# *Small Data for big insights in ecology*

Trends in
Ecology & Evolution

CellPress
OPEN ACCESS

Opinion

# ‘Small Data’ for big insights in ecology

Lindsay C. Todman [1,*] Alex Bush [2,*] and Amelia S.C. Hood [1]

**Big Data science has significantly furthered our understanding of complex systems by harnessing large volumes of data, generated at high velocity and in great variety. However, there is a risk that Big Data collection is prioritised to the detriment of ‘Small Data’ (data with few observations). This poses a particular risk to ecology where Small Data abounds. Machine learning experts are increasingly looking to Small Data to drive the next generation of innovation, leading to development in methods for Small Data such as transfer learning, knowledge graphs, and synthetic data. Meanwhile, meta-analysis and causal reasoning approaches are evolving to provide new insights from Small Data. These advances should add value to high-quality Small Data catalysing future insights for ecology.**

## Highlights

The specific context and characteristics of ecological studies mean datasets are often small and poorly suited to many advanced analytical approaches.

Some important ecological insights can only be derived from Small Data, so its collection should not be neglected.

Ongoing advances in study design, data analysis, and machine learning methods increase the value of ‘Small Data’ beyond the intrinsic purposes for which they are originally collected.

These methods include transfer learning, generating synthetic data, using causal model structures and new approaches to collating Small Data.

The greatest opportunities will arise if the reusability of Small Data is improved.

## The ‘Small Data’ trend

In a world that creates zettabytes of data each year, we are living in the ‘**Big Data**’ (see Glossary) era. ‘Big Data’ has enabled significant progress in many disciplines, including ecology, and continues to promise further advances [1]. New technologies are enabling ecological data to be collected at an unprecedented rate, and for some it is easy to assume that ‘Big Data’ is necessary to improve our understanding of complex problems. One may therefore question the value of expensive studies that culminate in a handful of data points. By contrast, many leading data scientists, companies, and trend analysts predict that the greatest advances in our capability and knowledge will be driven by methods that utilise ‘**Small Data**’ [2][i]. Such advances will greatly increase the range of problems available for analysis and broaden the insights that can be drawn. We hope this paper will help inspire ecologists to see that opportunities are emerging to integrate and maximise their research outputs at all levels.

Small Data and Big Data represent two ends of a continuum. Big Data is typically considered as high volume, composed of a complex mix of data types, and may also be generated and processed at high speed [3]. Conversely, Small Data contain a smaller volume of one or few measurement types, may have inconsistent structure, and may only be generated intermittently or even as a one-off. As a result, models analysing Small Data have a greater risk of overfitting because the number of features in a model are high relative to the degrees of freedom in the dataset. Overfitting leads to models that are influenced by patterns in the sample data that are not present in the population (i.e., noise), thus leading to spurious predictions that **generalise** poorly to new data. In very Small Data, this may mean no model can be fitted and trends in the data cannot be interpreted. As a result, the legacy value of Small Data to advance our collective knowledge is considered proportionately small, rather than recognising their value for offering unique insights about rare events, or providing tailor-made answers to specific questions.

In our opinion, Small Data is particularly important for the ecological community for several reasons. Firstly, ecology incorporates mission-oriented disciplines (e.g., conservation, agriculture), and collecting Small Data facilitates rapid, tailored, evidence-based decision-making. Secondly, understanding rarity (e.g., rare species) is fundamental to ecology, and Small Data are needed to do this. Thirdly, many data collection methods are costly and too time-consuming to scale-

[1]University of Reading, School of Agriculture, Policy and Development, Earley Gate, Whiteknights Road, PO Box 237, Reading RG6 6AR, UK
[2]Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK

*Correspondence:
l.todman@reading.ac.uk (L.C. Todman) and alex.bush@lancaster.ac.uk (A. Bush).

up to generate Big Data; this is particularly true in habitats where data collection is particularly arduous (e.g., rainforests, deep soil), and these are often understudied habitats that can provide unique insights. Application of new methods for Small Data, as well as amendments to enhance their usability, could therefore help to maximise the value of ecological data, either for stand-alone use, or as part of our collective understanding.

