Species Conservation in the Era of Digital Big Data

Thesis submitted for the degree of Doctor of Philosophy

University of Oxford

John C. Mittermeier

School of Geography and the Environment

University of Oxford

Oxford, UK

Trinity 2019

Abstract

The Earth is in the midst of a global biodiversity crisis. Populations of plants and animals are declining dramatically, and thousands of species are predicted to go extinct in the coming century. This thesis explores how a new technological resource, digital 'big data,' can help conservationists combat this challenge. Big data has led to revolutionary advances in multiple fields but until recently has received limited attention in conservation. I focus on one aspect of big data, the data generated by people's activity online, and on one type of conservation, the conservation of species. I develop methods to access and interpret data related to people's interest in species using a prominent online platform, Wikipedia, and identify methodological challenges associated with this. I highlight the ability of Wikipedia pageviews to address questions of public interest in species across broad social and geographic scales and over vast numbers of online interactions. After developing methods to access and interpret relevant Wikipedia data, I explore patterns that these data reveal. I identify the biological traits of species that influence public attention and highlight the importance of seasonal and geographic patterns in determining public interest in species. I find that the presence and abundance of a species in a region is a significant predictor of its public interest across temporal and geographic scales. I conclude by identifying ways in which the insights revealed through these methods can be applied to conservation. Specifically, I argue that online data can provide awareness as to why people prefer, and thus are more inclined to conserve, some species rather than others; be used to systematically identify species of high interest at global and regional scales; monitor temporal and spatial variations in public attention; and, in some cases, track the distribution and abundance of species. I conclude that online big data are a valuable compliment to existing conservation methods. Knowing how to access and interpret these data should be part of every conservationist's tool kit in the twenty-first century.

Acknowledgments

Many people helped me over the course of my time at Oxford and this DPhil was only possible as a result of their assistance. I would like to thank first and foremost my supervisor Rich Grenyer for taking me on as his student and for providing me with feedback and guidance throughout my time at Oxford. The Wilfrid Knapp Scholarship from St Catherine's College helped to support my studies at Oxford as did research and travel grants from Oriel College and the School of Geography and the Environment. Kate Jones, Dave Redding and Robin Freeman helped to spark my initial interest in this idea and encouraged me to pursue it. My fellow lab members and co-authors provided me with feedback on manuscript drafts and discussed ideas with me at various points along the way, specifically Ricardo Correia and Emma McIntosh. Ruth Saxton patiently answered numerous questions about visas, academic requirements and deadlines. Gonzalo Diaz was a vital part of making the first published paper in this thesis possible and taught me the basics of coding by allowing me to look over his shoulder while we worked together. Tom Matthews answered numerous questions on modelling and statistics and gave me encouragement and feedback at vital points throughout the thesis. Most importantly on the academic front, however, I would like to thank Uri Roll. Uri was a constant source of discussion, ideas, and feedback throughout my DPhil. He helped to provide funding for the publication of papers, helped to outline chapters and was there to offer encouragement when I needed it most. Much of my academic growth and learning happened thanks to his detailed comments on numerous paper drafts. Thank you, Uri. It is no understatement to say that none of this would have been possible without you.

In addition to the academic side of my time at Oxford, I need to acknowledge the many people who helped to keep me sane, and often even happy, during what I frequently found to be a challenging experience. A few of these people need to be mentioned specifically: Edouard

Gottlieb, Dan Antoun, Jan Stockmann, Will Tant, Martin Lesourd, Barclay Bram Shoemaker, Bryn Elesedy, Jaeyoung Lee, Filipa Soares, Erik Sandvig, and Sylvia Alvares-Correa, thank you all. Most of all, I would like to thank Sarah Waltcher who among much else never failed to brighten my day as I was writing and editing. Athletic activities of various kinds kept me grounded while spending much of my time in front of a computer and I would never have reached the end without the OUTFC, the Oriel College Boat Club, Henrik Hannemann, Jinwoo Leem and the EGBC, the Hard-os, and the Oxford University Athletics Club Short Sprints Team. Finally, my grandparents John and Sylvia Constable inspired me and helped to support me in my decision to pursue a DPhil and I would not have been able to do this without them. My grandmother in particular allowed me to hole up in a corner of her house for months while I did the final writing. Most significantly of all however, I want to thank my mother. Mom, no words can express how important you have been throughout every part of this. This thesis is dedicated to you.

Table of Contents

Introduction

The conservation of biological diversity has been recognized as a global priority for over thirty-five years (Ehrlich & Ehrlich 1981; Soule 1985). In 1984, the Pulitzer-prize winning biologist E.O. Wilson described the loss of biodiversity as "the folly our descendants are the least likely to forgive us," (Wilson 1984) and by 1992 recognition of the 'biodiversity crisis' was sufficient for representatives from 178 countries to meet at the Rio de Janeiro "Earth Summit" and enact the United Nations Convention on Biological Diversity (CBD). Ten years later, in 2002, the now 190 countries signatory to the CBD committed to achieving "a significant reduction of the current rate of biodiversity loss at the global, regional and national level" by 2010 (CBD 2002). Despite this international collaboration and these bold targets, however, biodiversity policy has largely failed to achieve its goals (Gilbert 2009; Walpole et al. 2009; Hoffmann et al. 2010; Rands et al. 2010). Extinction rates are projected to climb to 10,000 times the normal rate (De Vos et al. 2014), populations of once common and widespread species are declining precipitously (Dirzo et al. 2014; Rosenberg et al. 2019), and the loss of Earth's biodiversity is increasing and intensifying (Butchart et al. 2010).

Independent of conservation, many aspects of our daily lives together with an increasing number of research fields are undergoing a 'big data revolution' driven by exponential increases in the volume, velocity, and variety of data (Mayer-Schönberger & Cukier 2013; Kitchin 2014a). Big data analytics present us with customized shopping recommendations, suggest the music we listen to and the television shows we watch, and determine the stories that appear in our newsfeeds. When we contact customer service, we chat with bots that employ data analytics to process and reply to our complaints, and when we type in credit card details for a purchase, data analytics

assess whether or not to flag the activity as fraudulent. In research, the ability to aggregate and interrogate previously inconceivable volumes of digital information has provided insights into questions in genetics (Howe & Yon 2008), healthcare (Kayyali et al. 2013), economics (Einav & Levin 2013), and business management (McAfee & Brynjolfsson 2012), among others, and is progressively being used in qualitative fields such as history (Graham et al. 2016) and social science (Lazer et al. 2009). In the words of Lazer et al., the vast quantities of digital interactions in our lives "offer increasingly comprehensive pictures of both individuals and groups with the potential of transforming our understanding of our lives, organizations, and societies in a fashion that was barely conceivable just a few years ago" (Lazer et al. 2009, p. 722).

The aim of this thesis is to consider how digital big data can be applied to the challenge of conserving earth's biological diversity. There are many ways in which large quantities of digital data can, and in some case already are, being applied to biodiversity conservation (Hampton et al. 2013; Kelling et al. 2015; Ladle et al. 2016). Here I focus on a single type of big data, online text and internet pageviews. Pageviews are the more important of these two (more on this in Methods). They are the digital footprints that accrue as we create and search for information on the internet. There are a number of sources that can be used to extract online text and pageview data, Google Trends and Twitter are conspicuous examples, but for reasons that I explain later, I focus primarily on Wikipedia, the large, open-access encyclopaedia.

Just as there are many types of digital big data with relevance to conservation, there are also a wide array of questions and activities within conservation to which these data can be applied. Once again, I have narrowed things down. I focus my attention on species conservation. Species are the most frequently-used unit when assessing the state of biodiversity (Rodrigues et al. 2006; Hoffmann et al. 2010; Ladle & Jepson 2010), they are often employed as flagships to motivate

support for broader conservation policies (Caro 2010), and they are one of the principal ways in which people engage with and understand biodiversity. In the context of this thesis, I use the term "species conservation" in a broad sense to encompass the range of conservation activities and policies for which species are the units of measurement or engagement.[1]

Within the intersection between species conservation and online big data, my goals are threefold. First, I develop methods to access and interpret relevant online data. Second, I investigate some of the patterns that these methods reveal. Third, I propose how these methods and patterns can be applied to conservation practice and policy. The structure of thesis is arranged around these three goals with each chapter focussing on one of them (Methods, Pattern, Applications).

