Data Synthesis for Big Questions:
From Animal Tracks to Ecological Models

Rose Trappes, University of Exeter

Abstract

This paper addresses a relatively new mode of ecological research: data synthesis studies. Data synthesis studies involve reusing data to create a general model as well as a reusable aggregated dataset. Using a case from movement ecology, I analyse the trade-offs and strategies involved in data synthesis. Like theoretical ecological modelling, I find that synthesis studies involve a modelling trade-off between generality, precision and realism; they deal with this trade-off by adopting a pragmatic kludging strategy. I also identify an additional trade-off, the synthesis trade-off, between making data easy to synthesise for a particular project, on the one hand, and facilitating data reuse for other projects, on the other. In response to this synthesis trade-off, researchers create flexible datasets that are relatively easy to use for particular projects and can be adjusted to suit some other purposes. The flexibility compromise is also found in broader open data efforts, making it a significant element in the future of data-intensive ecology.

Keywords: open data; data-centric science; evidence synthesis; ecological model building; kludging; movement ecology; data journey

1   Introduction

Ecologists and conservation biologists frequently talk about the "big questions": questions about global and long-term changes, especially climate change, habitat loss, and mass extinction; questions about key ecological hypotheses, such as theories of succession or community structure; and questions about conservation strategies for complex ecological systems across regions and national borders. Researchers have recently set their sights on big data as a resource for answering these big questions. As two ecologists recently put it, "These large science questions generate a common requirement, data, lots of it collected on scales that require resources outside the range available to even the best-funded single principal investigator or group project." (Schimel and Keller 2015)

One emerging approach in ecology for acquiring and utilising big data to answer big questions is *data synthesis*. This approach capitalises on open science policies and practices, which have led to vast amounts of ecological data being published online (Hampton et al. 2015; O'Dea et al. 2021; Roche et al. 2021). Data synthesis involves gathering this data into large open datasets and using it to build large-scale and highly general models of complex phenomena. In this paper I ask what is involved in synthesising ecological data to answer big questions. What are data synthesis studies, how do they work, and what consequences do they have for ecological knowledge production?

Data synthesis studies are a kind of model building. In the philosophy of ecology, model building is seen to involve a trade-off between generality, realism and precision, in which modellers typically adopt a strategy of sacrificing one of the three desiderata (Levins 1966; Odenbaugh 2003; Weisberg 2006; Inkpen 2016; Elliott-Graves 2020). Data synthesis studies do produce models that seem to conform to this formula: highly general and realistic models that sacrifice the precision of capturing small effects or lower-level phenomena. However, I argue that the actual process of building models in data synthesis studies doesn't involve a strategy of sacrificing one modelling desideratum. Instead, modellers adopt a kludging strategy, incrementally adjusting their models to be sufficiently general, realistic, and precise to address a particular big question.

As well as building general models, synthesis studies also generate synthetic datasets that are made available to the community of ecologists for further reuse. I argue that the dual goals of synthesis studies create an additional challenge: what I call the *synthesis trade-off* between making heterogenous data easier to use for a particular general model, on the one hand, and supporting multiple different reuse scenarios on the other. In developing an account of the synthesis trade-off, I build on existing accounts of scientific data, such as data journeys (Leonelli 2020; Leonelli and Tempini 2020) and adequacy-for-purpose data evaluation (Bokulich and Parker 2021). The synthesis trade-off also relates to discussions of standardisation and coordination across datasets, in which choices about classificatory systems, spatiotemporal scales, or date and location format affect what information is retained, where emphasis is placed, and what sorts of analyses can be performed (Bowker 2000; Shavit and Griesemer 2009; 2011; Leonelli 2016; Sterner, Witteveen, and Franz 2020).

I find that biologists respond to the synthesis trade-off by developing flexible datasets that can be adjusted to suit some (but not all) purposes. This flexibility compromise is also seen in other open data efforts, making it a strategy that has broader implications for data-intensive ecological research. Ecologists face a future where decisions about methods and research questions are made at the intersection of building general models and developing multipurpose data

resources. The synthesis trade-off and flexibility compromise therefore deserve further philosophical attention as important new factors driving contemporary scientific practice.

To develop this analysis, I focus on a case from movement ecology: the study of organismal movement in its ecological context, chiefly using animal tracking technology such as radio tags or GPS devices. Recently, movement ecologists have been discussing the challenges of opening up animal tracking data so that it can be more effectively reused, especially to answer big questions (Campbell et al. 2016; Kays et al. 2021; Williams et al. 2020; Sequeira et al. 2021; Rutz 2022). They have also produced several open aggregated datasets as part of synthesis studies or regional projects (Sequeira et al. 2018; Tucker et al. 2018; Hindell et al. 2020; Davidson et al. 2020). I examine one particular case, a project synthesising Antarctic predator tracking data to answer a big question about conservation in the Southern Ocean (Hindell et al. 2020; Ropert-Coudert et al. 2020).

I begin in Section 2 by introducing synthesis studies and the case study. In Section 3 I consider this case in light of standard accounts of ecological modelling. Using resources from the philosophy of scientific data, in Section 4 I argue that synthesis studies involve the synthesis trade-off and discuss what strategies biologists adopt to deal with this trade-off. In Section 5 I elaborate how the same trade-off and strategies occur for open data, and section 6 closes the paper with a summary.

2   Data Synthesis Studies

Data synthesis is a kind of data reuse[1] in which data from multiple sources are combined into an aggregated dataset and used to produce a general model. In this section I characterise data synthesis studies as a distinctive kind of evidence synthesis with two goals: developing a model and providing a multipurpose dataset. I also introduce my case study, a synthesis study from movement ecology. First, however, I briefly consider why ecologists are interested in data synthesis at all.

Ecologists' big questions are a heterogenous bunch, ranging from understanding the effects of climate change through evaluating hypotheses about island biogeography to providing input for international conservation strategies. Such questions are big in at least three respects. First, they often concern large geographic areas and time periods, such as regions, continents, and

---

[1] Bokulich and Parker (2021, 16) distinguish data reuse from data repurposing. On their account, data reuse is restricted to using data for the same purpose it was originally collected for, for instance when researchers reanalyse a past dataset using new and improved data processing or analysis methods. Data repurposing, in contrast, refers to using data for new purposes not originally foreseen during data collection. However, this distinction is not easy to apply in the context of data synthesis studies. For instance, some synthesis studies are conducted to answer the same research questions that some of their data were also used to answer (just at a larger scale). On the other hand, some data are collected with the express intention to make that data available for other researchers to use for their own purposes. To avoid complications, I use the term "reuse" in a broad sense to cover any kind of reuse, whether for the same or different purposes.

hemispheres, and decades or even centuries of change. Second, and partly as a consequence of their broad scope, ecologists' big questions typically concern very heterogenous and complex ecological systems and processes. Third, big questions are often about pressing problems that require coordinated action amongst a diverse array of parties and sensitivity to a range of interests and values.

Big questions are, unsurprisingly, hard to answer. Ecological models can sometimes attain broad scope and high levels of generality while still being realistic enough to apply to real-world problems like habitat loss and conservation (Levins 1966; Elliott-Graves 2020). However, the complexity of many of ecological phenomena defy analytical methods, and the heterogeneity of ecological systems limits generalisation (Mitchell 2003; Elliott-Graves 2016; 2018). Recently, ecologists have focused on big data as a potential resource for addressing big questions. Big ecological data are surely not sufficient, especially as many big questions involve ethical, political, and social elements that require contributions from areas beyond the natural sciences (Efstathiou 2016). Nevertheless, the scope and detail of big data make them promising for studying the large-scale, general, and complex matters of big questions.