## Uses and value of Small Data

### Intrinsic value

Small datasets are often collected with a particular purpose in mind, for example, a research hypothesis or a monitoring exercise, that offer direct insights into the stand-alone dataset (Figure 1). Controlled experiments may only be able to collect a modest amount of information, but they are an important tool for facilitating rapid, tailored, evidence-based decision-making and add to scientific understanding of the causal interactions in ecology. **Counterfactuals** can be observed (as much as is practically possible, e.g., using matching method [4]), but experiments are often expensive and time-consuming leading to Small Data. These approaches are widely used in ecology, spanning 'manipulative' experiments where conditions are actively influenced and 'observational' experiments where differences between existing conditions are measured [5]. The design of these experiments is important to increase the value of the data that are collected and enable causal inference [6]. Equally, resource constraints mean trade-offs between the scale, frequency, and quality of measurements are implicit in all field studies. So, while empirical evidence to guide policies and mitigate the global decline of biodiversity is key, our understanding is also inherently constrained by the fact that most species are rare. A single approach would not work for all species, and therefore, the value of monitoring data collected for modelling rare species depends on a careful, potentially even bespoke, design [7].

In addition to study design, using analytical methods appropriate to Small Data can support new insights and increase the predictive performance of models. For example, regularisation methods can be used for linear models with small sample sizes [8] while in machine learning Support Vector Machines are a common method for Small Data [2]. Dimension reduction approaches can also improve model performance [2].

### Collating

Conceptually, the simplest method to reuse Small Data is to link Small Datasets together so they become more 'Big-Data like' (Figure 1) [9]. One increasingly popular method is **knowledge graphs**, which represent the relationships between different data in a flexible and machine-readable way (Table 1). Unlike **relational databases** in which the relationships between data are implicit to the data structure (e.g., data are stored in columns and relationships between columns are known with each row corresponding to an individual), instead knowledge graphs store the relational information for each data point as data in its own right [10]. They can integrate the data and related contextual information from multiple sources. By storing the data and the type of relationship between them as a series of 'nodes' and 'edges', knowledge graphs enable novel queries about the linkages between data [11,12].

Evidence synthesis such as meta-analysis is important for understanding causality in ecology, because when similar effects are seen in multiple studies, it builds evidence for the causal mechanisms behind the treatments. Conducting a meta-analysis can be time-consuming, particularly when linking several small studies, and one challenge is to increase evidence synthesis efficiency [13]. Approaches to do this include automating the data collation process [14,15] and dynamic meta-analysis [16] which improves the **transferability** to different contexts (Table 1). Another challenge for data collated from small studies is missing data and differences in study

## Glossary

**Bayesian Belief Network:** a probabilistic graphical model that captures the relationships between variables in a series of nodes and edges, useful for causal reasoning in uncertain domains.

**Big Data:** data with typical characteristics such as high volume, collected at high frequency and contains a complex mix of data types.

**Counterfactual:** expresses what would have happened if a treatment had not been applied, in designed experiments this is typically the control treatment to which any other treatments are compared.

**Data augmentation:** a data science method to create a kind of 'synthetic data' by distorting existing data.

**Deep learning model:** a complex neural network with multiple nodes per layer and multiple layers (typically >3).

**Generalise/generalisability:** the ability of a model to perform well for previously unseen data.

**Knowledge graph:** a graph-based structure that integrates data by encoding values of individual data points (nodes) and relationships between them (edges).

**Natural language processing:** a branch of machine learning/artificial intelligence that aims to learn the patterns in written or spoken language to enable computers to understand language.

**Neural network:** a machine learning method that loosely attempts to mimic processes in the human brain to learn patterns in data, 'neurons' (nodes) are arranged in a series of layers with each layer learning from a further transformation of the data to uncover more detail.

**Relational database:** a method for data integration in which data are organised according to predefined relationships, for example, in a series of columns with each row corresponding to an individual.

**Small Data:** data with typical characteristics such as low volume, collected as a one off or sporadically with a specific data structure that is not generalisable or has irregularities (e.g., variables that are not observed elsewhere, differences in measurement methods to other datasets).

**Structural equation model:** a multivariate statistical analysis technique that originated from causal modelling structure. Bayesian meta-analysis could overcome these challenges and is commonly used in disciplines such as medicine, but has received less attention in ecology (Table 1) [17].