The thesis follows a "paper route" and thus consists of a compilation of individually submitted publications. For each chapter, I provide a brief introduction to the chapter as a whole as well as introductions to each of the papers included in that chapter. The paper introductions contextualize how each paper fits into the broader themes of the thesis and, where necessary, provide some additional background information. I comment on the paper's submission status and its progress towards publication. All of these papers are multi-author collaborations. For purposes of the thesis, therefore, I also provide a summary of my personal contribution to the research. To help clarify how each of the published papers fit within this broader thesis structure, I list each paper, together with its relevant datasets and key conclusions in Table 1.

Chapter 1 is Methods. Here I explore specific methods for how online big data can be used to quantify interest in species. A vital component of these analyses is selecting the appropriate dataset, and I provide background on why I have chosen to use Wikipedia as the primary resource

[1] I make this distinction to avoid confusion with the narrower definition of "species conservation" as efforts solely focused on protecting the populations of individual plants and animals.

for quantifying online interest in species throughout the thesis. I also discuss some of the limitations and challenges associated with using online data in the context of conservation. This chapter contains one primary article (Methods 1) and an additional supplementary article (Methods 2). Methods 1 is the principal work of the chapter and covers Wikipedia as a dataset as well as important considerations for using Wikipedia data to quantify interest in species. Methods 2 is a brief contribution; I include it because it emphasizes an important point with relevance to the methods and to justifying my use of Wikipedia.

Chapter 2 focuses on Patterns. In this chapter, I align metrics of online interest with biodiversity datasets that quantify the physical traits, distribution, abundance, and extinction risk of species. I explore the relationships between online interest and these biological and ecology traits and discuss how the patterns revealed through these relationships are relevant to conservation. The chapter contains three papers, each varying in the species covered and the biodiversity datasets used. Patterns 1 explores how the physical traits of organisms correlate with online interest and compares Wikipedia pageviews for reptile species with data from the Global Assessment of Reptile Distributions. Patterns 2 looks at the role of timing, specifically seasonality, in determining online interest. It evaluates seasonal variations in pageviews for all of the nearly 32,000 species listed by the International Union for Conservation of Nature (IUCN) that have pages in Wikipedia. Finally, Patterns 3 explores the relationship between the regional abundance of a species and its online interest. It uses avian distributional and abundance from eBird.org for bird species in 25 different regions. Together these three papers comprise the primary analytical work of the thesis.

Chapter 3 deals with Applications. Here I explore ways in which these metrics of online interest and the patterns they reveal can be applied to biodiversity conservation. By virtue of the

format of the thesis, this topic of applications has already been discussed to varying extents in the previous chapters. The purpose of this chapter, together with the conclusion, is to summarise and synthesize some of these findings. Additionally, this chapter includes two papers that demonstrate specific examples of two of the applications discussed. The first paper (Applications 1) explores the potential to use online data to track the real-world movements of species. The second paper (Applications 2) provides a case study of using online data to systematically identify flagship species at a global scale.

The three primary chapters are book-ended by a section on Context and a Conclusion. In the Context section, I review background literature and offer some perspective and justification for the research as a whole. Since each of the submitted papers contains its own literature review, there is some unavoidable repetition between the Context and the papers. In the Conclusion, I summarise the main results of the thesis and reflect on some of the insights I have gained through the research process. Unlike the papers in the following three chapters, which have either been published or are submitted to be published, the Context and Conclusion sections exist solely for the purpose of this thesis. If you make the effort to read them, congratulations, you are probably one of the only people who will ever do so.

Placement in the thesis	Paper title	Taxonomic source	Data source	Key conclusions
Methods				
1. Methods 1	*What are the most popular birds in Wikipedia? Methods and considerations for measuring public interest in biodiversity using online data*	eBird/Clements List of World Birds	Wikipedia metadata	• Wikipedia is an excellent resource for analysing public interest in biodiversity • Specific biases need to be considered when using Wikipedia data • Of the many Wikipedia metadata, pageviews are the most useful for assessing public interest • Wikipedia pageviews can be used to identify global high interest species
2. Methods 2	*Inferring public interest from search engine data requires caution*	n/a	Google Trends	• Uncritical use of online big data can lead to misleading results • Sources that do not allow access to their raw data present challenges for research
Patterns				
3. Patterns 1	*Using Wikipedia page views to explore the cultural importance of global reptiles*	Global Assessment of Reptile Distributions	Wikipedia pageviews	• Reptile species that are large, venomous and endangered tend to receive more interest online
4. Patterns 2	*A season for all things: Phenological imprints in Wikipedia usage and their relevance to conservation*	IUCN listed species	Wikipedia pageviews	• Online interest in biodiversity follows seasonal patterns • The prevalence of seasonality varies according to languages and taxonomic clades

5. Patterns 3	*It is good to be common: birds that are more frequently encountered in the wild generate higher interest online*	eBird occurrence data for birds in 25 countries	Wikipedia pageviews	• Birds that are more frequently observed in the wild tend to attract greater public interest online • This pattern is prevalent across multiple linguistic and geographic contexts
Applications				
6. Applications 1	*Correspondence: The digital imprints of bird migration*	eBird occurrence data for migratory birds in 10 countries Bird ringing data in Japan and the USA	Wikipedia pageviews Google Trends	• For some species, temporal patterns in Wikipedia pageviews closely track their migratory movements
7. Applications 2	*Flagship species can deliver efficient conservation*	IOC World Bird List	Wikipedia pageviews	• Wikipedia pageviews can be used as a starting point to systematically identify global flagship species

TABLE 1. Overview of the seven papers included in this thesis. For each paper, I list the title at the time of its initial submission, the placement within the structure of the thesis, the biological and online big data resources that it uses, and give a brief overview of its key conclusions.

Context

As I mentioned in the Introduction, my thesis operates at the intersection of biodiversity conservation and the emergence of digital big data. The overlap between these two areas is broad and I investigate one component of it: how online big data can benefit species conservation. To set the stage for the thesis and justify my interest in this specific area, this section touches on three points. First, I review the importance of species conservation generally as well as the relevance of understanding human attitudes in the context of species conservation. Second, I provide some background on digital big data and big data research. And third, I discuss how big data methods are already starting to be applied to questions in conservation as part of the emerging research area of "conservation culturomics."

Species conservation

The conservation of individual species is one of the oldest and most-recognizable forms of conservation (Primack 2014). Efforts to manage other species have been part of human interactions with the natural world for millennia (Nash 1990), while concerns regarding the extinction of individual organisms date to at least the mid-1500s (Primack 2014). Species are at the core of some of the most recognizable and significant conservation legislation, campaigns, and organizations. While the term "biodiversity" may be overly abstract for many non-experts (Santana 2014), species form one of the key ways that people engage with and value biodiversity. They also play a fundamental role in how we define and quantify the scale of the biodiversity crisis. Claims that the 'sixth mass extinction' is underway rest on species data (Pimm et al. 1995; Barnosky et al. 2011), while the disappearance and decline of species forms one of the primary metrics for

assessing conservation progress or lack thereof (Rodrigues et al. 2006; Hoffmann et al. 2010; Ladle & Jepson 2010).

Given the history and significance of species conservation, it is worth reviewing what is known about it. I identify four broad conclusions. First, in some situations, conservation efforts can be very effective at saving species. For example, Butchart et al. estimate that 16 species of birds were saved from extinction in the decade between 1994-2004 as a direct consequence of conservation efforts (Butchart et al. 2006). In the US, meanwhile, implementation of the Endangered Species Act has played a prominent role in preventing the decline of species such as American Alligator (*Alligator mississippiensis*), Bald Eagle (*Haliaeetus leucocephalus*), and Whooping Crane (*Grus americana*) (Schwartz 2008; Primack 2014).