There are three main strategies to acquire big data in ecology and conservation biology: large collaborative research platforms, global citizen science projects, and data synthesis studies (Hampton et al. 2013; 2017). Large collaborative platforms, such as the Long Term Ecological Research (LTER) program, and global citizen science, like eBird and iNaturalist, can both be described as "big science" producing big data through centrally coordinated data collection (Sullivan et al. 2009; 2014; Devictor, Whittaker, and Beltrame 2010; Theobald et al. 2015; Di Cecco et al. 2021; Waide and Kingsland 2021).[2] In contrast, data synthesis studies utilise the "long tail" of science: the many smaller studies conducted for particular purposes (Hampton et al. 2013). The recent growth of the open science movement in ecology and related fields has led to large amounts of ecological data being available online. Synthesis studies make use of this open data, as well as contributing new open synthetic datasets.

Ecologists are interested in data synthesis as a way to bring together heterogenous data from many primary studies and thereby reach broader scopes and higher generality. As such, data synthesis studies fall under the broad category of *evidence synthesis*. Following its prominence in evidence-based medicine, evidence synthesis has found traction in many fields, including ecology and conservation biology (Carpenter et al. 2009; Nakagawa et al. 2020). Evidence synthesis

---

[2] Hampton et al. (2013) argue that big collaborative projects like the LTER program are inappropriate for solving grand challenges because of the homogeneity and centralisation of research methods and concepts, the lack of first-hand experience of environments due to automation, and the difficulty of obtaining serendipitous discoveries. I won't weigh in on this argument here, but it certainly adds more grist to the mill of open data and data synthesis advocates.

includes a number of ways of drawing different lines of evidence together, including systematic reviews, meta-analysis, and data synthesis. In general, evidence synthesis is used to systematically summarise and assess the current state of knowledge on a topic, often with a view to testing hypotheses on a more general level, developing new hypotheses, and providing guidance for decision making.

Like other types of evidence synthesis, data synthesis involves bringing together data from multiple primary studies to create a more general picture of a phenomenon or system. Yet data synthesis studies are distinctive in two ways. First, they are a kind of *synthetic modelling*, reusing data from multiple studies to develop complex, highly general models. Other kinds of evidence synthesis may also involve reusing data from multiple studies, but they don't use this data for modelling.[3] For example, meta-analysis reuses data from multiple studies to summarise evidence and test hypotheses, such as evaluating how robust an effect size is over many different studies. Philosophers studying meta-analysis have considered questions concerning evidence selection and evaluation (Stegenga 2011; Jukola 2015; Kovaka 2022). Data synthesis, in contrast, raises questions about model building, as I explore later in this paper.

Another distinctive feature of synthesis studies is their aim to produce *reusable aggregated* or *synthetic datasets*. Aggregated or synthetic datasets are datasets that incorporate data from multiple heterogenous studies. They are difficult, time-consuming and costly to produce, and there are few rewards for researchers who do so. This leads to a paucity of aggregated datasets, despite the importance of synthesis efforts for answering big questions. Researchers aim to enhance the usefulness of the few synthetic datasets they have by making these available for many researchers to reuse for their own specific research questions. Synthesis studies consequently aim to produce both a general model and a multipurpose synthetic dataset. Later in the paper I argue that these two goals create a trade-off specific to synthetic modelling, which I call the synthesis trade-off.

Data synthesis is a term originating in ecological research, and I focus on data synthesis studies in this field. However, many other fields have similar sorts of studies. For instance, data linkage, data mixes, or data mashups in the health sciences also bring together large datasets from disparate sources to generate both large-scale models and reusable synthetic data resources (Fleming et al. 2017; Tempini 2020). Similarly, climatology has long involved reanalysing data from many locations and sensor types to develop global models and aggregated datasets (Edwards 2010;

---

[3] Technically, meta-analysis and similar kinds of evidence synthesis do involve a sort of modelling, *data modelling*. That is, they transform data into a tractable form and use it to represent aspects of target systems and test hypotheses (Harris 2003; Leonelli 2019; Antoniou 2021; Bokulich and Parker 2021). In contrast to mere data modelling, data synthesis studies involve modelling in the sense that is more colloquial amongst scientists: engaging in more complex computational steps to produce something closer to a theoretical model, one that abstractly represents aspects of the world and can be used to explain and predict phenomena.

Bokulich and Parker 2021). My discussion of ecological data synthesis should therefore be understood within a broader context of novel approaches to gathering and manipulating big data from heterogenous sources.

To better understand data synthesis, I focus on a recent flagship case of a synthesis study performed on animal tracking data to answer a big question. Animal tracking data is the data produced when animal-borne devices are used to detect and record animal locations and movements, known as biologging, biotelemetry, or animal telemetry. Tracking devices include familiar technologies, such as radio transmitters, GPS devices, and accelerometers, as well as the less familiar, such as acoustic or light signalling devices. Devices may also include sensors for physiological or environmental parameters, such as heart rate sensors or depth sensors. Over time, tags have become smaller, more powerful, and more versatile, leading to an explosion of animal tracking data (Rutz and Hays 2009; Nathan et al. 2008; Brown et al. 2013; Börger et al. 2020; Williams et al. 2020; Nathan et al. 2022).

Given the abundance of tracking data, data synthesis has become a tantalising prospect to answer some of the big questions in ecology and conservation biology. For instance, a recent opinion paper argued that: "pooling data (across taxa, longer time periods or multiple locations) can reveal general patterns, aiding the design of particularly effective conservation strategies." (Rutz 2022, 221) Another review paper argued that "Efficient data sharing and archiving across many studies and authors will be key to answer the big questions in movement ecology, for example global responses to environmental change […]" (Williams et al. 2020, 195). This connection between synthesising animal tracking data and answering big questions is exemplified in a recent and very prominent synthesis study using animal tracking data, the Retrospective Analysis of Antarctic Tracking Data (RAATD) project, run by the Scientific Committee on Antarctic Research (SCAR).

The RAATD project involved producing and then using an aggregated dataset of over 20 years' worth of tracking data from Antarctic marine predators. The dataset and its production are described in the data paper "The retrospective analysis of Antarctic tracking data project", published by Yan Ropert-Coudert and 79 colleagues in the journal *Scientific Data* (Ropert-Coudert et al. 2020). The first and exemplary use of the RAATD dataset is presented in the paper "Tracking of marine predators to protect Southern Ocean ecosystems", published in *Nature* by Mark Hindell and 80 colleagues (Hindell et al. 2020). These two papers, which I refer to as the data paper and the Nature paper, together present the main elements of the RAATD project and exemplify the dual-purpose nature of synthesis studies.

The RAATD project brought together animal tracking data that had been originally collected for a variety of purposes by over 70 data contributors in the years 1991-2016. The dataset

includes 17 predator species: 12 seabird species (penguins, albatrosses and petrels) and 5 mammal species (seals and whales). Following data standardisation and filtering, the dataset contained data for 4060 individuals, with more than 2.9 million at-sea locations. This standardised and filtered data is held at the SCAR Antarctic Biodiversity Portal, which flows to OBIS (Ocean Biogeographic Information System) and GBIF (Global Biodiversity Information Facility), two of the largest biodiversity data collections in the world.

