Scientific Argument with Supervised Learning

Jeffrey W. Lockhart
University of Michigan
jwlock@umich.edu

Abigail Z. Jacobs
University of Michigan
azjabobs@umich.edu

Abstract

The use of machine learning (ML) for scientific discovery has enabled data-driven approaches to new and old questions alike. We argue that scientific arguments based on *algorithms for discovery* hold the potential to *reinforce* existing assumptions about phenomena, under the guise of testing them. Using examples from image-based biological classification, we show how scientific arguments using supervised learning can contribute to unintended, unrealistic, or under-evidenced claims.

ML for scientific discovery Life sciences often advance through creating and testing taxonomies or classifications of entities. Statistical methods to evaluate the theorized differences between groups offer the opportunity to validate such classifications. From machine learning, the introduction of large-scale, data-driven efforts to test and redefine biological classifications has led, for instance, to innovations in population genetics [1, 2], microbiology [3, 4], and pharmacology [5].

Scientists have long looked to the brain and physical body as the source of differences between people, focusing for example on how brain imaging could be used to differentiate men and women, trans and cisgender people, autistic and non-autistic people, criminals and non-criminals, and other social divisions [6, 7, 8]. Similarly, a raft of recent work from evolutionary psychology, criminology, and computer vision has looked to establish how photographs of faces can be used to measure the biological basis of "honesty, personality, intelligence, sexual orientation, political orientation, and violent tendencies" [9, 10, 11, 12, 13, 14], resurrecting questions from a branch of eugenics known as physiognomy [15, 16]; some of this work even explicitly endorses the discredited 19$^{\text{th}}$ century phrenologist Cesare Lombroso [17]. The recent availability of large, high-dimensional data sets such as face photographs and neuroimaging has enticed ML research to these questions [15].

For scientific fields focused on differences, supervised classification offers an appealing set of tools. When applied scientists believe that groups are "truly" categorically different, classification mirrors this assumption [18, 19]. For instance, to establish group differences using MRI data, many scientists have turned to classification. As one author explains, "If a binary classifier has good performance, then clearly the groups [men's and women's brains] have restricted overlap... a classifier can only achieve perfect classification if the data points are well separated (note... the data may be well separated, even if a particular classifier is no better than random guessing)" [20]. In other words, this author—and those of 48 other papers [6]—argue that classifiers give evidence to his claim that men and women have categorically different brains by showing that it is possible to assign the correct sex label to a person using brain MRI data. But as the quote indicates, supervised learning strategies cannot challenge or suggest alternatives to scientists' theories of group difference. Moreover, this approach to ML for discovery risks "machinic neoplatonism" [21], i.e., an overconfidence that the model represents real, fundamental truths of the world simply because the algorithmic approach to the problem is appealing—potentially despite poor empirical evidence, high error rates, and limited validity [6, 16, 21, 22].

Measurement in ML for scientific discovery An emphasis on prediction without an emphasis on measurement can seem to legitimize concepts that would not otherwise hold up to scrutiny [23, 24, 25]. Gelman and colleagues illustrate this in their critique of an algorithm predicting sexual orientation from face photographs: "stripping a phenomenon of its social context... give[s] the

NeurIPS 2021 AI for Science Workshop.

feel of scientific objectivity while creating serious problems for generalizing findings to the world outside the lab or algorithm", where sexual orientation is neither a boolean category nor a fixed, socially-unmediated biological phenomenon [26].

Beyond scientific insight, supervised learning for discovery carries larger societal risks and ethical challenges researchers should be aware of. For example, the neuroscientific question "do men and women have categorically different brains?" is still hotly contested among researchers [27, 28, 29, 30] and highly politicized in society as a means of excluding women from STEM jobs, denying trans people rights, and more [27, 28]. The argument about brain differences by sex was itself a central premise of the eugenics movement that compared women's brains to "gorillas... children and savages" in the 1870s [29, 31]. ML and even brain researchers are often unaware of the eugenic history and implications of claims like this, or related claims such as that we can infer personality, sexuality, or criminality from facial structure [15, 26, 32]. Nevertheless, building technology that inaccurately supposes an essential, categorical difference between human groups risks contributing to discriminatory, polarizing, and eugenic beliefs [33].