Second, while the above examples demonstrate that species conservation can be done successfully, it is infeasible at a scale necessary to 'save all species.' In Butchart *et al.*'s study, the 16 species that were saved represent only 1.3% of all those threatened, and in the same period ten times as many birds (164 species) moved closer to extinction (Butchart et al. 2006). While species conservation works, it is not, and perhaps cannot in its current form, be done at a scale sufficient to shift the overall downward trend in biodiversity (Hoffmann et al. 2010).

Third, the underlying assumption central to many early species conservation efforts that saving a species can be a one-off investment is often flawed (Scott et al. 2010); instead many species remain "conservation-reliant" and require continued investment in order to survive (Scott et al. 2005; Goble et al. 2012).

Fourth, in many situations the challenges facing species conservation projects are primarily social rather than biological (Kellert 1982; Tisdell 2014), and as such evaluations of species decline and extinction need to consider social and cultural variables in addition to biological ones

(Ladle & Jepson 2008). In the United States, for example, both the reduction in range of the Gray Wolf (*Canis lupus*) and challenges surrounding its reintroduction are driven primarily by public attitudes towards the species (Bright & Manfredo 1996; Williams et al. 2002). As Stephen Kellert points out: "most species are endangered not because of their biological inadequacies, but because of a variety of human, social, psychological, and cultural factors" (Kellert 1982).

This final point is worth considering in more detail. Seen broadly, Kellert's "human, social, psychological and cultural factors" influence how people interact with, prioritize, and manage particular species. As a result, these factors impact the extinction risk of species (Sheil & Meijaard 2010), effect the conservation investment a species receives (Walpole & Leader-Williams 2002; Martín-Forés et al. 2013), and often determine the long-term success or failure of conservation initiatives (Kellert 1982; Stokes 2007).

The importance of social variables is well recognized in the conservation literature. Myers, for example, notes 'cultural and aesthetic valuation' as one of six classes of attributes to consider in the setting of conservation priorities (Myers 1983); the concept of ecosystem services includes 'cultural services' consisting of components such as 'aesthetic information' and 'cultural and artistic inspiration' (Costanza et al. 1997; de Groot et al. 2002); Wilson addresses the contribution of other species to human culture and psychology through the concept of biophilia: 'the human tendency to relate and connect with the natural world' (Wilson 1984), and the Convention on Biological Diversity explicitly lists the "cultural, recreational and aesthetic values of biodiversity" as one of the core reasons for protecting biodiversity.

Despite this recognition, however, cultural factors often receive cursory attention in conservation and, in contrast to other biodiversity features, are not routinely incorporated in science-based conservation strategies (Kellert 1982; Gunnthorsdottir 2001a; Stokes 2007).

Quantifiable differences in the geographic distribution, population size, ecological role, and evolutionary distinctness of species underlie some of the most significant prioritization schemes in conservation (Vane-Wright et al. 1991; Faith 1992; Mills et al. 1993; IUCN 2014). Similar to these other variables, species are also unequal in their contributions to human culture. While a few authors have addressed this point (Cristancho & Vining 2004; Garibaldi & Turner 2004), the extent to which species vary in their cultural impact remains poorly studied and cultural variables are rarely addressed in prioritization schemes.

In addition to reviewing species conservation efforts specifically, it is important to contextualize how they fit into broader discussions in conservation. As mentioned above, the failure to achieve success in reaching targets has led to debates on the fundamental motivations of biodiversity conservation (Doak et al. 2013; Mace 2014; Marvier & Kareiva 2014; Tallis & Lubchenco 2014). Among other things, this has often led to an emphasis on more human-centric, instrumental approaches to conservation (Kareiva & Marvier 2012; Marvier & Kareiva 2014), a perspective reflected in the exponential growth of interest in the ecosystem services policy frame (Costanza et al. 1997; de Groot et al. 2002; Mace 2014). The ecosystem services concept advocates conserving biodiversity for the benefits it provides to humanity and, as a result, prioritizes conservation actions based on their potential to deliver these benefits (McCauley 2006; TEEB 2010).

The validity of instrumental as opposed to intrinsic justifications for conservation has been fiercely debated (McCauley 2006; Maguire & Justus 2008; Child 2009; Justus et al. 2009; Buscher et al. 2012). In the context of species conservation, both perspectives present challenges. Quantifying the economic benefits provided by individual species is difficult and many species have limited or no measurable economic benefit (McCauley 2006). As such species conservation

is often marginalized within the ecosystem services policy frame (Mace 2014). Conversely, intrinsic arguments which frequently centre on the idea that 'all species are equal' and that the goal of conservation should be to have "zero extinctions" suffer challenges. Species differ in their contributions to ecological systems and in the amount of evolutionary history that they represent, and species themselves are from a biological standpoint mutable, inconsistent entities (Vane-Wright et al. 1991; Hey 2001). When it comes to species conservation, instrumental approaches often fail to effectively capture the value of species while purely intrinsic ones frequently miss the realities of species inequalities.

Given what has been learned about species conservation to date, four priorities for future research are apparent: 1) *prioritization*, which species should be the focus of specific conservation projects? 2) *Engagement*, how can new stakeholders be incorporated in conservation activities? 3) *Social costs and benefits*, if species conservation is a long-term investment then who benefits from the conservation of a species and who should bear the cost of protecting it? 4) *New policy frames*, how do species conservation efforts fit into instrumental policy frames such as ecosystem services? Cultural and social variables are relevant to each of these challenges and as such devising methods to explicitly consider them could benefit modern species conservation efforts in multiple ways.

Big data

In order to fathom the relevance of digital big data, it is worth considering its *volume*, how much data is out there. The human brain can comprehend *megabytes* of data. The King James translation of the Bible, for example, contains 3,116,480 individual letters which in computing translates to just under three *megabytes* of data. Modern personal computers bring us into contact with *terabytes* of data, and these already exceed what we can manually process in terms of text.

Your one-terabyte hard-drive would hold just over 350,000 King James Bibles. At an average reading speed with no stops to eat, sleep, or visit the bathroom, these would take just over 2,102 years to read.[2] If someone had started non-stop reading a terabyte of Bibles at the time of Christ's death, they would still have nearly a century to go. To consider global scales of data, however, we need bigger units. One *zettabyte* is one trillion terabytes. The entire age of earth would be enough time to read 0.0002% of a zettabyte of text. Printed out and stacked one on top of another, a zettabyte of bibles would be nearly 2 light-years high, enough to cover the distance between the earth and sun more than 100,000 times or make it about halfway to our nearest star system.[3]

Those are the units, now here are the numbers: in 2013, a couple of years before I began my thesis, the total digital data on earth was around 4.4 zettabytes (EMC Digital Universe 2014). By 2020, the year after I finish, there will be around 50 zettabytes (Marr 2017). Fifty zettabytes is twenty-five stacks of bibles each high enough to reach the Alpha Centauri star system. While I have been writing, human society has every two days generated the same amount of data as it did over the entirety of human history up until 2003 (Marr 2017), and by the time I finish we will be generating an estimated twelve and half thousand terabytes of new digital data—a stack of bibles more than half-way to the moon—every second (Domo Inc 2018).[4] Whether we like it or not, big data together with the opportunities and challenges it offers, is here to stay.

---

[2]There are many examples of the relative size of digital data online, but the underlying numbers used in these examples are often unclear. Here is what I have used for this particular example: King James Bible = 783,137 words and 3,116,480 characters; one character = 1 byte; 3,116,480 bytes = 2.97 megabytes; 1 terabyte = 1,048,576 megabytes / 2.97 megabytes per Bible = 352,805.24 Bibles. Average reading speed = 250 words per minute; 783,137 words / 250 words per minute / 60 minutes per hour = 52.21 hours of reading per Bible; 52.21 hours x 352,805.24 Bibles / 24 hours per day / 365 days per year = 2,102.7 years to read a terabyte.

