities or agricultural communities beyond pure subsistence. What do you think about this?

KA: I think that is a valid observation. In some ways I think the tendency to rely on conventional categories emerges because many scientists are trained that way, particularly in fields such as molecular biology and microbiology. They use existing literature and data as a foundation from which to generate new ideas and questions. This approach is completely valid. In a way, these scientists are going out into their 'wild' of what is known already about a system. However, as you point out, biases exist in what has been studied already and how. These biases are often a result of influences from closely related fields and processes that are better studied as well as a tendency to rely on certain methods and approaches to generate data. Taking a more ethnographic approach to collecting background information during study design could complement and enrich these practices.

It is fascinating to me to think about how the field of microbiome research has developed over the past couple decades, and it seems relevant to your point. Sudden access to technology that could help us describe communities we had never studied before led to an explosion of exploratory research. As a result, the field was initially full of studies using a 'let's go see what we can find' approach that was less tethered by prior knowledge of the system since that knowledge mostly did not exist. At the same time, however, the people doing this research often came from molecular biology and microbiology backgrounds, which means –as you said—that there were biases in how that exploration was being done. Most host-associated microbiome research is human-centric with implications for health. There are standards of evidence and data that emerged early on that were very much driven by traditional microbiology frameworks such as Koch's postulates, and because many existing microbiologists at the time had been working on pathogens, even data interpretation had biases towards infectious disease biology. It has been fun to come into this field as someone with more of an ecology and zoology background and have discussions with people from microbiology and pathology about what we would and would not expect in host-microbe interactions. Not surprisingly, our studies end up looking fairly different, but I also think we are learning from each other and enriching our understanding of what we are studying. And this is just within 'science.' If we could move beyond different scientific perspectives and approaches and begin to include historical and biosocial perspectives as well, I think it would only benefit the knowledge building process.

RR: Thanks Katie. In your discussion you also raised another important issue which influences how data scientist work with data: scientists' credibility in times of post-truth and how it creates a certain reluctance from the part of scientists to make more nuances their data work.

KA: Yes. Unfortunately, I do think that there are probably scenarios in which data scientists downplay the gaps and limitations of their data as a result of the social and cultural landscape that they find themselves in. For example, in the US recently, there has been widespread doubt about whether we can trust experts or not. A large part of the population has decided that because everyone has biases, no one can be believed, including scientists. Along with this mistrust, there is a misunderstanding of how science works. Instead of acknowledging that scientists are always gathering more data and improving their assessment of whatever it is they are studying, there is a belief that scientists should come up with the correct answer the first time they address a question. And if they change their answer, or they do not sound sure about their answer when they are explaining it, then they can't be believed. As a result, there's cultural pressure put on data scientists to be over-confident. Maybe they know their answer isn't perfect, but they're being pressured to give a perfect answer so that they can be credible in the social and cultural spaces that they're inhabiting. I think anthropology can help identify how those processes are working, and perhaps create more space or better language for data scientists to be able to really engage with this phenomenon and say,"Yeah, it's not perfect, but here's the best we can do with what we know right now." How can we communicate uncertainty or imperfection to the people that we're working with in a way that it still is credible?

RR: Thanks Katie, yes, indeed what you propose is a process of transformation that includes not only scientists but also 'the public', always reminding that scientists too are part of it (Hinchliffe et al. 2018). And certainly, anthropology can help with this, especially if in alliance with scientists. What do you propose is pretty much in line with the anthropological program of relativizing truth. There is a long, and conflicting, history related to that aim. In recent year, the encounter between anthropology and science and technology studies has strengthened an approach that understands truth that is situated and therefore contingent, yet holding ontologically. Various authors[1] within the so-called 'ontological turn' propose doing away with the distinction between interpretation (constructionism) and reality (positivism), nature and culture, human and nonhuman, material and information, and similar opposites. This approach is an attempt to go beyond dualisms, categories and identities, but it also recognises the equal value of different sociotechnical ways of making sense of reality. Thus, an alliance between data scientists and anthropologists would be of advantage for both, because both are pursuing a similar epistemological endeavour. Which are, in your view, the main factors that limit this alliance to happen on a broad scale?

KA: Two main factors come to mind for me. First, there are differences in language and vocabulary. Each field has its particular way of communicating, and when people with training in different fields try to work together there are often communication barriers that must be overcome. Even when two parties are saying the same thing, they may use different words to describe it, and that can lead to misunderstanding. The more we can have the kinds of conversations that you and I are having, though, the easier it will be to avoid this pitfall. Even if we don't know where the language differences are, being aware that they exist will allow us to detect them and overcome them more quickly.