The Nature paper used the RAATD dataset to answer a decidedly big question: identifying areas in the Southern Ocean requiring conservation. The authors focus on so-called areas of ecological significance (AESs), areas worthy of protection due to their high biological productivity, high biodiversity, or importance for particular life-stages of species. Biological productivity and biodiversity can be difficult to measure in remote and large areas. The location of multiple predator species with different diets and foraging strategies serves as a proxy, indicating the presence of large amounts of prey and various prey types, and therefore high primary productivity and biodiversity. Hence, researchers can use tracking data to identify where predators are foraging and thereby locate AESs. Specifically, the RAATD study involved constructing habitat selection models for the 17 species under study, then combining these to provide an overall model of habitat importance across the Antarctic and sub-Antarctic. Using these models, researchers could then determine the exposure of AESs to stressors by mapping tracking data-based AESs against fisheries data and climate change models. This helps to pinpoint areas in need of special protection.[4]

Beyond its first use in the Nature paper, the RAATD dataset is open for reuse by other ecologists and conservation biologists. As the authors of the data paper state, "the dataset will be available for re-use to help address emerging research questions or pressing conservation issues" (Ropert-Coudert et al. 2020, 7). The RAATD dataset is focused on spatial occurrence and is therefore particularly suitable for analyses of species distribution, habitat choice, and resource use. The inclusion of different species, developmental stages, time periods, and geographic areas means that researchers can investigate many aspects of species distribution and resource use; the ready correlation with environmental and other types of data also facilitates the study of causes and consequences of animal movement as well as risks due to climate change and anthropogenic disturbances.

The RAATD project exemplifies the main features of data synthesis studies: data are gathered from multiple primary studies and processed to produce an aggregated dataset, this dataset is

---

[4] The RAATD findings have been used to inform conservation and management strategies. In 2020, the project and its recommendations were discussed at the 39th Meeting of the Scientific Committee of the Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR), an international commission responsible for setting conservation measures in the Antarctic (SC-CAMLR 2020).

used to produce a general model to answer a big question, and the dataset is also made available for further reuse. In the rest of the paper, I examine this case in light of the philosophy of ecological modelling and the philosophy of scientific data.

3   Building General Ecological Models

Data synthesis studies develop highly general models. Philosophers of ecology have long discussed trade-offs and strategies involved in building models of complex ecological systems, and in particular have argued that generality comes at the cost of either realism or precision (Levins 1966). In this section I use this discussion to make sense of trade-offs and strategies of synthetic modelling. By analysing the RAATD study, I argue that data synthesis involves a *kludging* strategy, making a complex series of iterative adjustments with respect to various desiderata at different points in the model building process.

In a now canonical paper, Richard Levins discusses different ways in which biologists develop models of complex systems (Levins 1966). Levins highlights three desiderata of models: generality, realism and precision. First, biologists want models that are general, that is, that apply to many different systems or instances in the world. They also want realistic models; realism can be equated roughly with accuracy, such that either the structure or the output of the model matches the world (Weisberg 2006). Finally, biologists want precise models, such that the model represents relationships quantitatively in exact mathematical forms rather than vaguely and qualitatively, accounts for factors that have small effects or rare large effects, and represents lower-level phenomena that feed into more general patterns (Levins 1966, 429–30).

Levins argues that it is impossible to maximise all three desiderata at once while still creating a useful model: a model that can be understood and used to explain and make predictions about ecological systems (Odenbaugh 2003; Matthewson 2011; Elliott-Graves 2018). When building models, then, biologists must adopt one of three different strategies, sacrificing either generality, realism, or precision. Some ecologists choose to sacrifice generality in order to develop very precise and realistic models of specific systems. Levins's examples of this strategy come from entomology and fisheries research which draw on large quantities of data for a specific area to develop multivariate mathematical models of population dynamics (Levins 1966, 422). Other ecologists sacrifice realism to generality and precision, developing general numerical equations that are also highly idealised, omitting many features of the systems being modelled. Many philosophers are familiar with this strategy, epitomised in the ubiquitous Lotka-Volterra models of population dynamics. In contrast, Levins advocates for the strategy of sacrificing precision in favour of highly general and realistic models: primarily qualitative, graphical models that represent the dynamics of

many different systems but make no pretensions to numerical precision or including smaller or more rare effects and lower-level phenomena. Nevertheless, Levins remains a pluralist, such that the adoption of a particular strategy depends on the specific research questions and purposes of the study, as well as the target system and the resources to hand (Odenbaugh 2006; Matthewson 2011; Goldsby 2013; Elliott-Graves 2020).

Levins focuses on building theoretical models, and many philosophers have followed suit (Odenbaugh 2003; Weisberg 2006; Matthewson 2011; Elliott-Graves 2018). However, Levins clearly thinks that more data-driven approaches to model building also adopt one of the three strategies, evidenced by his references to the entomological and fisheries models. Do these strategies capture the sort of modelling involved in synthesis studies? Is there a trade-off between generality, realism and precision when synthesising data to answer big questions, and if so, which model building strategy is most suitable?

The RAATD synthesis study created a model of habitat importance across the Southern Ocean. To do so, they used habitat selection models. Habitat selection models combine tracking data with environmental data to identify what factors are correlated with the tracked animals' movements. Other areas where the same sort of environmental conditions occur, and which are accessible to animals, can then be taken as likely spots for non-tracked animals to be found. In this way, habitat selection models enable researchers to extrapolate from their dataset to entire populations or species as well as across large geographic and temporal ranges. In the RAATD study, they then combined these species-specific habitat selection models to identify areas of importance for all 17 predator species studied.

The model of habitat importance produced in the RAATD synthesis study is certainly general: it covers many ecosystems in the Southern Ocean, many species with different foraging strategies, different seasons and life history stages, and so on. The RAATD model also seems to be fairly realistic: the structure and output of the model do presumably match the areas where predators forage. Realism is especially important given the ultimate goal of supporting conservation, because accurately representing a system is typically instrumental to successfully intervening on it. Finally, the RAATD model is fairly imprecise, since it ignores factors with small effects, is based on averages and maxima rather than representing all values, and ignores lower-level phenomena like foraging behaviour to focus on coarser-grained patterns of movement and location. Again, imprecision seems to fit with the goal of providing input for a large-scale conservation strategy, although precision can be important for more specific conservation interventions (Elliott-Graves 2016).

So, the RAATD model looks like the outcome of Levins' preferred third strategy, sacrificing precision for generality and realism. However, things get more complicated when we pay attention to the model building practice rather than just the model, as Levins instructs us to do (Weisberg 2006). Together, the data paper and the Nature paper (along with their respective appendices) describe in great detail the process of model building. Examining these details reveals that researchers didn't follow any of Levins' three model-building strategies. Instead, they performed a complex series of adjustments for various desiderata at different points in the modelling process, a *kludging* strategy.

Kludging, or developing kludges, is a strategy used in synthetic biology as well as other engineering disciplines (O'Malley 2009; 2011). Kludging involves making many iterative, often ad hoc adjustments aimed at making things work—in contrast to rational or efficient design, where systems are designed at the outset and then built in a linear, controlled fashion. Maureen O'Malley summarises the kludging strategy as such: "It does not matter how inelegant the process is to get there, or how inefficient the relationships between some of the componentry and circuitry. If the system works, that is the ultimate vindication of construction." (O'Malley 2009, 382) O'Malley argues that kludging is a common practice across biology and a good strategy for gaining knowledge and building functioning systems when faced with context-dependence and complexity.[5] Data synthesis studies do deal in complex combinations of heterogenous elements. It is therefore perhaps unsurprising that synthetic modelling involves a kludging strategy, making iterative, ad hoc adjustments with respect to modelling desiderata to develop a final model that works.

The RAATD modelling process exemplifies a kludging strategy. One sort of adjustment seen at several stages throughout the model building process was to sacrifice several desiderata at once. We see this in selective data collection and data deletion, where elements and areas of animal movement are ignored. For example, researchers did not aggregate any ecological, physiological, or genetic data from original studies, nor did they include movement data from devices that record depth or acceleration. Researchers also excluded or deleted data that had to do with on-land movement. This side-lines realism, since it results in a model that does not structurally represent all aspects of animal behaviour. It also sacrifices some generality, as it means that the model does not cover all of the many different systems in which movement is involved—foraging and social systems, for instance, or terrestrial and marine.