[3] More numbers: age of the earth = 4.543 billion years / 2,102.7 years to read one terabyte = 2.16 million terabytes read since the earth appeared; 1 trillion terabytes per zettabyte / 2.16 million terabytes = 0.000002 zettabytes x 100 = 0.0002%. Height dimensions: size of the "Holy Bible: King James Version Popular gift & award black leatherette edition" = 11.9 x 4.6 x 19 cm; distance between the earth and sun = 149.6 million km; one light year = 9.46 trillion km; distance to Alpha Centauri = 4.367 light years.

[4] And a final bit of numeric clarification: Data Never Sleeps 6.0 (Domo Inc 2018) estimates 1.7 megabytes of data created per human on earth per second; human population in 2020 = 7.758 billion x 1.7 = 13.189 billion megabytes

The term 'big data' originated in the early 1990s, but only in the last ten years has it become a popular buzzword (Diebold 2012; Kitchin 2014a). While big data broadly refers to the exponential increase in digital information, it does not have an explicit quantitative definition (De Goes 2013; Kitchin 2014a). There is no number of bits above which data become "big." Instead, big data is generally defined by the "3Vs" of volume, velocity, and variety. Volume refers to the massive quantity of data; velocity to the speed at which data are created and processed; and variety to the growth of different types of data (Zikopoulos et al. 2012; Kitchin 2014a). Veracity, implying that big data is more credible and reliable than traditional small data sets is sometimes also includes as a fourth "V" in definitions (Raghupathi & Raghupathi 2014). In addition to the Vs, Kitchin identifies four other "key characteristics" of big data (Kitchin 2014a). *Exhaustive scope*; rather than dealing with a subset of available information big data takes everything into account. *Fine scale resolution*; big data is extremely detailed in its coverage. *Relational nature*; big data uses fields that allow conjoining multiple data sets. And, *flexibility and scalability*; big data is built to grow and can easily add new fields and new data.

While the precise definition of big data may be debated, its impact is undeniable. The influence of big data is so significant that it has been described as making the current scientific method "obsolete" (Anderson 2008) and ushering in a new 'fourth paradigm' of scientific investigation (Hey et al. 2009). While traditional scientific methods require developing a hypothesis, testing it with available data, and then using those results to extrapolate beyond the test, big data make it possible to incorporate all or nearly all the available information. In doing so, it reduces the need to make hypotheses about causation (Prensky 2009). As Wired magazine's Chris Anderson argues, in the context of big data, "correlation is enough" (Anderson 2008). In

= 12,580 terabytes = 4.437 billion King James Bibles = 20.412 billion cm at 4.6cm per Bible = ca. 240,000km. Distance between the earth and moon = 384,000 km.

other words, when the scale of the data is sufficiently large and inclusive, patterns in the data can be enough to guide decision-making and make informed conclusions, even without understanding their underlying mechanisms. This point is important, particularly since identifying underlying mechanisms can be often challenging or even impossible at the scale of many big data analyses.

In addition to traditionally quantitative fields such as genetics and computing, big data has influenced research in the social sciences. The potential applications of big data in social science studies are vast; computational approaches have been used to forecast changes in the stock market based on people's moods (Bollen et al. 2011), track daily fluctuations in happiness (Dodds et al. 2011), and to look at changes in word usage over time (Michel et al. 2011; Aiden & Michel 2013), among others. As digital text archives continue to expand, the opportunities to apply big data analyses to questions in the social sciences will only increase.

In order to describe the impact of big data on the social sciences, Cioffi-Revilla *et al*. equate this technological jump to the invention of the telescope: "just like Galileo exploited the telescope as the key instrument for observing and gaining a deeper and empirically truthful understanding of the physical universe, computational social scientists…exploit the advanced and increasingly powerful instruments of computation to see beyond the visible spectrum of more traditional disciplinary analyses…it is the instrument of investigation that drives the development of theory and understanding" (Cioffi-Revilla 2010 p. 260). Examples of research "beyond the visible spectrum" that computational methods enable are increasingly prevalent (Lazer et al. 2009; Aiden & Michel 2013; Schich et al. 2014). First defined by Michel *et al.* (Michel et al. 2011), culturomics refers specifically to the use of computational methods to quantitatively study cultural trends using word frequencies in digitized text archives (Bohannon 2011; Michel et al. 2011). As an example of how these methods work, Aiden and Michel track how attitudes shifted from seeing the United

States as a coalition of states to a single nation based on the relative frequencies of the word sequences "United States is" as opposed to "United States are" in the Google Books archive (Aiden & Michel 2013).

Applying the quantitative approaches of big data analytics to social sciences presents several challenges. Many of these are technical; most social scientists are not trained in the skills required to manage large and dynamic datasets (Kitchin 2014b). Less obvious, but equally important, however, are the epistemological challenges presented by big data (Kitchin 2014b, 2014a). Three are particularly relevant to my research and emerge repeatedly over the course of the papers. First is the necessity to be cognizant of the limitations of big data methods (Kitchin 2014b, 2014a). Despite claims that big data represents "the end of theory" (Anderson 2008), Kitchin concludes that in reality these new methods are "unlikely to lead to the establishment of new disciplinary paradigms;" instead big data are better viewed as a set of tools that compliment rather than replace traditional, smaller scale studies (Kitchin 2014b). While big data approaches can detect trends and patterns "invisible" to other scales of analysis, they are poorly suited to detecting the ambiguities and intricacies that are often present in such studies.

Second, there is temptation amongst proponents of big data in quantitative fields to suggest that the results obtained are "objective" and, as a corollary to this, that more data leads to more objectivity; in reality, data interpretation and selection is never neutral irrespective of the scale at which it occurs (boyd & Crawford 2012; Rieder & Rohle 2017). Third, it is important to be cognizant of the context of data and types of knowledge it reflects (boyd & Crawford 2012; van Es et al. 2017). While context is important to any research, the fact that big data operates at scales that are human unreadable often makes it easy to lose track of context. Furthermore, since big data approaches rely on detecting correlations rather than hypothesis testing, context becomes even

more important when attempting to draw conclusions. The perils of this are neatly demonstrated in Leinweber's finding that fluctuations in the S&P 500 correlate with butter production in Bangladesh (Leinweber 2007). Despite claims to the contrary, simply looking at correlations without investigating the underlying mechanisms can lead to erroneous conclusions.

All three of these points emerge repeatedly in various forms in the papers included in this thesis. I will revisit them again at various stages, but as an overarching theme is it important to be cognizant of the limitations of this work. The approaches I use do not negate nor supersede the benefits of more traditional studies, but instead offer new perspectives that complement existing methods.

Conservation Culturomics

Concurrent with my thesis, the potential of digital big data to address questions in conservation has begun to gain more attention, and new concepts such as "digital conservation" (van der Wal & Arts 2015) and, in particular, "conservation culturomics" (Ladle et al. 2016) have been defined to describe research in this area. This growing recognition of the value of digital big data in conservation was highlighted in 2018 by "conservation culturomics" being identified as one of the top 15 emerging issues in global conservation (Sutherland et al. 2018). The timing of my thesis has allowed me to participate in this broader movement (see, for example, Methods 2), and these opportunities for collaboration have been one of the most rewarding aspects of my research.

One way to categorise recent culturomic studies in conservation is according to the digital data sources that used. Google Trends is one of the most popular and has been used to monitor public perceptions of conservation (Soriano-Redondo et al. 2017), assess the overlap between

where species occur and the amount of attention they receive online (Correia et al. 2016), and look at public awareness of broad conservation-related concepts (Proulx et al. 2013; Burivalova et al. 2018). While Google Trends data offer appealing opportunities for analysing questions of public interest and awareness, they also present challenges (Correia et al. 2019). This will emerge as an important theme in Methods 2. Another resource in culturomic studies is social media data (Di Minin et al. 2015; Toivonen et al. 2019). Among the social media platforms, Instagram and Flickr have been used to evaluate people's interests and activities in protected areas (Wood et al. 2013; Hausmann et al. 2018) as well as what motivates people to visit those areas (Hausmann et al. 2017). Meanwhile, Twitter data have been used to assess people's engagement with species (Roberge 2014), and monitor public awareness of threatened species (Kidd et al. 2018).