### Transfer learning

AI methods have typically required big training data sets to allow computers to 'learn the rules' for a classification problem from scratch (Figure 1) [18]. **Transfer learning** leverages existing big datasets with a lot of labelled information to train the architecture of a model (that may include millions of parameters), and that primed structure is then applied to solve a problem in an analogous way for Small Data (Table 1). Thus, while all models aim to generalise an understanding to unseen data based on patterns within the training data, transfer learning does so for an entirely new task, while retaining knowledge of the generic processes in the parent model.

Transfer learning is effective because **neural networks** start by identifying features that are common to a wide variety of tasks, whereas later layers are most specific to the features upon which the model was trained. Thus, base layers from an existing model are often a better starting point for a wide variety of tasks than random initialisations, but tuning is required to decide at which point to split the model for a new task [19]. In imagery, base features like texture waveform, vector recognition, and contiguous regions of contrast are all identified early on by neural networks (e.g., [20]). As a result, models trained on the generic image database ImageNet have then been tuned to classify other images such as photos of diseased plant leaves [21], and, by treating spectrograms as images, for identification of birds and frogs from audio (Table 1) [22].

One approach for transfer learning is to design a **deep learning model** that solves multiple tasks, sharing layers at first for all tasks and preserving generality, before subsequently splitting task-specific layers [23,24]. A more common alternative is to use an existing pre-trained deep learning model (e.g., Keras Applications[ii] or MicrosoftML packages[iii] for both R and Python users), and only replace the last layers to suit the new Small Data task [25,26]. This approach is particularly popular for image and audio processing because the pre-trained model can dramatically reduce the size of the input dataset and computation time.

### Synthetic data

In many situations, Small Data can be supplemented prior to the application of machine learning by the addition of computer-generated **synthetic data** (Figure 1). These data are generated in a systematic way to incorporate key characteristics of the Small Data, while the 'noise' in the data is varied. The simplest form of data generation is **data augmentation**, whereby the original Small Data are transformed or distorted to provide alternative instances of the data. This may be cropping or transposing for image processing [27], a change in pitch or noise overlay for an audio sample (e.g., to classify animal noises [28]), or applying a transformation, such as the radial basis function (e.g., applied to clinical trial data [29]). However, these augmentation methods are still based on the original sample, and **generalisability** can be limited.

More complex methods to generate synthetic data aim to create new samples with characteristics similar to the original data points. For image analysis, methods include Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) [30]. A statistical model can also be used to generate data with similar statistical characteristics to a Small Data set (Table 1). Synthetic data of this kind is still related to the Small Data set from which statistical characteristics have been estimated, but within the synthetic data set more examples of the expected noise are included, so that the noise is not overfitted in a machine learning model.

and uses a hierarchical structure to capture the relationship between measured and unobserved (latent) variables.

**Synthetic data:** data that are computer generated, generally to provide a larger dataset that mimics the characteristics of a smaller data set.

**Transferability:** the ability to reuse knowledge from relevant previous learning.

**Transfer learning:** a machine learning method where a similar but larger dataset is used to develop an initial model which is then 'fine-tuned' using a smaller data set.