The other key factor could be broadly described as variation in career deliverables. Each field has embedded practices with regard to knowledge generation and dissemination. Anyone that devotes time and resources to a project will want to disseminate their findings. However, dissemination may look very different for an anthropologist versus a data scientist. What happens then? Does the project support the creation of multiple products that can be disseminated in different ways? If not, is it professionally 'worth it' for data scientists to be publishing in anthropology journals and vice versa? Evaluation systems for jobs and promotions are differ across fields and do not incentivize interdisciplinarity. The university administration might not care if the data scientist created a rich collaboration with an anthropologist if it didn't result in grant money or publications. It also might not care if the anthropologist had their name on a paper in a top science journal if they haven't finished their book. As we begin to work more in these interdisciplinary teams, we have to be sure to think carefully about why the data scientist wants to work with the anthropologist and vice versa. For each project proposed, can they both get what they need intellectually/professionally from the interaction? We cannot assume it is the same goal for both.

RR: Thanks Katie. I totally agree but I think the new European funding scheme for research – Horizon Europe – is going into the direction of supporting that kind of interdisciplinary work. Yet, that does not immediately solve the problem of national rules of promotion and evaluation that still heavily rely on disciplinary differentiations. I think disciplinary specificity is important and should not be cancelled, but interdisciplinarity should be a feature to be added to formal evaluations.

In your discussion, however, you emphasized some zones of tensions also staying within one's discipline. You said that there may be some anxiety "where there's a breakdown between who's generating data and who's analyzing, using it or seeing it". You suggested that anthropologists may help data scientists in negotiating these transactions. Can you tell us something more about this?

KA: I think there's a lot of interesting things to be considered here in terms of how the data is being interpreted. I think it is useful to draw on what Douglas-Jones and colleagues (Douglas-Jones, Walford and Seaver 2021) have talked about - artful revelation - the idea that we're creating reality from the data. However, as part of this reality, we have the biases that Jennifer Jo Thompson was talking about, or kind of these tacit, practical reasonings that Ritwik Banerji was talking about. I would like to challenge data scientists to actually draw those biases and reasonings out and be more explicit about what perspectives we are using to make these interpretations. I think a lot of this process gets glossed over in data science, and it could be improved by increased interactions with anthropologists. I think this came up also in other contributions, also, with who gets to participate in different parts of a study on the researcher side. For example, do the data scientists actually get to go see the patients whose data they are analyzing? I think there's a real danger when we start to have those breakdowns in terms of big teams working on the same project, but everyone not having the same information, particularly qualitative information. In our practices of data science in my lab, we're trying to reduce these types of situations by encouraging what I would describe as interdisciplinary practices. For example, even students that want to focus only on data analysis are strongly encouraged to go to the field and see how the data are collected so that they're not analyzing the data in a vacuum. Without this type of experience, data scientists can end up sitting there and trying to interpret one odd point on a graph, knowing that important information is missing, but also having no idea what we should even be asking the people that collected it. To try to avoid this scenario, I encourage students and trainees actually go and experience the place and the participants so they have an appreciation of the qualitative environment and therefore all of the variables that might be at play, even if they weren't measured.

Even beyond a single research group, these challenges have become particularly daunting with the more widespread practice of open data availability. Different researchers can now go to public databases and simply download data to use without having to interact with the person that collected the samples or did the lab work. This practice can lead to similar gaps in interpretation and understanding. In my lab group, we try to address this by continuing to collaborate with the people that generated the data so that we don't completely misinterpret it. This practice is not always possible, but we try to prioritize it.

Finally, I will just mention that even without all of the added qualitative perspective the practices described above can provide, data scientists need to be really explicit about how much variation is explained by our data and where the error is. We need to be able to say "Here's the pattern we see, but here's how far it can go in terms of explaining the situation." Or "Here's how important we think this part of the pattern is". If we can make that information more explicit, it also by default leaves room for the other variables and complexities that were not captured. I think all of these approaches are attempts at trying to ensure that individual data scientists understand the extent to which they are oversimplifying the data. The more we can promote approaches that do this, the better our ability to trust the data interpretations in my opinion.

RR: Thanks Katie, so true. And making explicit what is missing in an analysis also may help in defining the terms used. For example, we go back to the issue of labelling populations as 'westernized' vs 'non-westernized'. Categorisations are "boundary objects" (Bowker et al. 2016), or rather, forms of representing reality often used by different scientific communities. But each community has its own interpretation that it takes for granted and naturalises. It is important to go beyond and across these naturalisations because ways of categorising have tangible effects. The question of how to categorise and hence name a population is a fundamental epistemological and ethical question because naming systems are not simply linked to the identity of individuals and groups, but to their relation to other groups and systems, as you rightly pointed out in your discussion about race/racism. At the same time, science need to be reductionistic to some extent to work. A pragmatic and functional approach that makes explicit the scientific priorities of the research and, at the same time, circumscribes the categorisation choices within broader socio-cultural and political considerations (which need being consistent with the scientific priorities) may be a good way to combine the two apparently contradictory requirements of complexity and reductionism in science, so stressing the importance of "looking both ways" (Powell and Dupré 2009, 63). What do you think about this?