Using classification to identify biases Even researchers trying to oppose racism often inadvertently reinforce racist, eugenic beliefs with supervised classification. Benjamin [34] offers numerous examples; more recently, a preprint [35] received widespread attention for showing that neural networks could classify race in medical imaging data such as hand X-rays. The authors publicized the paper to warn *against* racial bias in algorithms, and yet within a day the discussion of their findings was "swamped by racists" using it as proof of biological race differences [36]. This is why, we argue, it is not enough to involve ML and medical experts in such a project. Experts in the scientific and technological construction of race and other human identities are indispensable for designing ML to advance knowledge without falling into centuries-old traps.

Scientific claims with unsupervised learning In cases where the fundamental nature of groups is in question, or where historical harms have come from essentializing groups, unsupervised learning offers a variety of promising techniques to advance research while avoiding the risks we outlined. Unsupervised clustering approaches can offer a data-driven alternative to the biases and social baggage of human categories. They can also fulfill a promise of ML for discovery: to uncover patterns in data that we would not have expected, labeled, or sought out [37]. Clustering can also offer a better test of group difference hypotheses in neuroscience: Joel et al. [38] test the hypothesis that human brains come in two fundamental types, male and female, by first clustering brain imaging data and then comparing the clusters to participant sex. Unlike the classification approach [20], this approach could either support the hypothesis by finding that brains are best described by sex-segregated clusters, or reject it by finding otherwise. The authors find evidence to reject the hypothesis [38]. Other promising approaches turn to anomaly detection to evaluate categorical differences. Joel et al. [38] argue that if men and women's brains are as categorically different as some claim, then a model trained on women's brains would flag men's brains as anomalous and vice versa. This setup could produce evidence both for and against the hypothesis of categorical difference, and again the authors find evidence against it. These and related findings [19, 27, 29, 30] showing that sex is not a binary aspect of brains—and that furthering these constructions contributes to fairness-related harms—are of critical importance for ML researchers working with neuroimaging data, which at present uses sex classification as a 'gold standard' benchmark for evaluating new algorithms [6, 39, 40, 41].

Scientific claims with interpretative supervised learning The same interpretive approach that makes unsupervised learning useful for scientific discovery can be adapted for supervised learning. One such approach is illustrated by Sanchis-Segura et al. [30]: rather than uncritically reproducing group difference by asking *if* they can find differences, the authors show how researcher assumptions and data preparation choices influence findings about brains and sex. This work falls within a larger conversation about the implications of assumptions and data preparation and the emphasis on prediction in lieu of validation in algorithms for discovery [6, 22, 24, 42, 43]. (A related conversation looks to the harms made during the construction and labeling of data sets [44, 45, 46].) Other work compares performance across different model specifications, examining both what number of classes best models sex in brains and which (possibly inconsistent) features are important across models [19]. Broadly, instead of assuming a model is 'correct' and risking machinic neoplatonism—and its concomitant harms [6, 16, 21]— interpretative approaches learn from a diverse set of perspectives on a problem embodied in a diverse set of models [25, 37].

References

[1] D. R. Schrider and A. D. Kern, "Supervised Machine Learning for Population Genetics: A New Paradigm," *Trends in Genetics*, vol. 34, pp. 301–312, Apr. 2018.

[2] C. Battey, P. L. Ralph, and A. D. Kern, "Predicting geographic location from genetic variation with deep neural networks," *eLife*, vol. 9, p. e54507, June 2020.

[3] G. D. Hannigan, M. B. Duhaime, D. Koutra, and P. D. Schloss, "Biogeography and environmental conditions shape bacteriophage-bacteria networks across the human microbiome," *PLOS Computational Biology*, vol. 14, p. e1006099, Apr. 2018.

[4] S. Pollak, M. Gralka, Y. Sato, J. Schwartzman, L. Lu, and O. X. Cordero, "Public good exploitation in natural bacterioplankton communities," *Science Advances*, vol. 7, no. 31, p. eabi4717, 2021.

[5] K.-Y. Hsin, H. Kitano, Y. Matsuoka, and S. Ghosh, "Application of machine leaning approaches in drug target identification and network pharmacology," in *2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, pp. 219–219, Nov. 2015.

[6] J. W. Lockhart, "Because the Machine Can Discriminate: Machine Learning, Neuroimaging, and Epistemologies of Sex," tech. rep., SocArXiv, Aug. 2021.

[7] E. Llaveria Caselles, "Epistemic Injustice in Brain Studies of (Trans)Gender Identity," *Frontiers in Sociology*, vol. 6, 2021.

[8] O. Rollins, *Conviction: the Making and Unmaking of the Violent Brain*. Stanford: Stanford University Press, 2021.

[9] M. Kosinski, "Facial recognition technology can expose political orientation from naturalistic facial images," *Scientific Reports*, vol. 11, p. 100, Jan. 2021.