In addition to these dataset-specific studies, several review papers outlining the potential of online data to conservation and offering guidance for future research have been published during the course of my thesis research (Arts et al. 2015; Di Minin et al. 2015; Ladle et al. 2016; Sutherland et al. 2018; Toivonen et al. 2019). These have highlighted some of the possible applications of conservation culturomics, such as identifying constituencies that support conservation actions (Ladle et al. 2016) and measuring public interest in biodiversity over time (Sutherland et al. 2018). They have also identified priorities for furthering the integration of culturomics into conservation. Of these, the need to critically develop and refine methods to access and interpret culturomic data has emerged as the principal need in order to advance culturomic research in conservation in the immediate future (Toivonen et al. 2019; Correia et al. 2020).

My thesis provides two unique contributions to the emerging research area of conservation culturomics. First, it helps to address the methodological shortfall in the field by developing specific methods to access and interpret data from Wikipedia, a large and promising resource for

conservation culturomic analyses. Two of the papers in the thesis, Patterns 1 and Patterns 2, are the first studies to use Wikipedia data in conservation culturomics. In Methods 1, I provide background on Wikipedia, justify my decision to use it as the primary source of online big data in the thesis, and reference prior research that uses Wikipedia and investigates aspects of its use and creation. Second, my research compares interest in species across very large scales both in terms of numbers of users and numbers of species. Other culturomics studies have often focussed on concepts or place-based experiences (e.g. Nghiem et al. 2016; Hausmann et al. 2018) while those that have worked with species have done so with more limited geographic and taxonomic scopes; bird species in Brazil (Correia et al. 2016) or butterflies in the United Kingdom (Zmihorski et al. 2013), for example. As I described in the Methods chapter, Wikipedia is uniquely well-suited to investigating patterns at these broad scales; indeed, I would argue that it is the best available resource anywhere for addressing these types of questions. Thus, the questions I explore in the thesis and the methods I develop in it, are closely intertwined.

Chapter 1: Methods

There is a vast and growing quantity of information in the digital universe. Just because this information exists, however, does not mean it can be easily accessed. More importantly, even when it is accessible, interpreting information at the scale of big data can be extremely complex (boyd & Crawford 2012; Lazer et al. 2014; Sivarajah et al. 2017). In the context of this thesis, it is important to mention three methodological questions that need to be considered when working with big data: data access, data type, and data interpretation.

Access is related to what data are available to a researcher. For example, historical newspapers are interesting for answering a variety of conservation questions; however, many of these are either not digitized or, if digitized, they are not fully indexed and searchable. The data are there, but currently inaccessible. Another important component of access has to do with proprietary ownership. Facebook, Google, and Instagram area all widely used social media platforms that could be valuable resources for investigating questions related to human interest in particular species. As private companies, however, they limit the data that is publicly available. As of 2019, for example, Facebook allows access to raw data, but only a limited selection of it, whereas Google provides a greater range of data, but only makes data available in a pre-analyzed format (for more on this see Methods 2). Twitter allows open access to tweets but requires that users create their own dataset by collecting tweets over time. In this case, simply having the computing power and time to collect, download and maintain vast quantities of data can be a significant barrier to access.

Data type concerns the attributes of the data that get measured and considered. To use Twitter as an example, what components of the downloaded data would be most interesting to

address for conservationists: the content of a tweet or the number of "re-tweets"? Which is more important, how often something is "re-tweeted" or how often it is "liked"? Within a single dataset, these different data types can represent distinct users and forms of interactions. Thus, depending on the data source there will be a range of dataset-specific considerations to make regarding the most appropriate data type for the questions being asked.

Finally, data interpretation has to do with associating real-world sentiments and behaviours with the selected data. If someone "likes" a photograph of a baby tiger are they more prone to support tiger conservation or are they simply a fan of cute cat photos? At the large scales present in big data analyses, identifying the connection between action and causation can often be difficult and in some cases impossible. Understanding the links between online data and real-world actions and sentiments is one the principal challenges of conservation culturomics research and big data analytics in general.

Datasets frequently have specific requirements in terms of their data access, data types, and data interpretation. As a result, using a dataset for culturomic research requires domain-specific knowledge about that source. I use Wikipedia as my principal source for assessing online interest throughout this thesis. For the questions I am interested in, Wikipedia has unique advantages in terms of its popularity, coverage, structure, access, and the existing research that has been done with it. I describe each of these in Methods 1. In addition to concentrating on Wikipedia as a general dataset, I focus on Wikipedia pageviews as the primary datatype throughout the thesis. Pageviews reflect users' interest in expanding their knowledge about a particular subject, and increased knowledge has been shown to correspond with greater conservation ethic and interest (Lindemann-Matthies 2002, 2005). Again, this decision is justified in detail in Methods 1.

This chapter consists of two papers. Methods 1 is the more important of the two. It outlines an approach for using Wikipedia to systematically identify high interest species. In many cases, these high interest species could be good candidates to act as flagship species in conservation. More importantly in the broad context of the thesis, Methods 1 justifies the use of Wikipedia and specifically Wikipedia pageviews as a measure of online interest. Methods 2 is a collaborative paper on which I am one of several joint lead authors. Though it is only a brief communication, I have included it here because it demonstrates a specific point regarding the interpretation challenges associated with big data analyses in the conservation context.

Methods 1

This paper is the primary methodological contribution of the thesis. It justifies the use of Wikipedia, and specifically Wikipedia pageviews, as a measure of "online interest" and makes best practice recommendations regarding data access, the selection of data types and the interpretation of data. It also details methods for weighting Wikipedia pageviews so that they can be used to systematically identify high interest species both within and across languages. These methods are later demonstrated in Applications 2, which provides an example of how they can used in a conservation context.

Submission status: Submitted to PLoS Computational Biology 9 October 2019. Mittemeier, J.C., Roll, U., Correia, R., and R. Grenyer. *What are the most popular birds in Wikipedia? Methods and considerations for measuring public interest in biodiversity using online data.*

Personal contribution: Lead author. I devised the initial concept of the paper, downloaded and curated the data, conducted the formal analysis, developed and designed the methodology, created the visualizations, and wrote the initial draft of the text. Co-authors assisted with reviewing the methodology and editing drafts.

What are the most popular birds in Wikipedia? Methods and considerations for measuring public interest in biodiversity using online data

John C. Mittermeier[1], Uri Roll[2], Ricardo Correia[3,4], Rich Grenyer[1]

[1] School of Geography and the Environment, University of Oxford, South Parks Road, Oxford, OX1 3QY, UK

[2] Mitrani Department of Desert Ecology, The Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Midreshet Ben-Gurion 8499000, Israel

[3]DBIO & CESAM-Centre for Environmental and Marine Studies, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

[4] Institute of Biological and Health Sciences, Federal University of Alagoas, Maceió, Brazil

Abstract

Assessing public interest in biodiversity is important to conservation practice and policy development. However, at large scales measuring public interest is challenging and often prohibitively expensive using traditional methodologies such as interviews or survey questionnaires. Online data platforms have grown exponentially in recent years. These digital 'big data' offer the potential for conservationists to rapidly and inexpensively assess aspects of public interest across a broad cross-section of the internet-using public and large numbers of biodiversity features. Here we highlight the potential of online data for quantifying public interest in biodiversity by analyzing Wikipedia. Due to its large size, prevalent usage, underlying

data structure, and open-access nature, Wikipedia is uniquely well-suited to comparing interest over large sets of biodiversity features across broad linguistic and geographic scales. We identify and discuss six methodological considerations that should be addressed when comparing public interest in biodiversity using Wikipedia: data selection, temporality, taxonomy, language representation, biogeography, and the geography of Wikipedia. While we focus on Wikipedia specifically, the methodological considerations we identify are relevant to comparing public interest in biodiversity in other digital big data sources. As a case study, we analyze 757 million Wikipedia pageviews for 10,099 species of birds across 251 Wikipedia languages to identify globally popular bird species. These types of data-driven popularity indices could be used to help identify flagship species, pinpoint demographic groups likely to support a particular conservation initiative and investigate the underlying drivers of species charisma and popularity.