At other steps in the study, researchers made adjustments to increase several desiderata at once. For example, during data processing researchers removed inaccurate and imprecise

---

[5] Kludging may be even more widespread in model building. For example, similarly complicated and laborious processes of adjustment are found in the construction of data models in high energy physics (Antoniou 2021).

predictions of animals' locations and movements, deleting tracks that were too "noisy" (i.e., location estimates were too irregularly spread out over space), or too irregularly spaced over time, as well as data that implied an animal was travelling unrealistically fast (over 10 ms^{-1} for penguins and marine mammals and over 30 ms^{-1} for flying seabirds). Such data deletion increases the accuracy and precision of the resulting model, as well as creating a smaller, neater dataset that is easier to incorporate into a general model.

At still other stages, there were increases and decreases with respect to several desiderata. This is evident in the process of developing general habitat choice models for the whole year and for all members of a species, which relied on using maxima and averages. For example, researchers estimated the accessibility of locations for a species by generalising from the maximum distance travelled by tracked individuals to all individuals in the species. Similarly, the climatological models that researchers used to infer tracked individuals' habitat choice (and to generalise this to all individuals in the Southern Ocean) were based on monthly averages and patterns interpolated onto a grid. The use of maxima and averages involves sacrificing precision and realism in favour of generality. Yet researchers simultaneously made adjustments to increase realism and precision. For example, they developed separate habitat choice models for different life history stages, which provides a more precise picture of animals' movement patterns throughout the year by reducing the noise created by seasonal variation. In addition, researchers individually adjusted the breeding season timing, which increases the accuracy with which animals' movements are represented.

Kludging is also seen in the process of creating a multi-species model of habitat importance by averaging the species-specific habitat importance estimates. This averaging procedure may look like a simple strategy of sacrificing precision. However, researchers chose to separately average the habitat importance for Antarctic and sub-Antarctic species, in order to ensure that the relatively species-poor Antarctic region was still represented in the model. In addition, the multi-species model of habitat importance is still connected to the species-specific habitat choice models. This means that information about environmental covariates and how they affect animal movement was retained and could be reanalysed, for instance to predict the consequences of climate change. Adjustments like separately averaging habitat importance and preserving links to environmental variables show that the synthetic modelling process doesn't simply sacrifice precision, since some information about factors that have smaller or more rare effects as well as lower-level phenomena is retained.

Altogether, although the RAATD model may look like the output of Levins' preferred strategy of sacrificing precision for generality and realism, it in fact resulted from a more elaborate series of adjustments. Different desiderata were favoured throughout the process of data

collection, standardisation, filtering, and model construction, and this was done in a highly project-specific manner. In other words, data synthesis studies produce general models by kludging, performing a complex series of adjustments aimed at making heterogenous data work together to create a model that is general, accurate, and precise enough to serve its goal. This is true even if the goal is a general, realistic but imprecise model that can guide international conservation strategies; the goal doesn't determine the means, and kludging is an appropriate strategy to deal with the multiple, heterogenous elements involved in synthetic model building.

4   The Synthesis Trade-Off

As well as being model building practices, data synthesis studies are also data wrangling exercises. In this section I use recent accounts of scientific data as relational and historical to understand how synthesis studies work. I argue that synthesis studies involve what I call the synthesis trade-off, the trade-off between synthesising data for a particular model, on the one hand, and creating a dataset that can be used for multiple different projects, on the other.

Recent philosophy of data emphasises the *relational* nature of data, such that data serve a representational or evidential role only in relation to particular (potential) uses (Leonelli 2016). The relational view of data implies that data depend on the contexts in which they may be used, including research questions, goals, social settings, infrastructures, and so on. In addition, data rely on the practices of collection, processing, storage and dissemination that enable them to serve evidential roles in these contexts, forming *enriched evidence* (Boyd 2018). A corollary of the relational view is that data must be evaluated according to their *adequacy for purpose* (Bokulich and Parker 2021). Because they only represent the world in relation to specific contexts of use, data cannot be evaluated as good or bad representations simply based on internal features like errors or noise. Instead, data must be evaluated according to whether and to what extent they can be used to achieve specific circumscribed epistemic or practical goals. Adequacy for purpose in turn depends on other dimensions, such as the representational target, data users, methodologies, background circumstances, and what other resources are available.

Philosophers also highlight how data are created, transported and transformed in order to serve various representational or evidential roles, making data *historical* entities. For example, data depend on *data journeys* (Leonelli 2016; 2020; Leonelli and Tempini 2020), "the movement of data from their production site to many other sites in which they are processed, mobilised and re-purposed." (Leonelli 2020, 27) This concept highlights the importance of mobility for data, which can be used to generate knowledge only by moving along various routes from data collection through processing, analysis, storage, packaging, publication, reuse, and so on. Data journeys also

include the idea that data is mutable. Data are transformed to fit different uses as they travel, including data cleaning and processing as well as modifications to suit storage, dissemination or reuse. As entities that change over time, data can be understood as lineages (Leonelli 2020). This paves the way to think about the *phylogeny of data*: the history of reuse and repurposing of datasets can be represented as bifurcating lineages and analysed in terms of how data have been adapted over time to be adequate to particular purposes and contexts (Bokulich and Parker 2021).

The relational and historical accounts of data help to understand how data is repurposed for synthesis studies. Synthesis studies are a kind of data journey, one involving multiple steps of data collection, sharing, processing, analysis, and interpretation, and potentially changing the identity of the data en route (Tempini 2020, 259). Phylogenetically, data synthesis studies could be represented as (parts of) separate lineages converging to form a single lineage. In particular, synthesis studies involve *data integration*, "the activity of making comparable different data types from a huge variety of potentially inconsistent sources." (O'Malley and Soyer 2012, 61) And, as Sabina Leonelli points out, "Data integration happens locally to solve specific problems" (Leonelli 2016, 191).

In the RAATD study, data was collected and processed to suit the goal of modelling habitat choice and identifying AESs. For example, researchers were selective in data collection, focusing on at-sea movement data and relevant metadata and excluding other sorts of movement, ecological, physiological, and genetic data. Some of this excluded data could be relevant for studying habitat use. For instance, temperature, pressure, or other ecological data from local sensors can provide a highly localised picture of the conditions experienced by an animal. Similarly, gut content and scat analyses can indicate what an animal has actually eaten over a short period of time. Acceleration and depth data can also help to investigate where in the water column animals are foraging and what resources they may be consuming. However, all of these additional types of data are not consistently collected across all studies, and there are many different factors to measure and methods to measure them. Focusing just on tracking data helps to minimise the heterogeneity in the dataset and while still including data from many different studies. Researchers could thereby obtain a broader coverage of animal movement only by forgoing information about local ecological and organismal context.

Synthesis studies, like other sorts of data journeys (or lineages), are thus shaped by and adapted to their specific goals and resources: the goal of developing a general model to answer a big question using available data. Yet, as I discussed in Section 2, synthesis studies have an additional goal. Due to the context of limited data processing resources and the resource-intensive nature of synthesis work, researchers aim to produce aggregated datasets that can be reused for

multiple projects, not just for one particular model. In other words, synthesis studies aim to make data travel more widely than their initial reuse, joining the broader open data movement aiming at "the creation, dissemination and aggregation of vast datasets to facilitate their repurposing for as wide a range of goals as possible." (Leonelli 2020, 18)

The existence of two goals creates a tension for synthesis studies. On the one hand, researchers aim to synthesise data to suit a particular modelling project. This is best achieved by intensively tailoring the dataset to the resulting model, for instance by selecting only highly relevant data and including model-specific assumptions into the data processing stages. On the other hand, researchers aim to synthesise data so that it can be used for a wide variety of different projects. This is best achieved by including many different sorts of data in the dataset and allowing for considerable freedom in data processing to account for different assumptions and modelling approaches. Given their dual purposes, synthesis studies therefore face what I call the *synthesis trade-off*: the trade-off between synthesis to suit a particular model and synthesis to facilitate a plurality of reuses.

