AI & SOCIETY (2024) 39:799–801
https://doi.org/10.1007/s00146-022-01540-w

CURMUDGEON CORNER

The rise of machine learning in the academic social sciences

Charles Rahal[1,2] · Mark Verhagen[1] · David Kirk[1]

Received: 2 June 2022 / Accepted: 28 June 2022 / Published online: 23 August 2022

> ‘*In considering any new subject, there is frequently a tendency, first, to overrate what we find to be already interesting or remarkable; and, secondly, by a sort of natural reaction, to undervalue the true state of the case*’
>
> – Ada Lovelace, 1842

Machine Learning (ML) is gradually revolutionizing the social sciences as it has done for subjects like genomics and medicine. The new millennium brought an ambition to find the ‘Signal and the Noise’, followed by funding initiatives such as the creation of a working group in Computational Social Science by the Russell Sage Foundation. All aim to capitalize on ML’s ability to find intricate patterns; patterns which might have otherwise been missed in the traditional approach to model building. Figure 1 quantifies the ‘rise of machine learning’ via a regular-expression based search across all social science abstracts hosted on Scopus at the time of writing, calculating the prevalence of key words pertaining to ML over time. Growth in the use of (and discussion and debate around) ML methods in the immediate past has been remarkable; from 0.63% between 1960-2017, to nearly quadruple since (2.34%). We provide three explanations for this recent trend, and rationales for an even more optimistic view of the future:

1. **Historical Ideologies:** Social Scientists have previously had a preoccupation with parsimonious explanation and inferential ‘beta-hat’, as opposed to predictive ‘y-hat’ questions. However, the value of predictive algorithms is increasingly appreciated. The Fragile Families Challenge (Salganik et al. 2020) aimed to generate a better understanding of social determinism, but not every emergent application need be survey based. The use of optical character recognition (OCR) for digitizing archival population records (Cummins 2021) and the prediction of history (Risi et al. 2019) are prime examples of other recent and exciting applications of what ML makes possible. There are substantial public policy applications and opportunities for intervention based upon prediction, too; if we can more accurately predict rain tomorrow, we can better plan to bring an umbrella. There is also the essential realisation that ML can help with causal questions, and complement and improve classical tools designed for inference (Hofman et al. 2021), especially important given the rise of ‘Explainable Artificial Intelligence’ (XAI). The meticulous focus within ML on limiting over-fitting of the data also provides welcome encouragement for a renewed emphasis on reproducibility.
2. **Training and Accessibility:** Comparatively less attention has been paid to the development of ML skills for graduate social science candidates. Most degree-granting institutions – with exceptions such as the Oxford Internet Institute’s ‘MSc in Social Data Science’, and the University of Chicago’s ‘Masters in Computational Social Science' – maintain little emphasis on the training of ML skills. However, global initiatives like the Summer Institute in Computational Social Sciences and the data and software ‘Carpentries’ have emerged. Combined with the proliferation of ever increasing accessible ML libraries, this partially resolves concerns (Floridi 2012, p. 437) that such courses in advanced analytics (to overcome the ‘epistemological challenges’ of finding small patterns in ‘Big Data’) were ‘not exactly your standard degree at the university’.
3. **Data and Computing:** Constraints due to small-scale datasets and the ‘curse of dimensionality’ that have hampered social scientists in the past are rapidly changing, too. This is due to the enormous growth in large longitudinal surveys, long-term biobanks, and the availability of other administrative and unstructured ‘hidden’ data. Combined with substantial advances in high performance computing capacity (and the prospects of quantum computing more generally), this will allow social scientists to go beyond classical methods which were

✉ Charles Rahal
charles.rahal@sociology.ox.ac.uk

[1] Leverhulme Centre for Demographic Science and Nuffield College, University of Oxford, Oxford, UK