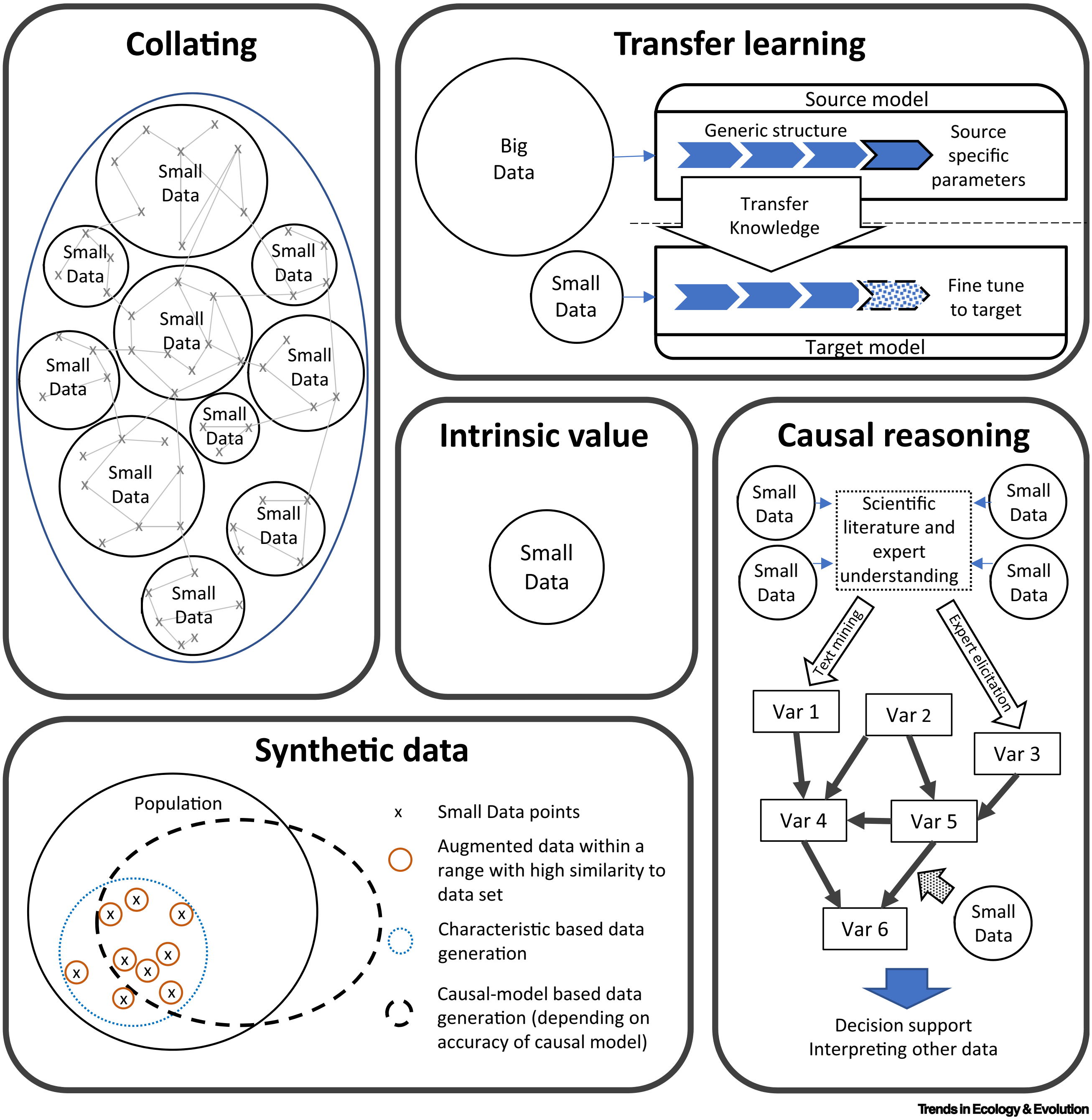

Figure 1. Uses of Small Data; see Table 1 for advances in these areas.

Another alternative is to use a causal model of the relevant process to capture the key signals that should also theoretically be present in the Small Data set. While this clearly requires some insight into the causal processes, these could be hypothetical and be tested with the Small Data set. Synthetic data could still be useful predictively even if it includes a number of realisations that are not present in the true population, as long as sufficient samples are present in the population,