Like the Levinsian modelling trade-off, the synthesis trade-off can be dealt with by adopting different strategies. Two strategies take either extreme: (1) creating highly specialised synthetic datasets that are easy to use and reuse for a particular modelling goal, but don't enable other sorts of reuse; and (2) creating large and rich datasets that can be used for many purposes but in each case will require considerable work getting to know the data and adjusting it to suit the project at hand. A third strategy makes a compromise between these extremes: (3) creating semi-specialised, flexible datasets that can be adjusted to suit a restricted range of projects. I call this third strategy the *flexibility compromise*.

The synthesis trade-off and the flexibility compromise are both evident in the RAATD synthesis study. During data processing, researchers used state-space models to transform tracking data into location estimates at regular time intervals; this smooths the inferred movement and thereby accounts for some measurement error in the tracking data. The time intervals for the state-space models were chosen based on typical sampling frequencies for different tracking devices: 1 hour for GPS tags, 2 hours for PTT (Platform Terminal Transmitter) tags, and 12 hours for GLS (Global Location Sensor) tags.

The authors of the data paper acknowledge that their choice of time intervals may lead to undersampling of data that had been collected at a higher frequency, especially for birds tracked with high-frequency GPS. They also note that different time intervals might be needed for different research questions. For instance, relatively coarse time intervals are appropriate for estimating where animals spend considerable time foraging—the goal of the Nature paper. In contrast, finer

time intervals might be needed to identify travelling routes or specific behaviours. The use of state-space models exemplifies the synthesis trade-off.[6] On the one hand, using typical sampling frequency created a dataset that could be easily used to model habitat choice and estimate habitat importance. On the other, it limited the potential for using the RAATD dataset to answer other questions about Antarctic predator behaviour and resource use.

The authors of the RAATD study recognised this trade-off and responded with the flexibility compromise. Although they recommend using the filtered data, they also provide the pre-filtered data along with the filtering code so that other users can tailor the state-space model to their own purposes. By including this flexibility, the RAATD synthesis study is better able to satisfy its dual goal of developing a general model and providing a new data resource. The flexibility is nevertheless not boundless. There are many questions to ask about the at-sea movements of Antarctic predators, and many general models to be constructed that capitalise on this standardised data. Yet the dataset clearly doesn't suit other sorts of projects, such as studies of on-land movements, social or reproductive interactions, fine-scale motor behaviour, or studies requiring data on animals' physiology, genetics or experienced environment. The flexible dataset is therefore a selective resource that enables some research questions and modelling approaches but restricts others.

This selective flexibility highlights the "scaffolded relationality" of scientific data, that is, how its epistemic value depends on and is shaped by the infrastructures through which data is processed and shared for reuse (Tempini 2020, 257). When adopting the flexibility compromise, researchers must therefore carefully reflect on how assumptions and expected uses are built into an open synthetic dataset in a way that scaffolds only some of data's potential uses as epistemic resources. It is also important to bear in mind that the flexibility compromise is a compromise. Just as kludging is a strategy to deal with (but not resolve) the Levinsian modelling trade-off, the flexibility compromise is a strategy to deal with, but not resolve, the synthesis trade-off. Faced with the inability to simultaneously maximise ease-of-use and diversity-of-uses for a synthetic dataset, the flexibility compromise provides distinct epistemic benefits, facilitating the pursuit of a modest variety of research questions with comparatively little effort. Researchers adopting the flexibility

---

[6] The choice of time intervals for state-space modelling also hints at another realm of trade-offs faced by researchers in movement ecology (and in other disciplines), to do with practical constraints. Which data are available for synthesis is dependent on technological developments, researchers' and funders' interests, how animals move, ethical restrictions, and many other factors. These factors often lead to further trade-offs for data, such as the trade-off between the length of time tracked and the frequency of tracking observations, which depends on the cost and size of tracking device (Trappes 2023). Thanks to a reviewer for suggesting this as an additional dimension of trade-offs faced in synthesis studies.

strategy must however decide what sorts of projects they would like their data resources to support and how to best facilitate those projects.

5 Flexible Open Data

The synthesis trade-off and flexibility compromise are especially acute for synthesis studies, but they are also evident in broader open data efforts. Apart from transparency and reproducibility, one major argument for open data is to enable reuse, whether through synthesis or other sorts of re-analysis. In this section I consider discussions in movement ecology around opening up tracking data, showing how researchers have adopted the flexibility compromise in their proposals to standardise open tracking data. This provides an example for how recognising the synthesis trade-off and flexibility compromise can help to understand not only synthesis studies but also the broader open data movement.

There are many efforts to share movement data online, especially within collaborative networks of researchers working on similar systems or organisms such as the Arctic Animal Movement Archive (Davidson et al. 2020; Nathan et al. 2022). In 2008 the foremost global animal tracking data resource, Movebank, was founded and now hosts a database as well as tools for processing and analysing movement data (Kays et al. 2021). Although many researchers use Movebank to privately archive their data and access the analytical tools, comparatively few publish their data publicly in the Movebank Data Repository (Max Planck Society 2021), and even when published, biotelemetry data are difficult to find and reuse (Campbell et al. 2016; Williams et al. 2020; Sequeira et al. 2021; Rutz 2022).

Heterogeneity has been identified as one of the key challenges to open data and data reuse in movement ecology (Campbell et al. 2016; Sequeira et al. 2021; Kays et al. 2021). Some of this heterogeneity comes from animal movement itself: different species, developmental stages, individuals, behaviours, times of day, environmental conditions, and so on lead to sometimes radically different patterns of movement. Heterogeneity also arises from data collection methods, such as different tracking devices, manufacturers, and device deployment.

Such heterogeneity makes sharing and reusing biologging data very difficult. First, it is challenging and time consuming to incorporate heterogenous data into a single repository while avoiding errors (Campbell et al. 2016). Second, finding and accessing relevant data is impeded by differing terminology and conflicting or missing metadata (Sequeira et al. 2021). Third, it can be difficult to understand a heterogenous dataset without insider knowledge about how it was produced (Leonelli 2016). Finally, conflicting, inconsistent, and messy data in heterogenous open datasets are difficult to process and analyse (Campbell et al. 2016). These problems disincentivise

sharing and using open data (Hampton et al. 2013; Roche et al. 2014; 2015; Rutz 2022). In response to such problems, researchers first called for and then developed proposals for the standardisation of biotelemetry data, arguing that reusable open data is essential for more efficient and effective research in ecology and conservation biology (Campbell et al. 2016; Sequeira et al. 2021).[7]

Standardisation efforts for open data face a tension much like the synthesis trade-off. On the one hand, greater standardisation makes data easier to include in a database as well as easier to find and reuse. On the other, standardisation shapes data in particular ways based on particular envisaged uses. As a number of theorists have argued with respect to biodiversity and biomedical data, standardisation can significantly affect data and its uses, for instance through de-emphasising types of biodiversity that are hard to classify, removing detailed location descriptions, or creating uncertainty in locations (Bowker 2000; Shavit and Griesemer 2011; Leonelli and Tempini 2021). Although many open data efforts concentrate on either rigid standardisation or unstandardised data sharing, standardisation proposals for animal tracking data have responded to the trade-off by making the flexibility compromise.