KK: I completely agree. In my language I would say that we need clear operational definitions, but these definitions should also be accompanied by acknowledgements of potential shortcomings. This practice would not add many words to a paper, which is sometimes a concern, and it would greatly increase clarity. It could be as brief as 'Because we are interested in the effect of X, we categorized our samples into groups 1 and 2 that exhibit [these specific] differences in X. However, it is important to note that factors A, B, and C also vary within and across groups 1 and 2."

RR: Also: above you rose a very important issue proposing an interdisciplinary method for data analysis. In anthropology, the issue of context is key. Yet, the concept of 'context' is misleading because it gives the wrong impression that exists a background on which things happens (Seaver 2015). What is conventionally called 'context', instead, emerges with and through the encounter with what or who act in it. The context would not exist without those encounters. In other terms, to know about the context it is not to add up to the knowledge that data can provide us, but context it is part and parcel of data, not something that can be eliminated. But the issue is that scientists cannot indulge in not considering the context on the assumption that it is already embedded in the data. This is a typical stance in scientists working with machine learning because "the data being modeled are themselves agents capable of modelling" (Kockelman 2020, 319). To know about the 'context', however, is a way not only to gain tools to make sense of data and enhance analysis but, especially, contextual knowledge broadens our range of possibilities for interacting with sociotechnical models (Kockelman 2020).

To put data into context, necessarily rises the issue of scale because contexts and data are typically multi-scalar. What can you tell us about scale?

KA: Scale has been really important in our data. We have found that there are different factors that influence the microbiome, but they become more or less important depending on the scale at which we are examining the data. For example, my group and others have found that diet has a big effect on the types of microbes animals have in their gut. However, this seems to be most true when we are considering a single animal species. If I am only looking at data from chimpanzees, diet will explain a large amount of the variation in microbial communities between different chimpanzees. However, if I then compare chimpanzees to humans, diet variation between individual chimpanzees and humans appears to be less important. The animal-species level differences between the human and chimpanzee gut microbiomes are much greater than the diet-induced differences among individuals of each species. As a result, my message about how important individual variation in diet is for shaping the microbiome is liable to change based on the scale I am using. Again, the more that data scientists can better communicate these intricacies, the better our understanding of a given process or relationship is likely to be. I think anthropologists have language and practices that can help with these issues, as you have mentioned. It very much is an issue of scale and 'context' being a part of the data itself. If we didn't have data from multiple animal species, we would have no idea that the importance of diet changed with scale.

RR: Thanks Katie. There is indeed a rich literature in anthropology and geography on scale that urges to take into account what happens every time a change of scale occurs (Tsing 2012) because the micro does not necessarily contains the macro, and viceversa (Hecht 2018, Irvine 2016) but scales are emergent, relational and performative (Clark 2012), yet real and pragmatic (Carr and Lempert 2016). In which way, specifically, do you think anthropologists may help scientists to think across scales?

KA: I think anthropologists are particularly good at exploring variation while data scientists are good at finding unifying principles. I think both fields acknowledge that shared patterns and unique patterns exist, but my feeling is that data scientists tend to start by looking for the commonalities before identifying deviations from them while anthropologists tend to start by looking at the unique patterns before finding commonalities. As a result, I think anthropologists can be helpful to data scientists by challenging them to see what happens if they add more data or collect the same data in another place or at another time. Essentially, I think anthropologists can help data scientists to have more 'encounters' with the data and to identify how this changes the 'reality' that the data are describing.

RR: Very interesting, Katie! Your mention to the 'reality' of data suggests me another question. So far, indeed, we have mainly dealt with the kinds of epistemological contribution that anthropology can provide to data science. But in your discussion's concluding remarks, you affirmed that "anthropology could influence the extent to which data becomes reality". I agree with that, proceeding from the idea that the epistemological (how we know things) and the ontological level (what things are) are interdependent. Could you elaborate on your statement?

KA: I think this relates back to my last point. Instead of accepting the idea that data are collected once and we should accept whatever interpretation is made from those data as reality, we should be thinking critically about how and when to include additional data as well as the shortcomings of the data generation process. By exposing potential avenues of bias and oversimplification, providing language to describe the relativity of interpretations, and prioritizing interdisciplinary explorations of the world, I think anthropology can help us more accurately understand the insight that data science is providing. Essentially, if we can further expose data science tools as tools and all of the models produced by data scientists as necessarily flawed at some level, hopefully we can allow more real realities to be created. We can acknowledge that a model produced by a single dataset is helpful for understanding a situation but also acknowledge that it does not completely describe the situation. I really liked what Adrianne Mannov was talking about with borrowing parts from each other to become a whole. And then what Ritwick Banerji was saying about kind of the accuracy of the system, and are we actually reflecting reality? And is that what we want? Or do we want to improve upon it? If we can help data science to embrace the idea that no one model will give us the answers about any single process or relationship, it will encourage more models to be created. This will simultaneously reduce the incorrect assumption that an oversimplified single model is sufficient for describing reality and increase our propensity to generate multiple models with different approaches that together can improve our understanding of reality.