Table 1. Inspiring advances in Small Data use

| | Method and use | Examples | Opportunities for ecology |
|---|---|---|---|
| Collating data | Knowledge graph, novel queries | Dorpinghaus *et al.* [48] developed an ontology to improve interoperability of clinical data from patients, combining it in a knowledge graph with contextual information and information from literature studies, thus linking numerous Small Datasets. They demonstrate how this knowledge graph enables them to answer queries, such as 'which patients are found most often in the context of a risk group?' combining aspects from data of different structures. | Structured linking of Small Data through common ontologies would enable novel queries. For example, rapidly locating data from studies with common contextual factors and direct queries on the linkages between data. |
| | Knowledge graph, meta-analysis | Tiddi *et al.* [14] developed a knowledge graph of social science studies on human co-operation in order to conduct a meta-analysis. They develop a structured schema for the data they collate and suggest this as a format for data formatting of future studies to allow the meta-analysis to be updated. | Similar structures for ecological data used for new studies would facilitate rapid integration into meta-analyses. |
| | Dynamic meta-analysis | Dynamic meta-analysis [16] enables users to interactively filter and weight data based on their own criteria (e.g., climate) via an online interface. The user can tailor their analysis easily by clicking buttons which produce statistical outputs (e.g., subgroup analyses) and plots (e.g., funnel plots). | This approach could increase the impact of Small Data, as it makes analysis of collated data easily accessible to decision-makers to answer questions in specific contexts by reusing collated datasets. |
| Transfer learning | Image processing, species identification | Knausgård *et al.* [49] used existing frameworks for object identification and feature extraction to automated marine fish identification. | There are a range of species identification applications for these methods and many more opportunities for their use for additional species. |
| | Natural language processing | Models used to analyse wildlife Twitter hashtags [50] or to automate the mining of taxonomic information from scientific literature [51] were transferred from test processing pipelines trained on Wikipedia entries and medical literature, respectively. | The general layers from natural language processing algorithms could greatly increase opportunities to collate insights from ecologically specific texts. |
| | Natural language processing, genomics | Transfer learning from text processing has improved the ability to infer gene functions of microbes from complex sequence datasets [52,53]. | Provides opportunities to make greater use of public omics databases [54]. |
| Synthetic data | Synthetic data, based on statistical properties | Kantidakis *et al.* [55] generated synthetic data for 1000 individuals with the statistical characteristics of clinical trial data to model expected survival time for a bone cancer (osteosarcoma). The synthetic data was then used to train a neural network that had comparable performance to an existing modelling approach. | These approaches could be applied to generate synthetic data from designed experiments so that machine learning can be used to develop predictive models, potentially improving predictive capacity compared with regression models. |
| | Synthetic data, using causal model | Trafton *et al.* [56] generated synthetic data for 20 000 individuals using a process-based cognitive model of human decision-making to predict behaviour when managing multiple unmanned arial vehicles, and combined this with observations of ten individuals to improve predictive performance of a deep neural network.<br>Mazumder *et al.* [57] used a process-based model of a heart to generate synthetic data to combine with observed data to improve coronary artery disease classification. | There are numerous causal models in ecology that could be used to supplement experimental data for a range of applications, for example, game theory models of human or animal behaviour that could be supplemented with observations. The challenge will be to develop approaches to use these robustly, respecting the limitations of uncertain causal models. |
| | Causal model, natural language processing | Ancin-Murguzur and Hausner [35] used text mining to develop a map of causal relationships in an artic tundra ecosystem. | Causal structures from mining the scientific literature could be used to improve model structures used for analysis of Big Data and collated datasets. |

although this would decrease algorithm efficiency. Where reasonable causal models exist, and risks of extrapolation of knowledge are considered, this could provide a powerful way to supplement Small Data in a way that goes beyond the characteristics of the original data by drawing on broader process understanding.

### Causal reasoning

Causal inference can be used in a number of different approaches from machine learning to decision support tools (Figure 1). It is of increasing interest in machine learning to overcome the criticism that the structure of these approaches is a 'black box', with growth in 'explainable AI' and 'causal AI' that make algorithmic reasoning more explicit by drawing on existing understanding or by inferring causality from data [31]. Bayesian networks have been used in a range of contexts in ecology and offer a way to combine empirical data with a wider body of knowledge for applications such as risk assessment [32]. In ecology, the causal structure is often elicited from experts [33]. In many cases, these experts are researchers in the topic, whose expertise has developed through their previous work including the collection and analysis of Small Data. **Structural equation modelling** [34] and **Bayesian Belief Networks** [33] are important methods for quantitative causal modelling, and Small Data can play an important role in both establishing the causal structure of a model and in parameterising specific links within the model. Similarly, **natural language processing** is increasingly being applied to scientific text to learn causal model structures (Table 1) [35,36]. Rather than using Small Data directly, an algorithm takes a researcher's description of their data sets, analysis, and results, and recognises causal words and phrases (e.g., 'X increased Y') to develop a model of causal interactions. While this relies on the accuracy of statement in published papers, justifiable approaches of this kind are linking the individual insights drawn by researchers and can be used to look for common patterns across multiple sources.

