Making Arguments with Data

Savic, Selena; Martins, Yann Patrick

Empfohlene Zitierung / Suggested Citation:
Savic, S., & Martins, Y. P. (2023). Making Arguments with Data. In B. Herlo, & D. Irrgang (Eds.), *Proceedings of the Weizenbaum Conference 2022: Practicing Sovereignty - Interventions for Open Digital Futures* (pp. 199-208). Berlin: Weizenbaum Institute for the Networked Society - The German Internet Institute. https://doi.org/10.34669/wi.cp/4.19

Proceedings of the Weizenbaum Conference 2022:

Practicing Sovereignty. Interventions for Open Digital Futures

MAKING ARGUMENTS WITH DATA

Savic, Selena
FHNW Academy of Art and Design
Basel, Switzerland
selena.savic@fhnw.ch

Martins, Yann Patrick
FHNW Academy of Art and Design
Basel, Switzerland
yannpatrick.martins@fhnw.ch

KEYWORDS

visual data study, situated knowledge, data observatories, machine learning, correlationism, critique from within

DOI: 10.34669/wi.cp/4.19

ABSTRACT

Whether we are discussing measures in order to "flatten the curve" in a pandemic or what to wear given the most recent weather forecast, we base arguments on patterns observed in data. This article presents an approach to practicing ethics when working with large datasets and designing data representations. We programmed and used web-based interfaces to sort, organize, and explore a community-run archive of radio signals. Inspired by feminist critique of technoscience and recent problematizations of digital literacy, we argue that one can navigate machine learning models in a multi-narrative manner. We hold that the main challenge to sovereignty comes from lingering forms of colonialism and extractive relationships that easily move in and out of the digital domain. Countering both narratives of techno-optimism and the universalizing critique of technology, we discuss an approach to data and networks that enables a situated critique of datafication and correlationism from within.

1 INTRODUCTION

The outputs of machine learning algorithms trained on large datasets (often referred to as “big data”) play an increasingly important role in decisions that concern personal as well as global, socio-political and economic choices. The patterns and trends observed in machine learning models are taken as sources of truth and reason in recruitment and admission processes. They guide policymakers in deciding on measures to take in the pandemic, and they help individuals decide whether to purchase an item or not. Nevertheless, we have very limited access to examine the datasets that inform such decisions. Tools and frameworks such as Google’s Colab[102], OpenAI’s GPT3,[103] or design-specialized RunwayML[104], come with pretrained models and assumptions of correlation between data points. Professionals outside of computer science field work with these frameworks to quickly prototype experimental and innovative projects that are informed by machine learning (ML) and artificial intelligence (AI) tools.

Artistic practice has repeatedly demonstrated how hard it is to counter the assumptions and biases that permeate through automated training processes on datasets. For example, in a recent experimental theatre piece by artist Simon Senn and developer Tammara Leites titled *dSimon,*[105] an artificial personality was performed as a conversant, artistic advisor and as a stand-in for Elon Musk and Simon Senn himself. The dramatic unfolding of inappropriate behavior by the *dSimon* conversation agent, trained on Simon Senn’s personal data using the GPT-3 artificial intelligence engine, engaged the audience as witnesses to the bizarre and unsettling propositions. The imaginary of neutrality in vast collections of internet-based text is quickly dispelled, revealing the inherent sociality of anyone’s or anything’s ability to understand and compose language.

This article combines the technical and artistic perspectives on the bias, and other forms of structural inequality in applications of machine learning models, informed by critical data studies and the critique of contemporary aspirations to objectivity in machine learning applications. The central argument engages the critique of scientific aspiration to universal objectivity most notably addressed by Donna Haraway (1988, 2016) and focuses on the critique of contemporary aspirations to objectivity in working with data, particularly in terms of information representation. This points to the need to envision different ways of working with datasets and machine learning models: These

[102] Colaboratory: browser-based machine learning environment, funded by Google; visit https://colab.research.google.com/ [accessed 15 February 2022].

[103] GPT-3 neural network machine, funded by Microsoft and Elon Musk; https://openai.com/ [accessed 15 February 2022].