Impact statement

Online data platforms offer exciting opportunities to measure public interest in biodiversity, but it is important to develop effective methods for conducting these analyses in the context of conservation. We discuss methods for using Wikipedia to compare public interest in biodiversity and as a case to assess interest in global bird species across 251 languages.

Introduction

Effective conservation requires public support. Therefore, it is important to understand which aspects of biodiversity attract the most public interest. For example, knowing which species are popular is relevant to selecting appropriate flagship species (Caro, 2010), and

devising conservation marketing campaigns (Smith, Veríssimo, Isaac, & Jones, 2012). Knowing which species attract public interest can also aide in identifying species with a high cultural significance (Garibaldi & Turner, 2004). When coupled with temporal and geographic information, measures of species popularity can detect seasonal changes in interest (Mittermeier, Roll, Matthews, & Grenyer, 2019), identify which demographic groups are likely to support a specific conservation initiative (Correia, Jepson, Malhado, & Ladle, 2016), and help reveal situations where species may be popular amongst one group of stakeholders, but not another (Bowen-Jones & Entwistle, 2002; Mulder, Schacht, Caro, Schacht, & Caro, 2009).

The quantity and availability of digital data has expanded exponentially since the start of twenty-first century (Kitchin, 2014), and the emergence of this digital 'big data' has had a transformative impact across a range of research fields (Mayer-Schönberger & Cukier, 2013). The ways in which people navigate this ever-accumulating online information can provide understandings of human preferences and patterns of behavior (Lazer et al., 2009). Among other examples, Twitter posts have been used to predict changes in the stock market (Bollen, Mao, & Zeng, 2011), and Wikipedia pageviews to identify the timing and location of disease outbreaks (Generous, Fairchild, Deshpande, Valle, & Priedhorsky, 2014; Hickmann et al., 2015). In conservation, the potential to apply digital big data to questions of human interest in biodiversity has led to the development of the new research area of 'conservation culturomics' (Ladle et al., 2016). The digital data generated by human activities online has been used to quantitatively compare interest in reptile species (Roll et al., 2016) and measure interest in different species of birds in Brazil (Correia et al., 2016). While these approaches hold promise for a range of conservation applications, methods for applying these analyses to conservation are still in their infancy and require further development and critique (Ladle et al., 2016; Sutherland et al., 2018).

Interpreting the outputs from digital data can be complex and using this information uncritically has the potential to result in misleading conclusions (Correia et al., 2019).

Here we explore the popularity of global bird species in Wikipedia, the large online encyclopedia. Wikipedia has several features that make it well-suited to compare interest across different species (see below). Though several of these are unique to Wikipedia, many of the insights gained from Wikipedia are applicable to online data in general. As such, we provide relevant methodological considerations for making quantitative comparisons of interest in biodiversity using online digital data in general.

Dataset selection—why Wikipedia?

A variety of digital information archives are amenable to culturomic analyses; Bing search data, Web of Science, Google Trends, Twitter, Facebook, Instagram, YouTube or digitized newspapers to name a few. The choice of dataset is crucial as each may reflect a distinct user base as well as specific types of engagement with online material. For example, content consumption (readership, views in places such as digitized newspapers) differs from content generation (present on blogs or micro-blogs such as Twitter) and content engagement (likes and comments on sites such as Instagram or Facebook). Additionally, online platforms differ in their inherent biases, technical challenges and often the type of data that are made accessible to researchers. As of 2019, for example, Facebook allows access to raw data, but only a limited selection of it, whereas Google provides a greater range of data but only makes data available in a pre-analyzed format. Unsurprisingly, some datasets will be better suited to certain questions than others.

Here we focus on Wikipedia for five reasons: (a) *popularity*; as of 2019, Wikipedia is the tenth most popular site on the internet in terms of global web traffic and is the most visited site online that is not a search engine, social network or online shopping site (Alexa, 2019). Currently it receives upwards of 16 billion pageviews per month across its associated projects (https://stats.wikimedia.org/EN/TablesPageViewsMonthlyAllProjects.htm). (b) *Coverage*; in addition to the sheer number of visits Wikipedia receives, the scope of its coverage is noteworthy. Wikipedia includes 304 language editions with over 203 million total pages that have received more than 2.58 billion edits (https://meta.wikimedia.org/wiki/List_of_Wikipedias) at the time of this writing. Furthermore, Wikipedia has integrated with taxonomic databases such as the Global Biodiversity Information Facility (www.gbif.org), the Encyclopedia of Life (www.eol.org), and the Integrated Taxonomic Information System (www.itis.gov), and as a result contains pages for tens of thousands of biodiversity-related entities. This breadth of coverage is reflected in the fact that we were able to extract pages for over 95% of bird species listed in a widely used global taxonomy (Clements et al., 2018). (c) *Structure*; Wikipedia pages are categorized and linked together via Wikidata, the encyclopedia's underlying data structure, and metadata about each page in the encyclopedia is collected and summarized in a standardized format. This organization enables comparisons between entities and across large numbers of languages that would be difficult, if not impossible, in other digital text formats. (d) *Access*; Wikipedia is open-access and all of its raw data are available for download. This enables transparent and repeatable research. Other large online platforms limit direct access to their raw data or only provide analytics produced through their own, often secret, search algorithms (Ladle et al., 2016; Malcevschi, Marchini, Savini, & Facchinetti, 2012). Of the ten most trafficked sites on the internet, Wikipedia is the only one to allow open access to its raw data. (e) *Existing*

research; Wikipedia is the subject of a large and growing body of research that investigates aspects of its coverage (Messner & DiStaso, 2013; Samoilenko & Yasseri, 2014), contributor demographics (Wilson, 2014), and user dynamics (Yasseri, Spoerri, Graham, & Kertész, 2014; Yasseri, Sumi, Rung, Kornai, & Kertész, 2012). Pertinent to the question of comparing the interest across species, Wikipedia has been used to quantitatively compare the fame and cultural impact of individual people (Skiena & Ward, 2014a; Yu, Ronen, Hu, & Hidalgo, 2016). The Pantheon project by Yu *et al*., in particular, proposes a method for indexing globally significant people using Wikipedia and provides a template for accounting for variation in the number of language editions in which a page appears and how Wikipedia pageviews are distributed amongst languages (Yu et al., 2016).

In this context, Wikipedia data is well-suited to making broad-scale comparisons in popularity across large numbers of entities and languages. A challenge with Wikipedia data for this type of analysis is the lack of geographic information associated with the pageview data (a feature that Wikipedia deliberately does not make public for privacy reasons) which makes it impossible to precisely locate where pageviews originate. Following previous studies, we use language as a coarse proxy for geography in Wikipedia (Generous et al., 2014; Mittermeier et al., 2019).

Methods

We used the Wikidata Query Service (https://query.wikidata.org) to extract a list of 12,855 Wikidata entities tagged with an "eBird taxon ID" (Wikidata property: P3444) on 29 April 2019, and matched these entities to the eBird/Clements checklist of birds of the world, v2018 (Clements et al., 2018) using eBird species codes. For each Wikidata entity in our list, we

scraped the associated sitelinks with rvest (Wickham, 2019). We obtained daily pageviews for the resulting urls generated by users (bot and spider views were excluded) on all platforms (desktop, mobile-app and mobile-web) for the period between 01 July 2015 and 01 May 2019 (1,401 days) using 'pageviews' (Keyes & Lewis, 2016). To avoid confusion, it is important to note the discrepancy in dates here: 29 April marks the date that we extracted the list of pages, while 01 May marks the day up to which we counted views for those pages. In other words, a page created on 30 April, though it may have received views on 01 May, would not be included in our dataset. Meanwhile, a page removed from Wikipedia on 30 April would appear in our dataset but would be counted as having zero views on 01 May. The discrepancy between the two dates results from the time needed for our code to run.