## Reusability of data and insights

Many of the methods presented in this paper depend on the ability to integrate Small Data with other datasets: big or small. Such integration is easiest if datasets meet the FAIR data principles, and are: Findable, Accessible, Interoperable, and Reusable [37]. Significant efforts have been made to increase the FAIRness of datasets, such as many funders and ecology journals now mandating public data archiving [38], but the majority of datasets in ecology still do not meet FAIR research principles: for example, 64% of authors in ecology and evolution archive their data in a way that prevents reuse [39]. Reporting guidelines, such as checklists, have become increasingly common in the field of medicine in the past two decades [40], and these guidelines have successfully improved reporting standards [41–43]. The wide uptake of reporting guidelines in medicine has been facilitated by their widespread endorsement in medical journals[iv] and the development of an online library of searchable guidelines[v]. Ecologists also need support to reformat their data in a way that makes it reusable, as the format used for an individual study is rarely the most appropriate structure to enable reuse [44]. An infrastructure that provides access to data management specialists is needed along with time for data formatting, and this requires sufficient funding [45]. Improved infrastructure, wider uptake of reporting guidelines, and FAIR data principles in ecology would increase the impact of Small Data by increasing the ability to integrate and amalgamate datasets.

Meanwhile, we recognise that reusability may not be pragmatic for the most nuanced datasets for which it is difficult for the data to conform to existing standards. Yet, we propose that the causal understanding derived from these data is a 'knowledge fragment' that could be reused. Clear reporting of the key findings and contextual factors using common ontologies and phrasing for causal interactions could enable mining of the literature to develop causal models. In Bayesian network models, 'network fragments' or 'idioms' are developed to capture patterns of reasoning that are commonly used within a discipline so that they can be used elsewhere [46,47]. While Small Data may not be capturing common patterns, capturing the understanding from a Small Data study in a machine-readable format such as a knowledge graph would allow understanding to be integrated by looking for recurring, overlapping, or contradicting causal networks. This could enable rapid integration of causal understanding from Small Data studies and highlight areas where understanding is most uncertain.

## Concluding remarks

Significant advances in data science and machine learning now allow Small Data in ecology to be viewed differently, and given the active interest in Small Data methods across many fields, new methods are likely to emerge in the near future with applicability to ecology (see Outstanding questions). This paper aims to (i) highlight the growing range of uses of Small Data available to the ecological research community; (ii) add further support for good data management to improve data reusability, be this raw datasets or key 'knowledge fragments' drawn from complex research studies; and (iii) emphasise the continuing importance of 'Small Data' of high quality so that valuable data collection of this kind is not neglected due to new methods of data collection. Though widely available, 'Small Data' is underused and deprioritised by data scientists and funders; we recommend that it is reprioritised. In the words of John Ruskin 'It is small, if you will; but when you begin to think of things rightly, the ideas of smallness and largeness pass away'.

## Outstanding questions

Where else can these methods for Small Data be applied to draw new insights in ecology?

How can synthetic data from numerous process-based models in ecology be used appropriately to enable new insights from Small Data while respecting the uncertainty in process-based models? And how should Small Data be collected to increase the effectiveness of synthetic data methods?

How can the understanding from analysis of Small Data be captured in a way that enables reuse of this 'knowledge fragment'?

How can we effectively incentivise data sharing to build further momentum to increase good data management and sharing practices?

### Acknowledgments

We acknowledge the support of 'Championing INnovation In the DIgital ENvironment' (NE/S016287/1) for L.C.T. and A.B. No new data were created in this work.

### Declaration of interests

No interests are declared.

### Resources

[i] www.gartner.com/smarterwithgartner/gartner-top-10-data-and-analytics-trends-for-2021