In their proposal, Sequeira et al. (2021) include many levels of processing, each of which is shared with open code and documentation. For instance, sharing interpolated data (the products of state-space models) makes it easy to conduct synthesis studies without having to create state-space models for huge amounts of data. Yet, as the authors put it prosaically, "There are many different ways to apply state-space models to data." (Sequeira et al. 2021, 1002) To facilitate other kinds of reuse, the authors recommend also sharing un-interpolated data with the code to enable researchers to construct their own state-space models to suit their research questions. Similarly, presenting tracking data in a grid format with specific spatiotemporal resolution makes it easier to link movement and environmental factors. The authors suggest picking a spatiotemporal resolution that suits available and commonly used environmental data, such as the monthly satellite readouts of chlorophyll-*a*. However, they also recommend sharing the code used to grid the data, so that researchers can adjust the spatial and temporal resolution to suit their own needs. But again, the flexibility is still limited. The processing steps favour projects relating to spatial distribution and are less suitable for studying other phenomena, such as behaviours related to depth, social interaction, or more fine-grained motor behaviour. The flexibility of the proposed standardisation thus makes for open tracking data that can be easily used for some purposes and fairly readily adjusted to suit a limited number of other projects.

---

[7] The proposal by Sequeira et al. (2021) is the first and to date only multipurpose standardisation framework for animal tracking data. However, the International Bio-Logging Society, founded in 2016, has set standardisation as one of its key objectives and formed a data standardisation working group (Newman, Cagnacci, and Davidson 2019).

Beyond movement ecology, the flexibility compromise is a likely strategy for data sharing in many different contexts. For instance, it seems that a similar sort of compromise has been reached in data mashup infrastructure in the health sciences, which incorporate the flexibility to make project-specific combinations of data from different sources despite the fact that this flexibility makes it more difficult to perform these mashups (Tempini 2020). The flexibility compromise also resonates well with philosophical accounts of data. First, it turns on the way that data are processed to suit the purposes for which the data can and will be used, emphasising the relationality of data. Second, by including multiple levels of processed and analysed data, along with the code for processing and analysis, flexible open data makes the historical nature of data particularly explicit. The flexibility compromise also makes apparent the complexities involved in data journeys or lineages, in which data are used for many different purposes and in the process taken and shaped in different directions and with different consequences.

6 Conclusion

Big questions loom large in ecology and conservation biology. Aside from large collaborations and citizen science, researchers have identified the synthesis of ecological data as a promising way to address big questions. In this paper I considered this approach of data synthesis studies, focusing on the field of movement ecology and a specific synthesis study of animal tracking data.

Data synthesis studies are a kind of evidence synthesis, but they have two distinctive goals: creating general models and creating a reusable aggregated dataset. To better understand these practices, I examined the trade-offs and strategies involved in data synthesis. I found that researchers adopt a kludging strategy to deal with the Levinsian trade-off between generality, precision and realism, engaging in a complex sequence of adjustments in order to produce a model that serve their purposes. In addition to the modelling trade-off, researchers performing synthesis studies also face the synthesis trade-off: making data easier to synthesise for a particular project, on the one hand, and making data suitable for reuse for other projects, on the other. The synthesis trade-off is faced by biologists across the board with the rise of open data. It is especially evident in ecology and conservation biology, where the pressing nature of the big questions and the limitations on funding in many parts of the world hasten the move towards more open data and more synthesis projects.

Movement ecologists have taken up the flexibility compromise as a promising strategy for dealing with the synthesis trade-off. Rather than sharing just one very specific standardised dataset that is easy to use but highly restricted, or one very unstandardised, generic dataset that is difficult to use but less restricted to particular purposes, researchers are sharing multiple layers of processed

and filtered data along with adjustable code. The more processed datasets are easier to use for particular questions, but for other scenarios researchers can still access the minimally processed data and adjust the processing and analysis code to suit their purposes. The flexibility compromise is still a compromise: the data is not suitable for every purpose, and certain uses of the data are privileged because using the pre-processed data requires more effort and additional justification. The increasing availability of flexible open datasets is therefore likely to have epistemic consequences in terms of the kinds of research questions asked and the kinds of projects pursued, paving the way for some lines of research and leaving others less well-travelled.

Data synthesis studies and the flexibility compromise fit well with existing philosophical accounts of scientific data. They highlight both the relationality and historicity of data, with data being transformed and transported to serve certain purposes. As a data journey or a data lineage, however, data synthesis studies, and open datasets more broadly, are tangled and complex. Opening up ecological data to facilitate data reuse, and especially to enable synthesis studies to answer big questions, means facing up to the synthesis trade-off, and that means creating flexible data resources that can be shaped and pulled in different directions to suit a variety of projects. These messy but fruitful practices will likely only become more prevalent as ecologists and conservation biologists move towards more data-intensive research and as the big questions about global change and conservation become ever more pressing.

References

Antoniou, Antonis. 2021. "What Is a Data Model?" *European Journal for Philosophy of Science* 11 (4): 101. https://doi.org/10.1007/s13194-021-00412-2.

Bokulich, Alisa, and Wendy Parker. 2021. "Data Models, Representation and Adequacy-for-Purpose." *European Journal for Philosophy of Science* 11 (1): 31. https://doi.org/10.1007/s13194-020-00345-2.

Börger, Luca, Allert I. Bijleveld, Annette L. Fayet, Gabriel E. Machovsky-Capuska, Samantha C. Patrick, Garrett M. Street, and Eric Vander Wal. 2020. "Biologging Special Feature." *Journal of Animal Ecology* 89 (1): 6–15. https://doi.org/10.1111/1365-2656.13163.

Bowker, Geoffrey C. 2000. "Biodiversity Datadiversity." *Social Studies of Science* 30 (5): 643–83. https://doi.org/10.1177/030631200030005001.

Boyd, Nora Mills. 2018. "Evidence Enriched." *Philosophy of Science* 85 (3): 403–21. https://doi.org/10.1086/697747.

Brown, Danielle D., Roland Kays, Martin Wikelski, Rory Wilson, and A. Peter Klimley. 2013. "Observing the Unwatchable through Acceleration Logging of Animal Behavior." *Animal Biotelemetry* 1: 20. https://doi.org/10.1186/2050-3385-1-20.

Campbell, Hamish A., Ferdi Urbano, Sarah Davidson, Holger Dettki, and Francesca Cagnacci. 2016. "A Plea for Standards in Reporting Data Collected by Animal-Borne Electronic Devices." *Animal Biotelemetry* 4 (1): 1. https://doi.org/10.1186/s40317-015-0096-x.

Carpenter, Stephen R., E. Virginia Armbrust, Peter W. Arzberger, F. Stuart Chapin, James J. Elser, Edward J. Hackett, Anthony R. Ives, et al. 2009. "Accelerate Synthesis in Ecology and Environmental Sciences." *BioScience* 59 (8): 699–701. https://doi.org/10.1525/bio.2009.59.8.11.

Davidson, Sarah C., Gil Bohrer, Eliezer Gurarie, Scott LaPoint, Peter J. Mahoney, Natalie T. Boelman, Jan U. H. Eitel, et al. 2020. "Ecological Insights from Three Decades of Animal Movement Tracking across a Changing Arctic." *Science* 370 (6517): 712. https://doi.org/10.1126/science.abb7080.

Devictor, Vincent, Robert J. Whittaker, and Coralie Beltrame. 2010. "Beyond Scarcity: Citizen Science Programmes as Useful Tools for Conservation Biogeography: Citizen Science and Conservation Biogeography." *Diversity and Distributions* 16 (3): 354–62. https://doi.org/10.1111/j.1472-4642.2009.00615.x.

Di Cecco, Grace J, Vijay Barve, Michael W Belitz, Brian J Stucky, Robert P Guralnick, and Allen H Hurlbert. 2021. "Observing the Observers: How Participants Contribute Data to INaturalist and Implications for Biodiversity Science." *BioScience* 71 (11): 1179–88. https://doi.org/10.1093/biosci/biab093.