We filtered our Wikidata entities to only include those that eBird/Clements v2018 classified as "species," and for each species-page, we calculated the sum views over the time period as well as a weighted mean of daily views using Tukey's biweight (Bunn et al., 2018). For pages in English-language Wikipedia we scraped twelve additional attributes related to the page's size, edit history, and creation. Following these methods, we obtained a list of 218,226 pages for 10,137 bird species (95.7% of the total species listed in eBird/Clements v2018) across 274 Wikimedia and Wikipedia platforms. These pages received over 765M pageviews over the time period sampled.

Similarity between lists of species in different languages and amongst metadata types was assessed using Euclidean distance matrices. For each variable, data were scaled to a mean of 0 and standard deviation of 1. Distance matrices were visualized with agglomerative hierarchical clustering dendrograms using Ward's minimum variance method (Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2019). For geographic comparisons, we defined the distribution of a language

as the borders of the country responsible for the highest proportion of pageviews in that language (Mittermeier et al., 2019; Zachte, 2018). All data analyses and visualizations were done using R (R Core Team, 2019).

Methodological considerations in comparing the popularity of species using Wikipedia

We identify six considerations that need to be taken into account when using Wikipedia data to compare the popularity of species.

1. Data selection

Wikipedia pages have a wide array of associated data which can be quantified and compared across pages. As of May 2019, for example, each Wikipedia page includes measures of at least 32 continuous variables in its general statistics that relate to its size, edit history, edit frequency, and links to other pages in Wikipedia (accessible via https://xtools.wmflabs.org). These metadata do not necessarily reflect the actions of the same communities of uses nor the same forms of online engagement. Wikipedia editors or "Wikipedians," for example, skew strongly male and are a significantly smaller community than the people who view the encyclopedia (Lam, Uduwage, Dong, Sen, & David, 2011). Specific metadata types may also reflect distinct forms of engagement; such as contributors hoping to gain recognition for their editing (Iba, Nemoto, Peters, & Gloor, 2010). Page metadata also vary in the quantity of interactions they contain. In our dataset, English-language bird pages received more than 291 million pageviews over nearly four years compared to 901,000 edits counted over the entire history of Wikipedia. Other metadata categories may have even less data or may be influenced by the actions of only a small number of users or the activity of 'bots', automated programs that edit and contribute data to

Wikipedia. The sequence in which articles were added to Wikipedia (i.e. their age), for example, often the results from bot activity rather than the input of human users (nearly half of all English-language bird-pages were created over a four-day period in 2007 by a single bot).

In this context, it is unsurprising that the highest-ranking species differ across metadata types (Fig. 1) and that there is often limited correlation between the rankings created by different metadata types (Fig. 2). Simply compiling information across metadata types without critically considering which inputs are used may reveal meaningless outputs. Instead, certain types of data will be useful for answering particular questions, such as edit histories being reflective of controversial pages (Yasseri et al., 2012).

Pageviews have three clear advantages over other metadata types in measuring human interest in species. 1) Pageviews reflect a distinct type of interaction (learning more about a particular subject); 2) pageviews capture the actions of the widest community of users (there are many more Wikipedia readers than editors) and are least likely to be impacted by the activities of a small number of individuals; 3) pageviews contain the largest quantity of data amongst Wikipedia metadata types; and have a published precedent for being used to compare the popularity (Skiena & Ward, 2014b; Yu et al., 2016). Wikipedia's reclassification of its pageviews in 2015 allows for differentiation of user as opposed to bot-generated 'views' circumventing a previous challenge with this source. One drawback of pageviews pertinent to many conservation questions is that they do not capture sentiment. It is not possible to distinguish whether a viewer reached a page because they felt positively or negatively about a particularly subject.

2. Temporal variation

As an open-access, user-generated resource, Wikipedia is constantly being updated and revised. As of spring 2019, Wikipedia receives nearly 50 million edits and gains 33 GB of content per month (https://stats.wikimedia.org/v2/#/all-projects). An extraction of information made in January therefore will differ from one made in May, and older versions will have fewer pages and less content than newer ones. In order to be repeatable, methods using Wikipedia should identify the date a list of pages in the encyclopedia was identified as well as the timeframes over which the metadata associated with those pages was collected. In addition to its overall patterns of growth, Wikipedia activity can follow seasonal patterns and comparisons made over short temporal spans may be influenced by intra-annual fluctuations (Mittermeier et al., 2019). As Mittermeier *et al.* (2019) point out, the level of interest in species in Wikipedia depends, in part, on the time of year. These seasonal patterns can be compensated for by extracting data over multi-year periods as opposed to shorter intervals. In addition to seasonal patterns, Wikipedia pages can undergo short bursts of attention and the pages that receive the highest sum views over a given period of time may not necessarily be the ones that consistently receive the most daily views (Fig. 3). For specific questions, such as assessing the impact of a publicity campaign, this will be of interest, while for other questions, such as measuring consistent interest in a species it is worth considering measures that weight against short bursts of interest (such as a weighted mean of daily views).

3. Taxonomy

When comparing interest across species in Wikipedia the underlying taxonomy used to determine which pages are included in the assessment may influence the results. For example, the widespread Barn Owl is considered a single species by some taxonomies (Clements et al.,

2018) and a cluster of three closely related species by others (Gill & Donsker, 2019). Wikipedia has pages for both the more inclusive species ("Barn owl") and the three split ones ("Eastern barn owl", "Western barn owl," and "American barn owl"), but pageviews are highly unevenly distributed amongst them. Over the time frame of our study, the English-language page for "Barn Owl" received >1.35 million pageviews whereas none of the pages for the split species received >24,000 views. Even when added together, the total views for the three split species-pages only amount to 0.05% of those received by the more inclusive page. An assessment following the splitting taxonomy might conclude that barns owls have relatively low public appeal, whereas one using the taxonomy with the single species would identify barn owls as being amongst the most popular of all birds.

Taxonomies also differ in their degree of integration with Wikipedia. Querying Wikidata for all species marked with an "eBird taxon ID" (Wikidata property: P3444) returned 12,855 entities at the time of our study. Other potential tags for extracting lists of birds include the "Internet Bird Collection species ID" (P3099; 794 entities as of 29 May 2019), the "Avibase ID" (P2026; 8745 entities as of 29 May 2019), and the BirdLife taxon ID (P5257; 11,509 entities as of 29 May 2019). The most appropriate taxonomy will depend on the questions being asked. Furthermore, since public interest does not necessarily correspond to precise taxonomic units, in some cases compiling information from multiple pages may be the best approach. Whatever the case, the choices made in terms of the underlying taxonomy used should be explicitly stated and justified.

4. Language representation

Wikipedia contains 304 different language editions as of May 2019, and these vary dramatically in size from zero articles (i.e. only an introductory main page) to over 5.87 million articles in the English language Wikipedia. The distribution of information amongst languages is highly skewed. The ten most-viewed language editions account for ca. 88% of all pageviews (https://stats.wikimedia.org/EN/TablesPageViewsMonthlyCombined.htm). English, in particular, is dominant in Wikipedia and receives 49% of all pageviews and 25% of all edits (https://stats.wikimedia.org/EN/TablesPageViewsMonthlyCombined.htm, https://meta.wikimedia.org/wiki/List_of_Wikipedias). Given this inequality, multilingual comparisons that sum data across language editions will be strongly influenced by English. It is worth noting that many English-language pageviews originate from countries where English is not the majority language (together the world's six large majority English-speaking countries account for just under 60% of the English Wikipedia pageviews; Zachte, 2018) and thus many viewers to English language Wikipedia probably do not speak English as their first language.

Given these linguistic inequalities, unscaled multilingual comparisons derived from Wikipedia data, though they may contain information from nearly 300 languages, will predominately reflect only a small subset of those languages. Multilingual comparisons that seek to represent information from smaller language editions will have to weight results by language in order to better capture data in those languages. It is also important to keep in mind that the ca. 300 languages in Wikipedia are less than 5% of 7,111 recognized languages currently spoken on Earth (Eberhard, Simons, & Fennig, 2019). Clearly, many perspectives are not captured by Wikipedia data and as with any data sources, the interpretation of results from Wikipedia need to be carefully contextualized. As Funk and Rusowky (2014) point out in the context of Google "results from Google searches exclude this substantial group of people [who do not use the

internet] and are, thus, not a 'worldwide' representation and should, therefore, not be used as a single proxy for human interest and as a main pillar to inform conservation communication" (Funk & Rusowsky, 2014). Clearly the same applies to Wikipedia, and the linguistic and geographic context of the users represented needs to be taken into account.