[ii] https://keras.io/

[iii] www.nuget.org/packages/Microsoft.ML

[iv] www.prisma-statement.org/Endorsement/PRISMAEndorsers

[v] www.Equator-Network.Org

### References

1. Farley, S.S. *et al.* (2018) Situating ecology as a Big-Data science: current advances, challenges, and solutions. *Biosci. J.* 68, 563–576
2. Kokol, P. *et al.* (2021) Machine learning on small size samples: a synthetic knowledge synthesis. *Sci. Prog.* 105, 368504211029777
3. Younas, M. (2019) Research challenges of big data. *Serv. Oriented Comput. Appl.* 13, 105–107
4. Ribas, L.G.S. *et al.* (2021) Estimating counterfactuals for evaluation of ecological and conservation impact: an introduction to matching methods. *Biol. Rev.* 96, 1186–1204
5. Wiersma, Y.F. (2022) A review of landscape ecology experiments to understand ecological processes. *Ecol. Process.* 11
6. Kimmel, K. *et al.* (2021) Causal assumptions and causal inference in ecological experiments. *Trends Ecol. Evol.* 36, 1141–1152
7. Jeliazkov, A. *et al.* (2022) Sampling and modelling rare species: conceptual guidelines for the neglected majority. *Glob. Chang. Biol.* 28, 3754–3777
8. Finch, W.H. and Hernandez Finch, M.E. (2016) Regularization methods for fitting linear models with small sample sizes: fitting the lasso estimator using R. *Pract. Assess. Res. Eval.* 21, 7
9. Kitchin, R. and Lauriault, T.P. (2014) Small data in the era of big data. *GeoJournal* 80, 463–475
10. Hogan, A. *et al.* (2022) Knowledge graphs. *ACM Comput. Surv.* 54, 1–37
11. Tuhin, I.A.K. *et al.* (2022) Smart cybercrime classification for digital forensics with small datasets. In *International Symposium on Cyber Security, Cryptology, and Machine Learning*, pp. 270–280, Springer International Publishing
12. Dimitrova, M. *et al.* (2021) Infrastructure and population of the OpenBiodiv biodiversity knowledge graph. *Biodivers. Data J.* 9, e67671
13. Sutherland, W.J. and Wordley, C.F.R. (2018) A fresh approach to evidence synthesis. *Nature* 558, 364–366
14. Tiddi, I. *et al.* (2020) Fostering scientific meta-analyses with knowledge graphs: a case-study. In *The Semantic We. ESWC 2020. Lecture Notes in Computer Science*, pp. 287–303, Springer International Publishing
15. Futia, G. and Vetrò, A. (2020) On the integration of knowledge graphs into deep learning models for a more comprehensible AI—three challenges for future research. *Information* 11, 122
16. Shackelford, G.E. *et al.* (2021) Dynamic meta-analysis: a method of using global evidence for local decision making. *BMC Biol.* 19, 33
17. Pappalardo, P. *et al.* (2020) Comparing traditional and Bayesian approaches to ecological meta-analysis. *Methods Ecol. Evol.* 11, 1286–1295
18. Weiss, K. *et al.* (2016) A survey of transfer learning. *J. Big Data* 3, 1
19. Yosinski, J. *et al.* (2014) *How transferable are features in deep neural networks?* Proceedings of the 27th International Conference on Neural Information Processing Systems Volume 2 pp. 3320–3328
20. Krizhevsky, A. *et al.* (2017) ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90
21. Atila, Ü. *et al.* (2021) Plant leaf disease classification using EfficientNet deep learning model. *Ecol. Inform.* 61, 101182