[104] Runway ML, machine learning platform for visual tasks; https://runwayml.com/about/ [accessed 15 February 2022]

[105] More information on the performance of the *dSimon* theatre piece in Vidy theatre in Lausanne, in December 2021 is available at: https://vidy.ch/en/dsimon-0 [accessed 15 February 2022].

should entail enabling people to formulate arguments based on relations they actively discover in the data and trained models. A machine learning model is the output of a machine learning algorithm run on a specific dataset, which establishes data structures and relationships that can be applied to further datasets to infer similarities and predictions. Because such models are dependent on the training process and datasets, they tend to re-encode pre-existing determinisms and beliefs. In her work on race and technology, sociologist Ruha Benjamin identified this as engineered inequity and default discrimination (Benjamin, 2019). Furthermore, as computer scientist Cathy O’Neil has observed (O’Neil, 2016), decisions to take correlations between data on, for example, employment histories and addresses at face value, is at the root of the discriminatory operations of algorithms. In order to change this, we argue that we have to start from the dataset and reimagine the expectations of truth and reason from training processes and trained models.

Seeing things in data has historically been of interest in art and in engineering. The aspiration to make visible and public that which is measured and documented can be traced back to the 19th century, when it was first used to denote making visible the information that was not actually present at sight.[106] Data representation is rendered ever more accessible and efficient with the use of information technologies, and with this grows the responsibility to maintain the specificity and situatedness of the assumptions and inferences one makes when working with datasets. We believe that one can learn to do this with a carefully crafted digital tool that makes it possible to navigate datasets in experimental, non-essentialist ways, in order to practice a critique of datafication and correlationism from within. Based on the experiences with SNSF-funded research project *Radio Explorations*, discussed in more detail below, we expect that this way of working with data can result in meaningful arguments, engagements, and stories.

2 Resisting Colonial Relations in Machine Learning Processes

The current landscape of machine learning [ML] includes a large number of tools that are becoming increasingly accessible for people without a computer science background or any coding skills. These tools run on remote servers, as the computing power they require to solve the complex statistical analysis cannot be achieved with regular laptop or PCs.

Tools such as Runway ML provide artists and designers with prebuilt models and a pay-as-you-go system to deploy heavy computing ML on remote servers. For creative practitioners, the software

[106] The nonavailability to sight is mentioned in the entry on Visualization at the Oxford English Dictionary https://www.oed.com/view/Entry/224009. For a good historical overview of cultural importance of data visualizations, see Orit Halpern’s book *Beautiful Data* (Halpern, 2014).

offers access to several basic ML models: text generation, image synthesis, and object detection. The result displays predictions by their trained algorithms based on an input by the user but without showing either the data nor the process of the prediction. Such approaches could be called “arboreal,” to borrow the term from Deleuze and Guattari’s *rhizomatic theory* (1976). It preserves a tree-like hierarchical conception of knowledge and information with discreet categorization. This confronts the user with a tool that does not grant access to the underlying technology of ML such as statistical analysis, data clustering, and prediction. By refusing access, such tools reproduce a colonial-like relationship of entitlement: Resources, such as computational power and algorithms, are claimed by those who operate them in their best self-interest, simultaneously organizing and extracting the work of their nomadic[107] users.

Scholars in social studies of science and technology have addressed the problems that arise with the use of pretrained ML algorithms as decision-making and forecast tools. These models tend to reproduce biases encoded in the data they are trained on. Such biases have already made their presence felt in automated decision making, which tends to exhibit racial and gender preferences in which job candidates to select, who to admit to a college program, who to incarcerate or grant parole to, or whom to give loan approval.

To counter such biases, US-based artist and researcher Caroline Sinders has led many workshops to create feminist datasets[108]. Data collection informed by intersectional feminist practices aspires to mitigate the effect of biases in ML algorithms by critically engaging in the data collection process (Sinders, 2020). Sinders’ workshops invited the public to explore the meaning of data and its use for protest and social justice. In a related gesture, Crag Dalton and Jim Thatcher called for counter data actions (Dalton & Thatcher, 2014). Dalton and Thatcher offered provocations to the regime of “big” data that recognized its situatedness and the risk of technological determinism and challenged the notion of data being “raw.” While current software for ML algorithms often lacks access to the data they are built upon, critical approaches to data collection in academic settings, or workshops within festivals and seminars, promote a discursive approach to the topic but lack a more technical approach.