[2] Department of Sociology, University of Oxford, Oxford, UK

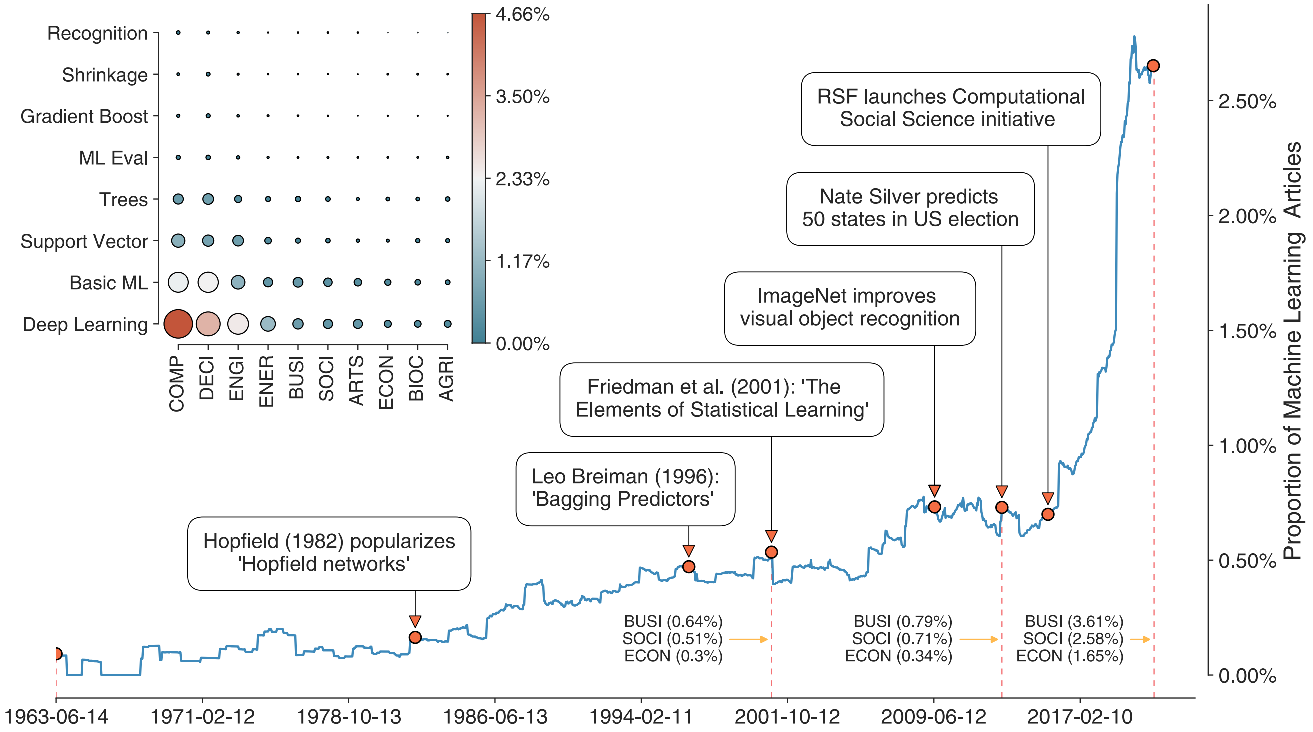

Fig. 1 The Rise of Machine Learning in the Academic Social Sciences. The blue line indicates the rolling one-year frequency at which a range of ML based terms are observed in abstracts for the `SOCI` (social sciences), `BUSI` (business) and `ECON` (economics) Scopus subject areas, with orange annotations indicating each individual long-run average for the previous year at that point in time. The inset scatter plot indicates the frequency of use of various clusterings of terms compared to a selection of other subject areas indexed by Scopus, where 'Basic ML' indicates a simple mention of 'machine learning' or 'artificial intelligence'. The x-axis of the inset relates to subject areas. For example, 'Trees' pertains to a variety of tree-based methods. Further information is available at github.com/crahal/ML_in_SocSci and via a DOI on Zenodo: 10.5281/zenodo.5918226

> – in part – designed with computational limitations in mind.

However, the social science community still has an important role to play. We must acknowledge that many of the ground-breaking yet, by now, more 'classical' methodological advances that occurred across the 20th century were made with wholly different restrictions in place: we should embrace new methodological trajectories accordingly. Social scientists need to actively ensure that ambitions which have been central to our discipline are maintained in our further development of ML methods, such as through a continued emphasis on explainability and causal reasoning (Athey and Imbens 2016). Immense care also needs to be taken to ensure that the algorithms which we develop are fair and unbiased (Mehrabi et al. 2021). Unacceptable levels of bias have already been observed in criminal justice and healthcare, and are quickly emerging in the area of recruitment, all acting in a way which amplifies existing biases and inequalities within society. Indeed, there have already been more than reasonable high profile arguments 'Against Prediction' in certain settings, unless it can be done in a socially responsible way (Harcourt 2008). Alongside all relevant ethical concerns regarding individual-level prediction, we call for further theoretical work that attempts to understand what the 'predictive ceiling' of social variables substantively represents as we further eliminate reducible error. If we take these steps, we might postulate that the use of ML in the academic social sciences is at the beginning of a sharp incline across the technologist's S-Curve. Indeed, social scientists may be beginning a wholesale change in the nature of the research process, or – at the very least – are moving from a 'peak of inflated expectations' to a 'plateau of productivity'.

Supplementary Material Further information pertaining to the methods, derivative data for visualisation, and the replication materials more broadly is available at github.com/crahal/ML_in_SocSci and via a DOI on Zenodo: https://doi.org/10.5281/zenodo.5918226.

Acknowledgements The authors are grateful of support from the Leverhulme Trust, the Leverhulme Centre for Demographic Science (LCDS) and Nuffield College. Comments gratefully received from Felix Tropf, Per Engzell, Saul Newman and Kyla Chasalow and members of the LCDS more broadly.

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? –Editor

Declarations

Conflict of interest The authors declare no competing interests.

References

Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. Proc Natl Acad Sci 113(27):7353–7360

Cummins N (2021) Where is the middle class? evidence from 60 million English death and probate records, 1892–1992. J Econ Hist 81(2):359–404. https://doi.org/10.1017/S0022050721000164

Floridi L (2012) Big data and their epistemological challenge. Philos Technol 25(4):435–437

Harcourt BE (2008) Against prediction: profiling, policing, and punishing in an actuarial age. University of Chicago Press, Chicago

Hofman JK, Duncan JW, Susan A, Filiz G, Thomas LG, Jon K, Helen M, Sendhil M, Matthew JS, Simine V et al (2021) Integrating explanation and prediction in computational social science. Nature 595(7866):181–188

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv (CSUR) 54(6):1–35

Risi J, Amit S, Rohan S, Matthew C, Duncan JW (2019) Predicting history. Nat Hum Behav 3(9):906–912

Salganik MJ, Ian L, Alexander TK, Caitlin EA, Khaled A-G, Abdullah A, Drew MA, Jennie EB, Nicole BC, Ryan JC et al (2020) Measuring the predictability of life outcomes with a scientific mass collaboration. Proc Natl Acad Sci 117(15):8398–8403

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.