Edwards, Paul N. 2010. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming.* Cambridge, Mass: MIT Press.

Efstathiou, Sophia. 2016. "Is It Possible to Give Scientific Solutions to Grand Challenges? On the Idea of Grand Challenges for Life Science Research." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 56 (April): 48–61. https://doi.org/10.1016/j.shpsc.2015.10.009.

Elliott-Graves, Alkistis. 2016. "The Problem of Prediction in Invasion Biology." *Biology & Philosophy* 31 (3): 373–93. https://doi.org/10.1007/s10539-015-9504-0.

———. 2018. "Generality and Causal Interdependence in Ecology." *Philosophy of Science* 85 (5): 1102–14. https://doi.org/10.1086/699698.

———. 2020. "The Value of Imprecise Prediction." *Philosophy, Theory, and Practice in Biology* 12 (20220112). https://doi.org/10.3998/ptpbio.16039257.0012.004.

Fleming, Lora, Niccolò Tempini, Harriet Gordon-Brown, Gordon L. Nichols, Christophe Sarran, Paolo Vineis, Giovanni Leonardi, et al. 2017. "Big Data in Environment and Human Health." In *Oxford Research Encyclopedia of Environmental Science*, by Lora Fleming, Niccolò Tempini, Harriet Gordon-Brown, Gordon L. Nichols, Christophe Sarran, Paolo Vineis, Giovanni Leonardi, et al. Oxford University Press. https://doi.org/10.1093/acrefore/9780199389414.013.541.

Goldsby, Michael. 2013. "The 'Structure' of the 'Strategy': Looking at the Matthewson-Weisberg Trade-off and Its Justificatory Role for the Multiple-Models Approach." *Philosophy of Science* 80 (5): 862–73. https://doi.org/10.1086/673728.

Hampton, Stephanie E., Sean S. Anderson, Sarah C. Bagby, Corinna Gries, Xueying Han, Edmund M. Hart, Matthew B. Jones, et al. 2015. "The Tao of Open Science for Ecology." *Ecosphere* 6 (7): art120. https://doi.org/10.1890/ES14-00402.1.

Hampton, Stephanie E., Matthew B. Jones, Leah A. Wasser, Mark P. Schildhauer, Sarah R. Supp, Julien Brun, Rebecca R. Hernandez, et al. 2017. "Skills and Knowledge for Data-Intensive Environmental Research." *BioScience* 67 (6): 546–57. https://doi.org/10.1093/biosci/bix025.

Hampton, Stephanie E, Carly A Strasser, Joshua J Tewksbury, Wendy K Gram, Amber E Budden, Archer L Batcheller, Clifford S Duke, and John H Porter. 2013. "Big Data and the Future of Ecology." *Frontiers in Ecology and the Environment* 11 (3): 156–62. https://doi.org/10.1890/120103.

Harris, Todd. 2003. "Data Models and the Acquisition and Manipulation of Data." *Philosophy of Science* 70 (5): 1508–17. https://doi.org/10.1086/377426.

Hindell, Mark A., Ryan R. Reisinger, Yan Ropert-Coudert, Luis A. Hückstädt, Philip N. Trathan, Horst Bornemann, Jean-Benoît Charrassin, et al. 2020. "Tracking of Marine Predators to Protect Southern Ocean Ecosystems." *Nature* 580 (7801): 87–92. https://doi.org/10.1038/s41586-020-2126-y.

Inkpen, S. Andrew. 2016. "Like Hercules and the Hydra: Trade-Offs and Strategies in Ecological Model-Building and Experimental Design." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 57 (June): 34–43. https://doi.org/10.1016/j.shpsc.2016.02.019.

Jukola, Saana. 2015. "Meta-Analysis, Ideals of Objectivity, and the Reliability of Medical Knowledge." *Science & Technology Studies* 28 (3): 101–21. https://doi.org/10.23987/sts.55344.

Kays, Roland, Sarah C. Davidson, Matthias Berger, Gil Bohrer, Wolfgang Fiedler, Andrea Flack, Julian Hirt, et al. 2021. "The Movebank System for Studying Global Animal Movement and Demography." *Methods in Ecology and Evolution*, December, 2041-210X.13767. https://doi.org/10.1111/2041-210X.13767.

Kovaka, Karen. 2022. "Meta-Analysis and Conservation Science." *Philosophy of Science*, June, 1–21. https://doi.org/10.1017/psa.2022.68.

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study.* Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226416502.001.0001.

———. 2019. "What Distinguishes Data from Models?" *European Journal for Philosophy of Science* 9 (2): 22. https://doi.org/10.1007/s13194-018-0246-0.

———. 2020. "Learning from Data Journeys." In *Data Journeys in the Sciences*, edited by Sabina Leonelli and Niccolò Tempini, 1–24. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-37177-7_1.

Leonelli, Sabina, and Niccolò Tempini, eds. 2020. *Data Journeys in the Sciences.* Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-37177-7.

———. 2021. "Where Health and Environment Meet: The Use of Invariant Parameters in Big Data Analysis." *Synthese* 198 (S10): 2485–2504. https://doi.org/10.1007/s11229-018-1844-2.

Levins, Richard. 1966. "The Strategy of Model-Building in Population Biology." *American Scientist* 54 (4): 421–31.

Matthewson, John. 2011. "Trade-Offs in Model-Building: A More Target-Oriented Approach." *Studies in History and Philosophy of Science Part A* 42 (2): 324–33. https://doi.org/10.1016/j.shpsa.2010.11.040.

Max Planck Society. 2021. "The Movebank Data Repository." Movebank. 2021. https://www.move-bank.org/cms/movebank-content/data-repository.

Mitchell, Sandra D. 2003. *Biological Complexity and Integrative Pluralism*. Cambridge: Cambridge University Press.

Nakagawa, Shinichi, Adam G. Dunn, Malgorzata Lagisz, Alexandra Bannach-Brown, Eliza M. Grames, Alfredo Sánchez-Tójar, Rose E. O'Dea, et al. 2020. "A New Ecosystem for Evidence Synthesis." *Nature Ecology & Evolution* 4 (4): 498–501. https://doi.org/10.1038/s41559-020-1153-2.

Nathan, Ran, W. M. Getz, E. Revilla, M. Holyoak, R. Kadmon, D. Saltz, and P. E. Smouse. 2008. "A Movement Ecology Paradigm for Unifying Organismal Movement Research." *Proceedings of the National Academy of Sciences* 105 (49): 19052–59. https://doi.org/10.1073/pnas.0800375105.

Nathan, Ran, Christopher T. Monk, Robert Arlinghaus, Timo Adam, Josep Alós, Michael Assaf, Henrik Baktoft, et al. 2022. "Big-Data Approaches Lead to an Increased Understanding of the Ecology of Animal Movement." *Science* 375 (6582): eabg1780. https://doi.org/10.1126/science.abg1780.

O'Dea, Rose E., Timothy H. Parker, Yung En Chee, Antica Culina, Szymon M. Drobniak, David H. Duncan, Fiona Fidler, et al. 2021. "Towards Open, Reliable, and Transparent Ecology and Evolutionary Biology." *BMC Biology* 19 (1): 68. https://doi.org/10.1186/s12915-021-01006-3.

Odenbaugh, Jay. 2003. "Complex Systems, Trade-Offs, and Theoretical Population Biology: Richard Levin's 'Strategy of Model Building in Population Biology' Revisited." *Philosophy of Science* 70 (5): 1496–1507. https://doi.org/10.1086/377425.

———. 2006. "The Strategy of 'The Strategy of Model Building in Population Biology.'" *Biology & Philosophy* 21 (5): 607–21. https://doi.org/10.1007/s10539-006-9049-3.