22. LeBien, J. *et al.* (2020) A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. *Ecol. Inform.* 59, 101113
23. Pouyanfar, S. *et al.* (2018) Multimodal deep learning based on multiple correspondence analysis for disaster management. *World Wide Web* 22, 1893–1911
24. Tian, H. *et al.* (2018) Multimodal deep representation learning for video classification. *World Wide Web* 22, 1325–1341
25. Molchanov, P. *et al.* (2016) Pruning convolutional neural networks for resource efficient inference. *arXiv* Published online November 19, 2016. https://doi.org/10.48550/arXiv.1611.06440
26. Tian, H. *et al.* (2020) Evolutionary programming based deep learning feature selection and network construction for visual data classification. *Inf. Syst. Front.* 22, 1053–1066
27. Seib, V. *et al.* (2020) Mixing real and synthetic data to enhance neural network training – a review of current approaches. *arXiv* Published online July 17, 2020. https://doi.org/10.48550/arXiv.2007.08781
28. Nanni, L. *et al.* (2020) Data augmentation approaches for improving animal audio classification. *Ecol. Inform.* 57, 101084
29. Izonin, I. *et al.* (2021) Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method. *Math. Biosci. Eng.* 18, 2599–2613
30. Moreno-Barea, F.J. *et al.* (2020) Improving classification accuracy using data augmentation on small data sets. *Expert Syst. Appl.*, 161113696
31. Shao, Z. *et al.* (2022) Tracing the evolution of AI in the past decade and forecasting the emerging trends. *Expert Syst. Appl.* 209, 118221
32. Kaikkonen, L. *et al.* (2020) Bayesian networks in environmental risk assessment: a review. *Integr. Environ. Assess. Manag.* 17, 62–78
33. Marcot, B.G. and Penman, T.D. (2019) Advances in Bayesian network modelling: integration of modelling technologies. *Environ. Model. Soft.* 111, 386–393
34. Pearl, J. (1998) Graphs, causality, and structural equation models. *Sociol. Methods Res.* 27, 226–284
35. Ancin-Murguzur, F.J. and Hausner, V.H. (2021) Replication data for: causalizeR: a text mining algorithm to identify causal relationships in scientific literature. *PeerJ* 9, e11850
36. van Bilsen, J.H.M. *et al.* (2020) Seeking windows of opportunity to shape lifelong immune health: a network-based strategy to predict and prioritize markers of early life immune modulation. *Front. Immunol.* 11, 644
37. Wilkinson, M.D. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 1–9
38. Mislan, K.A.S. *et al.* (2015) Elevating the status of code in ecology. *Trends Ecol. Evol.* 31, 4–7
39. Roche, D.G. *et al.* (2015) Public data archiving in ecology and evolution: how well are we doing? *PLoS Biol.* 13, e1002295
40. Simera, I. and Altman, D.G. (2009) Writing a research article that is "fit for purpose": EQUATOR Network and reporting guidelines. *BMJ Evid. Based Med.* 14, 132–134
41. Plint, A.C. *et al.* (2006) Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Med. J. Aust.* 185, 263–267
42. Turner, L. *et al.* (2012) Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Syst. Rev.* 1, 1–7
43. Stevens, A. *et al.* (2014) Relation of completeness of reporting of health research to journals' endorsement of reporting guidelines: systematic review. *BMJ* 348, g3804–g3804
44. Poisot, T. *et al.* (2019) Ecological data should not be so hard to find and reuse. *Trends Ecol. Evol.* 34, 494–496
45. Perrier, L. *et al.* (2020) The views, perspectives, and experiences of academic researchers with data sharing and reuse: a meta-synthesis. *PLoS One* 15, e0229182
46. Kyrimi, E. *et al.* (2020) Medical idioms for clinical Bayesian network development. *J. Biomed. Inform.* 108, 103495
47. Carriger, J.F. *et al.* (2019) An introduction to Bayesian networks as assessment and decision support tools for managing coral reef ecosystem services. *Ocean Coast. Manag.* 177, 188–199
48. Dörpinghaus, J. *et al.* (2021) An efficient approach towards the generation and analysis of interoperable clinical data in a knowledge graph. In *Proceedings of the 16th Conference on Computer Science and Intelligence Systems*, pp. 59–68
49. Knausgård, K.M. *et al.* (2021) Temperate fish detection and classification: a deep learning based approach. *Appl. Intell.* 52, 6988–7001
50. Edwards, T. *et al.* (2022) Identifying wildlife observations on twitter. *Ecol. Inform.* 67, 101500
51. Guillarme, N.L. and Thuiller, W. (2021) TaxoNERD: deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods Ecol. Evol.* 13, 625–641
52. Hamid, M.-N. and Friedberg, I. (2020) Transfer learning improves antibiotic resistance class prediction. *bioRxiv* Published online April 18, 2020. https://doi.org/10.1101/2020.04.17.047316
53. Hoarfrost, A. *et al.* (2022) Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nat. Commun.* 13, 1–12
54. David, M.M. *et al.* (2022) Revealing general patterns of microbiomes that transcend systems: potential and challenges of deep transfer learning. *mSystems* 7, e0105821
55. Kantidakis, G. *et al.* (2021) A simulation study to compare the predictive performance of survival neural networks with Cox models for clinical trial data. *Comput. Math. Methods Med.* 2021, 1–15
56. Trafton, J.T. *et al.* (2020) Using cognitive models to train big data models with small data. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1413–1421
57. Mazumder, O. *et al.* (2022) Synthetic PPG signal generation to improve coronary artery disease classification: study with physical model of cardiovascular system. *IEEE J. Biomed. Health Inform.* 26, 2136–2146