[10] K. Wolffhechel, J. Fagertun, U. P. Jacobsen, W. Majewski, A. S. Hemmingsen, C. L. Larsen, S. K. Lorentzen, and H. Jarmer, "Interpretation of Appearance: The Effect of Facial Features on First Impressions and Personality," *PLOS ONE*, vol. 9, p. e107721, Sept. 2014.

[11] Y. Wang and M. Kosinski, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images," *Journal of Personality and Social Psychology*, vol. 114, no. 2, pp. 246–257, 2018.

[12] T. F. Stillman, J. K. Maner, and R. F. Baumeister, "A thin slice of violence: distinguishing violent from nonviolent sex offenders at a glance," *Evolution and Human Behavior*, vol. 31, pp. 298–303, July 2010.

[13] X. Wu and X. Zhang, "Responses to Critiques on Machine Learning of Criminality Perceptions (Addendum of arXiv:1611.04135)," *arXiv:1611.04135 [cs]*, May 2017. arXiv: 1611.04135 version: 2.

[14] A. Kachur, E. Osin, D. Davydov, K. Shutilov, and A. Novokshonov, "Assessing the Big Five personality traits using real-life static facial images," *Scientific Reports*, vol. 10, p. 8487, May 2020.

[15] B. A. y. Arcas, M. Mitchell, and A. Totorov, "Physiognomy's New Clothes," May 2017.

[16] J. Goldenfein, "The profiling potential of computer vision and the challenge of computational empiricism," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 110–119, 2019.

[17] K. Amjad, P. D. A. A. Malik, and D. S. Mehta, "A Technique and Architectural Design for Criminal Detection based on Lombroso Theory Using Deep Learning," *LGURJCSIT*, vol. 4, pp. 47–63, Sept. 2020.

[18] N. E. Anderson, K. A. Harenski, C. L. Harenski, M. R. Koenigs, J. Decety, V. D. Calhoun, and K. A. Kiehl, "Machine learning of brain gray matter differentiates sex in a large forensic sample," *Human Brain Mapping*, vol. 40, pp. 1496–1506, Apr. 2019.

[19] C. Sanchis-Segura, N. Aguirre, Á. J. Cruz-Gómez, S. Félix, and C. Forn, "Beyond "Sex Prediction": Estimating and Interpreting Multivariate Sex Differences and Similarities in the Brain," tech. rep., July 2021.

[20] J. D. Rosenblatt, "Multivariate revisit to "sex beyond the genitalia"," *Proceedings of the National Academy of Sciences*, vol. 113, pp. E1966–E1967, Apr. 2016.

[21] D. McQuillan, "Data Science as Machinic Neoplatonism," *Philosophy & Technology*, vol. 31, pp. 253–272, June 2018.

[22] A. Z. Jacobs, "Measurement as governance in and for responsible AI," *arXiv:2109.05658 [cs]*, Sept. 2021. arXiv: 2109.05658.

[23] U. Von Luxburg, R. C. Williamson, and I. Guyon, "Clustering: Science or art?," in *Proceedings of ICML workshop on unsupervised and transfer learning*, pp. 65–79, JMLR Workshop and Conference Proceedings, 2012.

[24] A. Z. Jacobs and H. Wallach, "Measurement and fairness," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385, 2021.

[25] L. Hancox-Li and I. E. Kumar, "Epistemic values in feature importance methods: Lessons from feminist epistemology," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 817–826, Mar. 2021. arXiv: 2101.12737.

[26] A. Gelman, G. Mattson, and D. Simpson, "Gaydar and the Fallacy of Decontextualized Measurement," *Sociological Science*, vol. 5, pp. 270–280, 2018.

[27] L. Eliot, "Neurosexism: the myth that men and women have different brains," *Nature*, vol. 566, p. 453, Feb. 2019.

[28] J. W. Lockhart, "Paradigms of Sex Research and Women in STEM," *Gender & Society*, vol. 35, no. 3, pp. 449–475, 2021.

[29] G. Rippon, *The gendered brain: the new neuroscience that shatters the myth of the female brain*. London: The Bodley Head, 2019.

[30] C. Sanchis-Segura, M. V. Ibañez-Gual, N. Aguirre, Á. J. Cruz-Gómez, and C. Forn, "Effects of different intracranial volume correction methods on univariate sex differences in grey matter volume and multivariate sex prediction," *Scientific Reports*, vol. 10, pp. 1–15, July 2020.