O'Malley, Maureen A. 2009. "Making Knowledge in Synthetic Biology: Design Meets Kludge." *Biological Theory* 4 (4): 378–89. https://doi.org/10.1162/BIOT_a_00006.

———. 2011. "Exploration, Iterativity and Kludging in Synthetic Biology." *Comptes Rendus Chimie* 14 (4): 406–12. https://doi.org/10.1016/j.crci.2010.06.021.

O'Malley, Maureen A., and Orkun S. Soyer. 2012. "The Roles of Integration in Molecular Systems Biology." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1): 58–68. https://doi.org/10.1016/j.shpsc.2011.10.006.

Roche, Dominique G., Loeske E. B. Kruuk, Robert Lanfear, and Sandra A. Binning. 2015. "Public Data Archiving in Ecology and Evolution: How Well Are We Doing?" *PLOS Biology* 13 (11): e1002295. https://doi.org/10.1371/journal.pbio.1002295.

Roche, Dominique G., Robert Lanfear, Sandra A. Binning, Tonya M. Haff, Lisa E. Schwanz, Kristal E. Cain, Hanna Kokko, Michael D. Jennions, and Loeske E. B. Kruuk. 2014. "Troubleshooting Public Data Archiving: Suggestions to Increase Participation." Edited by Jonathan A. Eisen. *PLoS Biology* 12 (1): e1001779. https://doi.org/10.1371/journal.pbio.1001779.

Roche, Dominique G., Rose E. O'Dea, Kecia A. Kerr, Trina Rytwinski, Richard Schuster, Vivian M. Nguyen, Nathan Young, Joseph R. Bennett, and Steven J. Cooke. 2021. "Closing the Knowledge-action Gap in Conservation with Open Science." *Conservation Biology*, November, cobi.13835. https://doi.org/10.1111/cobi.13835.

Ropert-Coudert, Yan, Anton P. Van de Putte, Ryan R. Reisinger, Horst Bornemann, Jean-Benoît Charrassin, Daniel P. Costa, Bruno Danis, et al. 2020. "The Retrospective Analysis of Antarctic Tracking Data Project." *Scientific Data* 7 (1): 94. https://doi.org/10.1038/s41597-020-0406-x.

Rutz, Christian. 2022. "Register Animal-Tracking Tags to Boost Conservation." *Nature* 609 (7926): 221–221. https://doi.org/10.1038/d41586-022-02821-6.

Rutz, Christian, and Graeme C. Hays. 2009. "New Frontiers in Biologging Science." *Biology Letters* 5 (3): 289–92. https://doi.org/10.1098/rsbl.2009.0089.

SC-CAMLR. 2020. "Report of the Thirty-Ninth Meeting of the Scientific Committee." Hobart: Scientific Committee for the Conservation of Antarctic Marine Living Resources. https://meetings.ccamlr.org/system/files/e-sc-39-rep.pdf.

Schimel, David, and Michael Keller. 2015. "Big Questions, Big Science: Meeting the Challenges of Global Ecology." *Oecologia* 177 (4): 925–34. https://doi.org/10.1007/s00442-015-3236-3.

Sequeira, Ana M. M., Malcolm O'Toole, Theresa R. Keates, Laura H. McDonnell, Camrin D. Braun, Xavier Hoenner, Fabrice R. A. Jaine, et al. 2021. "A Standardisation Framework for Bio-logging Data to Advance Ecological Research and Conservation." *Methods in Ecology and Evolution* 12 (6): 996–1007. https://doi.org/10.1111/2041-210X.13593.

Sequeira, Ana M. M., J. P. Rodríguez, V. M. Eguíluz, R. Harcourt, M. Hindell, D. W. Sims, C. M. Duarte, et al. 2018. "Convergence of Marine Megafauna Movement Patterns in Coastal and Open Oceans." *Proceedings of the National Academy of Sciences* 115 (12): 3072–77. https://doi.org/10.1073/pnas.1716137115.

Shavit, Ayelet, and James Griesemer. 2009. "There and Back Again, or the Problem of Locality in Biodiversity Surveys." *Philosophy of Science* 76 (3): 273–94. https://doi.org/10.1086/649805.

———. 2011. "Transforming Objects into Data: How Minute Technicalities of Recording 'Species Location' Entrench a Basic Challenge for Biodiversity." In *Science in the Context of Application*, edited by Martin Carrier

and Alfred Nordmann, 274:169–93. Boston Studies in the Philosophy of Science. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-90-481-9051-5_12.

Stegenga, Jacob. 2011. "Is Meta-Analysis the Platinum Standard of Evidence?" *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 42 (4): 497–507. https://doi.org/10.1016/j.shpsc.2011.07.003.

Sterner, Beckett W., Joeri Witteveen, and Nico Franz. 2020. "Coordinating Dissent as an Alternative to Consensus Classification: Insights from Systematics for Bio-Ontologies." *History and Philosophy of the Life Sciences* 42 (1): 8. https://doi.org/10.1007/s40656-020-0300-z.

Sullivan, Brian L., Jocelyn L. Aycrigg, Jessie H. Barry, Rick E. Bonney, Nicholas Bruns, Caren B. Cooper, Theo Damoulas, et al. 2014. "The EBird Enterprise: An Integrated Approach to Development and Application of Citizen Science." *Biological Conservation* 169 (January): 31–40. https://doi.org/10.1016/j.biocon.2013.11.003.

Sullivan, Brian L., Christopher L. Wood, Marshall J. Iliff, Rick E. Bonney, Daniel Fink, and Steve Kelling. 2009. "EBird: A Citizen-Based Bird Observation Network in the Biological Sciences." *Biological Conservation* 142 (10): 2282–92. https://doi.org/10.1016/j.biocon.2009.05.006.

Tempini, Niccolò. 2020. "The Reuse of Digital Computer Data: Transformation, Recombination and Generation of Data Mixes in Big Data Science." In *Data Journeys in the Sciences*, edited by Sabina Leonelli and Niccolò Tempini, 239–63. Cham: Springer.

Theobald, E.J., A.K. Ettinger, H.K. Burgess, L.B. DeBey, N.R. Schmidt, H.E. Froehlich, C. Wagner, et al. 2015. "Global Change and Local Solutions: Tapping the Unrealized Potential of Citizen Science for Biodiversity Research." *Biological Conservation* 181 (January): 236–44. https://doi.org/10.1016/j.biocon.2014.10.021.

Trappes, Rose. 2023. "How Tracking Technology Is Transforming Animal Ecology: Epistemic Values, Interdisciplinarity, and Technology-Driven Scientific Change." *Synthese* 201: 128. https://doi.org/10.1007/s11229-023-04122-5.

Tucker, Marlee A., Katrin Böhning-Gaese, William F. Fagan, John M. Fryxell, Bram Van Moorter, Susan C. Alberts, Abdullahi H. Ali, et al. 2018. "Moving in the Anthropocene: Global Reductions in Terrestrial Mammalian Movements." *Science* 359 (6374): 466–69. https://doi.org/10.1126/science.aam9712.

Waide, Robert B., and Sharon E. Kingsland, eds. 2021. *The Challenges of Long Term Ecological Research: A Historical Analysis*. Vol. 59. Archimedes. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-66933-1.

Weisberg, Michael. 2006. "Forty Years of 'The Strategy': Levins on Model Building and Idealization." *Biology & Philosophy* 21 (5): 623–45. https://doi.org/10.1007/s10539-006-9051-9.

Williams, Hannah J., Lucy A. Taylor, Simon Benhamou, Allert I. Bijleveld, Thomas A. Clay, Sophie Grissac, Urška Demšar, et al. 2020. "Optimizing the Use of Biologgers for Movement Ecology Research." Edited by Jean-Michel Gaillard. *Journal of Animal Ecology* 89 (1): 186–206. https://doi.org/10.1111/1365-2656.13094.