To work with data means to take a position and to formulate a clear goal. Even if correctly translated as “given” from the original Latin term, data is not simply given and is always collected with certain logics of measurement and observation. Data and analysis never speak for themselves, as anticolonial pollution scholar Max Liboiron has poignantly illustrated (Liboiron, 2021). The

[107] Nomadic is used here to stress the non-settled status of online platforms users, who come and go, register, and depart; at the same time, the problem of user uprootedness resonates with Rossi Braidotti’s nomadic theory, which addresses nomadic subjects resisting “deterritorialization” in Deleuzian terms (Braidotti, 2011).

[108] For an overview on Caroline Sanders’s work, see: https://carolinesinders.com/feminist-data-set/ [Accessed 15 February 2022].

presumption of unproblematic and unaccountable (a special way of saying objective) data collection reproduces colonial relations to resources and reality. Liboiron also emphasized the importance of the care for the subject of critique. We therefore search for ways to develop and work with a digital tool that encourage critical engagement with data, involve formulating the questions that one wants answered prior to observing patterns in data, and clearly expressing one's position in regard to the question. We developed practical approaches to dataset making and interpretations of machine learning models, starting from the aforementioned archive of radio signals. With this, we hope to contribute a clear example for working with large dataset and machine learning technologies in an informed way that promotes participation and intentionality.

3 Data Observations, Projections, and Comparisons

What patterns can we observe in data with our eyes? Our eyes provide us with an embodied, finite point of view (Haraway, 1988). Such point of view embodies limitations that are interpreted as polluting or disqualifying bias from a universalist, objective position. But a universally objective position implies having a way of being everywhere equally, the so-called "god's trick," which carries with it a denial of responsibility, to paraphrase Haraway. Our work with datasets and data representation offers a refreshed reading of Haraway's insistence on the importance and persistence of vision, in the face of the visualization of digital data and matters of representing data objectively.