[31] S. J. Gould, *The mismeasure of man*. New York: Norton, rev. and expanded ed., 1996.

[32] J. W. Lockhart, "'A Large and Long Standing Body': Historical Authority in the Science of Sex," in *Far Right Revisionism and the End of History: Alt/Histories* (L. D. Valencia-García, ed.), pp. 359–386, New York: Routledge, 2020. doi: 10.4324/9781003026433.

[33] O. Keyes, Z. Hitzig, and M. Blell, "Truth from the machine: artificial intelligence and the materialization of identity," *Interdisciplinary Science Reviews*, vol. 46, pp. 158–175, Apr. 2021.

[34] R. Benjamin, "Catching Our Breath: Critical Race STS and the Carceral Imagination," *Engaging Science, Technology, and Society*, vol. 2, pp. 145–156, July 2016.

[35] I. Banerjee, A. R. Bhimireddy, J. L. Burns, L. A. Celi, L.-C. Chen, R. Correa, N. Dullerud, M. Ghassemi, S.-C. Huang, P.-C. Kuo, M. P. Lungren, L. Palmer, B. J. Price, S. Purkayastha, A. Pyrros, L. Oakden-Rayner, C. Okechukwu, L. Seyyed-Kalantari, H. Trivedi, R. Wang, Z. Zaiman, H. Zhang, and J. W. Gichoya, "Reading Race: AI Recognises Patient's Racial Identity In Medical Images," *arXiv:2107.10356 [cs, eess]*, July 2021. arXiv: 2107.10356.

[36] Luke Oakden-Rayner, "Hi everyone. This thread has been swamped by racists. I'm probably gonna miss your replies, but I'll still be here in a few days when they move on or you can reach out through other channels. We appreciate all the wonderful support we've received from the community," Aug. 2021.

[37] L. K. Nelson, "Computational Grounded Theory: A Methodological Framework," *Sociological Methods & Research*, vol. 49, pp. 3–42, Feb. 2020.

[38] D. Joel, A. Persico, M. Salhov, Z. Berman, S. Oligschläger, I. Meilijson, and A. Averbuch, "Analysis of Human Brain Structure Reveals that the Brain "Types" Typical of Males Are Also Typical of Females, and Vice Versa," *Frontiers in Human Neuroscience*, vol. 12, 2018.

[39] L. Yuan, X. Wei, H. Shen, L.-L. Zeng, and D. Hu, "Multi-Center Brain Imaging Classification Using a Novel 3D CNN Approach," *IEEE Access*, vol. 6, pp. 49925–49934, 2018.

[40] W. Huf, K. Kalcher, R. N. Boubela, G. Rath, A. Vecsei, P. Filzmoser, and E. Moser, "On the generalizability of resting-state fMRI machine learning classifiers," *Frontiers in Human Neuroscience*, vol. 8, July 2014.

[41] M. Nieuwenhuis, H. G. Schnack, N. E. van Haren, J. Lappin, C. Morgan, A. A. Reinders, D. Gutierrez-Tordesillas, R. Roiz-Santiañez, M. S. Schaufelberger, P. G. Rosa, M. V. Zanetti, G. F. Busatto, B. Crespo-Facorro, P. D. McGorry, D. Velakoulis, C. Pantelis, S. J. Wood, R. S. Kahn, J. Mourao-Miranda, and P. Dazzan, "Multi-center MRI prediction models: Predicting sex and illness course in first episode psychosis patients," *NeuroImage*, vol. 145, pp. 246–253, Jan. 2017.

[42] J. M. Hofman, A. Sharma, and D. J. Watts, "Prediction and explanation in social systems," *Science*, vol. 355, no. 6324, pp. 486–488, 2017.

[43] J. M. Hofman, D. J. Watts, S. Athey, F. Garip, T. L. Griffiths, J. Kleinberg, H. Margetts, S. Mullainathan, M. J. Salganik, S. Vazire, *et al.*, "Integrating explanation and prediction in computational social science," *Nature*, vol. 595, no. 7866, pp. 181–188, 2021.

[44] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis) contents: A survey of dataset development and use in machine learning research," *arXiv preprint arXiv:2012.05345*, 2020.

[45] K. Peng, A. Mathur, and A. Narayanan, "Mitigating dataset harms requires stewardship: Lessons from 1000 papers," *arXiv preprint arXiv:2108.02922*, 2021.

[46] V. U. Prabhu and A. Birhane, "Large image datasets: A pyrrhic win for computer vision?," *Proc. WACV*, 2021.