RR: thanks Katie. I agree and I think that interdisciplinary work between data scientists and anthropologist may encourage a better design of research, the formulation of good research questions, to reveal the limit of the customary focus on specific set of variables, people, time or places, exposing hidden features and the un/intended consequences of one choice vs another. The integration of socio-cultural anthropology and in-depth, into-the-wild interdisciplinary analysis is not intended to be normative on how to conduct science, nor is aimed at offering easy solutions for current biosocial challenges. Rather, it serves to understand the ambiguities and compromises between abstract goals and (un)intended effects, and to expose the relationships that science entertains with and against (in)equities. This, as you said, would strengthen the 'reality' of our research.

Acknowldgements

KRA receives funding as a Fellow in the CIFAR 'Humans and the Microbiome' program. RR has written this Afterword as part of a project that has received funding from the European Union's Horizon 2020 Research and Innovation Programme (GA n. 949742 ERC-HealthXCross).

Notes

1. The 'ontological turn' is a wide-ranging and diversified approach in anthropology difficult to do justice to in a footnote. It emerges from different fields as indigenous studies, science and technology studies and material culture.

References

Benn Torres, Jada, 'Anthropological Perspectives on Genomic Data, Genetic Ancestry, and Race', *American Journal of Physical Anthropology*, 171/S70 (2020), 74–86

Bowker, G.C., Timmermans, S., Clarke, A.E., and Balka, E., *Boundary Objects and Beyond: Working with Leigh Star* (Cambridge, Massachussets and London, 2016) <https://books.google.it/books?id=UWOkCwAAQBAJ>

Carr, E. Summerson, and Lempert, Michael, *Scale: Discourse and Dimensions of Social Life* (2016)

Clark, Timothy, 'Derangements of Scale', in *Elemorphosis: Theory in the Era of Climate Change*, ed. by T. Cohen (Ann Arbor, Mich., 2012), 148–66

Douglas-Jones, Rachel, Walford, Antonia, and Seaver, Nick, 'Introduction: Towards an Anthropology of Data', *Journal of the Royal Anthropological Institute*, 27/S1 (2021), 9–25

Hecht, Gabrielle, 'Interscalar Vehicles for an African Anthropocene: On Waste, Temporality, and Violence', *Cultural Anthropology*, 33/1 (2018), 109–41

Hinchliffe, Stephen, Jackson, Mark A., Wyatt, Katrina, Barlow, Anne E., Barreto, Manuela, Clare, Linda, et al., 'Healthy Publics: Enabling Cultures and Environments for Health', *Palgrave Communications*, 4/57 (2018) <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5978671/> [accessed 24 December 2022]

Ioannidis, John P. A., Powe, Neil R., and Yancy, Clyde, 'Recalibrating the Use of Race in Medical Research', *JAMA*, 325/7 (2021), 623–24

Irvine, Judith, 'Going Upscale: Scales and Scale-Climbing as Ideological Projects', in *Scale: Discourse and Dimensions of Social Life*, ed. by E. Summerson Carr and M. Lempert (Oakland, 2016), 213–31

Kockelman, Paul, 'The Epistemic and Performative Dynamics of Machine Learning Praxis', *Signs and Society*, 8/2 (2020), 319–55

Kuthyar, Sahana, and Reese, Aspen T., 'Variation in Microbial Exposure at the Human-Animal Interface and the Implications for Microbiome-Mediated Health Outcome', *MSystems*, 6/4 (2021), e00567-21

Powell, A., and Dupré, J., 'From Molecules to Systems: The Importance of Looking Both Ways', *Studies in History and Philosophy of Biological and Biomedical Sciences*, 40 (2009), 54–64

Seaver, Nick, 'The Nice Thing about Context Is That Everyone Has It', *Media, Culture & Society*, 37/7 (2015), 1101–9

Tsing, A.L., 'On Nonscalability: The Living World Is Not Amenable to Precision-Nested Scales', *Common Knowledge*, 18/3 (2012), 505–24

Yudell, Michael, Roberts, Dorothy, DeSalle, Rob, and Tishkoff, Sarah, 'Taking Race out of Human Genetics', *Science*, 351/6273 (2016), 564–65