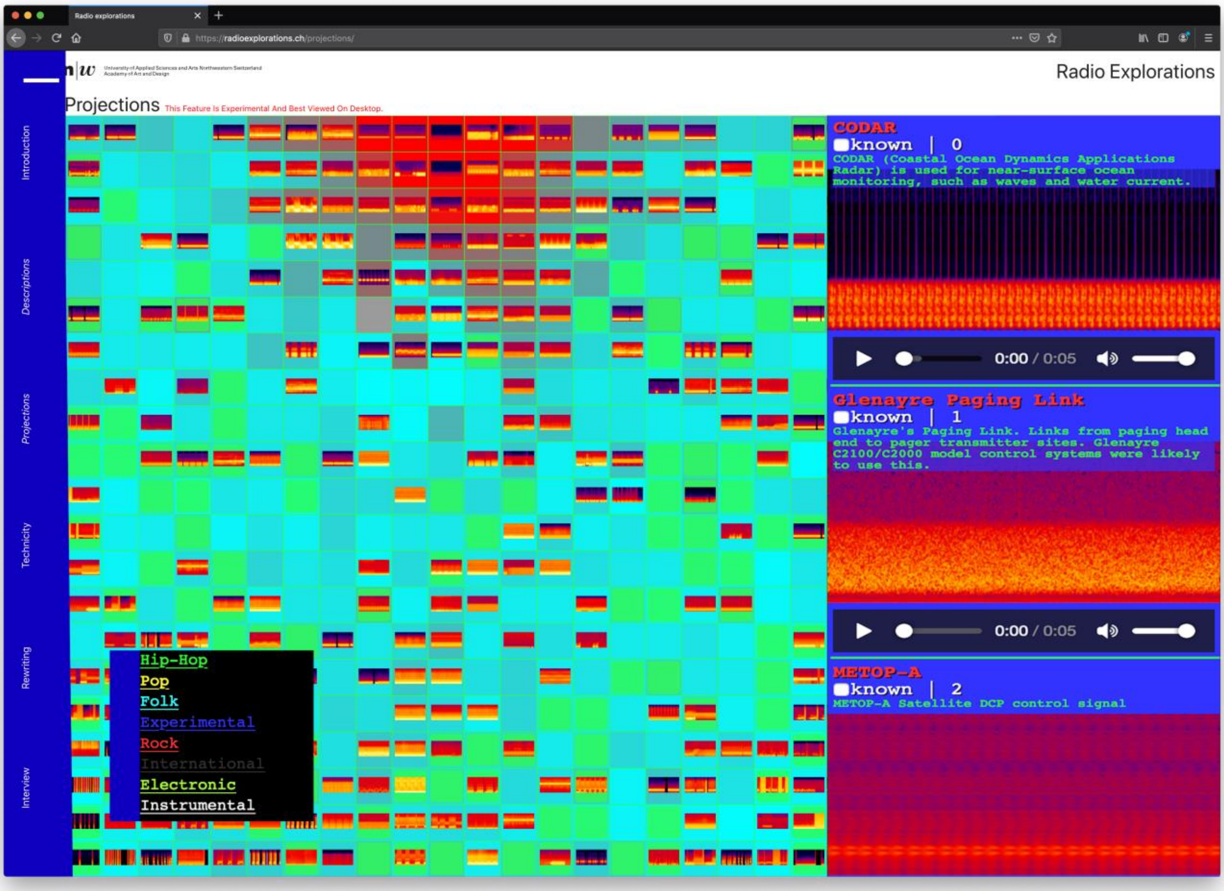
In the *Radio Explorations* project, we designed data observatories as intuitive tools for orienting and navigating. The principal aim was to develop and practice techniques for working with digital data in a way that is ethically sensitive to biases and universalism and that highlights material and symbolic connections with the world. We developed and used a tool that enabled comparisons between patterns in datasets. Such practices foster a unique relationship between the data (given), the method of comparison and the questions one brings to the data.

For example, we created a dataset from the digital archive of radio signals by focusing on specific aspects of these situated recordings of radio transmissions. We computed features such as noisiness or the probability of silence in samples of radio signals found in the database. We then compared the measurement of similarity across signals—as established by a machine learning algorithm called Self-Organizing Map[109]—to those in other, not directly related datasets, such as a Free Music Archive (FMA) dataset for music analysis. By looking at music and radio signals from a comparable point of abstraction, we created a shared landscape of properties whereby data is

[109] SOM is an unsupervised machine learning technique introduced in the 1980s by a Finnish computer scientist Teuvo Kohonen (Kohonen, 1982). It is known for its ability to classify data in an intuitive manner, emergent from the data.

organized according to the conditions of the comparison. In the process, it becomes important how radio signal samples are placed next to each other: A direct similarity between radio signals on the map should reflect their likeness in an aspect that is shared with audible information on music. These comparisons opened up new readings of relationships that can be established across datasets and that refuse to lend themselves to causal interpretations and superficial correlations. While certain signals are similar to other signals in terms of protocols or applications (military, satellite communication, etc.), the setup described here makes it possible to disregard the instrumental qualities of telecommunications and focus on the way digital data can be articulated in its own terms. This means that digital data regarding radio signals is comparable to data on musical genres and that a certain inherent property of data can emerge from the comparison. Radio is therefore not understood in terms of its capacity to transmit messages, which recalls the problematic assumption of access and use of electromagnetic waves as a resource, as opposed to the capacity to conceptualize radio signals in terms of the digital traces they leave and how they interact with recording equipment—which is a perspective we develop in this paper.

The visual aspect of comparison and navigation is important. For example, Figure 1 illustrates the organization of the two previously mentioned datasets—that is, radio signals juxtaposed over musical genres. The visual qualities of radio signal spectrograms facilitate taking a distance from an instrumental perspective on radio signals and their usual categorization according to application or frequency. Signals are represented here in terms of abstract visual patterns that preserve partial qualities related to these instrumental concerns. Visual interpretation goes both ways: It is helpful to compare signals but also to perceive how the tool itself operates and question whether the connections proposed by algorithms actually make sense.



CODAR
□known | 0
CODAR (Coastal Ocean Dynamics Applications Radar) is used for near-surface ocean monitoring, such as waves and water current.

Glenayre Paging Link
□known | 1
Glenayre's Paging Link. Links from paging head end to pager transmitter sites. Glenayre C2100/C2000 model control systems were likely to use this.

METOP-A
□known | 2
METOP-A Satellite DCP control signal

Figure 5. Radio Explorations. Signals are "projected" onto a preorganized map of musical samples, labelled according to the genre (overlay, bottom left). Each genre "highlights" some cells among which certain radio signals can be found. Highlighted here is the "Hip-Hop" genre.

We practiced working with different datasets and the idea of remaining open to the relationships in data, to the interpretability of statistics, and to data clusters. By selecting the data we worked with, and choosing a relationship we wanted to explore, we made it possible to search beyond correlations and establish meaningful comparisons across datasets so that people can make a visual/verbal argument that relates to their question and not to a "neutral" pattern in the data.

4 CONCLUSIONS

The recent rise of data driven technologies, like classifiers and recommender systems, has drawn attention to the problem of biases within data, and it has prompted vocal criticism of automated machine-learning-powered technologies. Nevertheless, such criticism often precludes alternative ways to use technologies that can be steered towards new modes of expression and argumentation. We hold that the main challenge for digital sovereignty and active participation in digital transformations actually comes from lingering forms of colonialism and extractive relationships that easily move in and out of the digital domain. With this paper, we want to invite the reader to rethink ways of engaging with data so that people can take the space and structure to assert their own questions in relation to data.

We developed a technical framework that comprised a digital tool for data processing and analysis within (redacted) project and used it to explore multi-threaded narratives of music and telecommunication, of power and efficiency, encoded in datasets we worked with. We insisted on the visual organizing aspect of this practice. This is not meant as a call to improve ways of visualizing data but rather to innovate on ways to interpret and work with data using visual and other means to represent relationships always previously established in code (i.e., machine learning algorithms). Combining the concern for the importance and persistence of vision and its access to complex relations in the data, with the concern for digital sovereignty expressed as a resistance to colonial relations that haunt digital tools and knowledge of technical artefacts, we suggest paying attention to data in a carefully critical way.

5 ACKNOWLEDGMENTS

The *Radio Explorations* research project was generously supported by the SNSF-Spark funding grant number 190310. We are grateful to our research workshops guests: Carl Colena (SIGID wiki), Miro Roman (ETHZ), Simone Conforti (IRCAM), Sarah Grant (Kunsthochschule Kassel), and Roberto Bottazzi (The Bartlett) for their invaluable input. Special gratitude goes to Miro Roman for numerous informative discussions on working with SOM, as well as to Carl Colena for his support and discussions on radio signals beyond workshops and interviews.

6 REFERENCES

1. Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity.
2. Braidotti, R. (2011). *Nomadic Subjects: Embodiment and Sexual Difference in Contemporary Feminist Theory, Second Edition*. Columbia University Press.
3. Dalton, C., & Thatcher, J. (2014, May 12). What Does A Critical Data Studies Look Like, And Why Do We Care? *Society and Space*. https://www.societyandspace.org/articles/what-does-a-critical-data-studies-look-like-and-why-do-we-care
4. Deleuze, G., & Guattari, F. (1976). *Rhizome: Introduction*. Éditions de Minuit.
5. Halpern, O. (2014). *Beautiful data: A history of vision and reason since 1945*. Duke University Press.
6. Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, *14*(3), 575. https://doi.org/10.2307/3178066
7. Haraway, D. (2016). *Staying with the trouble: Making kin in the Chthulucene*. Duke University Press.
8. Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69. https://doi.org/10.1007/BF00337288
9. Liboiron, M. (2021). *Pollution is colonialism*. Duke University Press.
10. O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy* (First edition). Crown.
11. Sinders, C. (2020, May 5). Rethinking Artificial Intelligence through Feminism. *CCCB LAB*. https://lab.cccb.org/en/rethinking-artificial-intelligence-through-feminism/