5. Wikipedia language geography

Wikipedia languages are uneven in their geography, and their distribution does not follow overall patterns of linguistic diversity. While over 60% of spoken languages occur in Africa and Asia with tropical countries such as Nigeria and Papua New Guinea being particularly diverse (Eberhard et al., 2019), the majority of Wikipedia languages are European. Languages from sub-Saharan Africa, the Pacific and South America are underrepresented in Wikipedia. With the notable exception of China, this pattern mirrors the distribution of global internet access in general (Graham, 2014). Home to the world's largest population of internet users, China has intermittently blocked access to Wikipedia (something it has also done for a variety of other widely used internet platforms) and so, like several other datasets, data from China are not well represented in Wikipedia-derived metrics ("Censorship of Wikipedia," 2018; Miniwatts Marketing Group, 2018).

In comparisons of public interest in biodiversity, the distribution of Wikipedia languages is relevant to being aware of who's activities are reflected in the data. Furthermore, multi-lingual comparisons made from Wikipedia data will produce results that are strongly influenced by the geographic distribution of Wikipedia languages (Fig. 4, 5).

6. Biogeography

In addition to the geography of Wikipedia languages, the distribution of the features being compared (in this case species) may impact measures of popularity. In our data this is reflected in the geographic structure of bird popularity amongst languages. Across languages, species that occur within the country that accounts for the majority of a language's pageviews feature prominently amongst the most popular species-pages in that language (Fig. 4), and languages that are geographically proximate to one another tend to create more similar rankings of pageviews (Fig. 5). As a result of this biogeographic structure, comparisons of popularity that do not account for the geographic distribution of the species will be biased towards those species whose distribution overlaps with the distribution of the languages being assessed.

Case study application: what are the most popular birds in Wikipedia?

To put these methodological considerations into practice, we use Wikipedia to compare methods for identifying the most popular global bird species. This type of data-driven assessment of species popularity could be used to identify flagships species, investigate the traits that lead to interest in species, or heighten attention for specific conservation measures. Furthermore, making this assessment at a global scale is relevant since many conservation actions seek to gain funding and support for international activities.

Case study: methods

We obtained raw data from Wikipedia and Wikidata using the methods previously described. In order to provide a template for best practice, we identify each of the methodological choices used in our approach.

1. *Choice of metadata*—to assess the popularity of species amongst a wide group of users we used Wikipedia pageviews. We limited our data to pageviews from human users (removing bot-generated 'views') and obtained user-generated views from desktop, mobile-app and mobile-web sources (Keyes & Lewis, 2016). We filtered our analysis to pageviews from Wikipedia language editions and excluded pageviews to other Wikimedia projects (e.g. Wikibooks, Wikiquote, Wikispecies, etc.).

2. *Temporal variation*: we specify the date that our list of Wikipedia pages was extracted (29 April 2019) and obtained information for all pages in our dataset concurrently to facilitate comparability. To minimize the impact of seasonal variations, we selected a long time series of pageviews (1,401 days).

3. *Taxonomy*: we used the "eBird taxon ID" (Wikidata property: P3444) as the basis for our definition of what qualifies as a bird in Wikidata, and verified our species list by matching it to the eBird/Clements world bird list v. 2018 (Clements et al., 2018). The eBird/Clements taxonomy is a well-established and widely used global taxonomy and had the highest degree of integration with Wikidata at the time of our study. As discussed in the context of the barn owl example above, choosing a different taxonomy could potentially alter some of our results. Within the eBird/Clements list we restricted our analysis to pages for species (category "species" in eBird/Clements). Though public interest in biodiversity does not necessarily aggregate to species units (subspecies, races or broader family pages may be popular as well), we selected species because they are the most frequently used taxonomic unit in biodiversity assessments. However, this methodological choice could lead to groups that have been subject to recent taxonomic revisions or groups where public interest coalesces at different levels of the taxonomic hierarchy being underrepresented.

4. *Language representation*: in order to identify species that attract high interest across a range of languages, we adapted the method of Yu et al. (2016) to weight our results for species that (a) had pages in multiple language editions (assessed by the total number of language editions per species), and (b) had views to pages evenly distributed amongst languages rather than concentrated in a single edition (assessed using the entropy in pageviews and coefficient of variation across languages). Furthermore, we accounted for the predominance of English in Wikipedia by providing an additional weighting for views in non-English Wikipedia editions. These methods are described in detail by Yu et al. (2016).

5. *Wikipedia geography*: we did not provide a measure to weight languages based on their geographic distribution. Wikipedia-usage and Wikipedia-languages are disproportionately skewed towards Europe and the Global North, and therefore our results primarily the views of an internet-using public that resides in Europe, North America and parts of east and south Asia.

6. *Biogeography*: to identify species that are attract high interest irrespective of their local abundance, we assigned the species in our list to one of fifteen 'general regions' based on their breeding distribution listed in Gill and Donsker (2019). Languages were assigned to the same geographic regions using information from the CIA World Factbook field listing for languages (Central Intelligence Agency, 2019) and Glottolog (Hammarstrom, Forkel, & Haspelmath, 2019). Each species was then given an additional score for pageviews received in languages outside of its area of distribution.

Considering each of these variables, we compared three metrics for indexing interest in species across languages. First, a raw score of species that received the most total views with no weighting.

(a) $Popularity = \sum pageviews$

Second, we used an adapted version of Yu et al.'s (2016) method for weighting species according to a) the number of language editions they occur in, b) the distribution of pageviews across those language editions, and c) the pageviews they receive in non-English language editions.

(b) $Popularity = ln(L) + ln(L^{*}) + ln(pageviews^{non\text{-}English}) - ln(CV)$

In this equation, L = the total number of language editions per species; L^{*} = the effective number of language editions taking into account the entropy [spread] of views amongst languages; $pageviews^{non\text{-}English}$ = views received to Wikipedia editions other than English; CV = the coefficient of variation in views amongst languages for each species.

Finally, we further modified Yu et al.'s method by adding an additional weighting for views that a species received outside the biogeographic region where it occurs (pageviews non-range).

(c) $Popularity = ln(L) + ln(L^{*}) + ln(pageviews^{non\text{-}English}) - ln(CV) + ln(pageviews^{non\text{-}range})$

For equations (b) and (c) we calculated the relative contribution of each variable in the equation on the overall score using R^2 partitioned by averaging over orders (Gromping, 2006; Lindeman,

Merenda, & Gold, 1980). For further description of variables (aside from *pageviews* $^{non\text{-}range}$) see the discussion of methods in Yu et al. (2016).

Case study: results

Our initial Wikidata query returned 12,855 entities tagged with an "eBird taxon ID." Of these, 12,829 (99.8%) matched to an "eBird species code" in the v2018 eBird/Clements world list. After filtering our list to only include birds in the "species" category of eBird/Clements, we were left with a 10,099 eBird/Clements bird species that had a page in at least one Wikipedia language edition. This represents 95.4% of the species listed in the eBird/Clements v2018 world list (the exclusion of non-Wikipedia platforms results in these totals being lower than those described above). Our dataset included 199,699 pages across 251 Wikipedia language editions. These pages received nearly 757 million pageviews over the period of our study, with weighted mean of 486,000 pageviews per day across all pages.

The distribution of pageviews across languages was highly uneven both in terms of the total overall views per language edition (range 1-290 million, mean 3.01 million) and the mean daily views per language edition (range 0-190,000, mean 1,940). English was by far the largest language edition with views to English-language pages accounting for 38.3% of all pageviews. Together with English, the ten most-viewed languages in our dataset (English, German, Spanish, Russian, Japanese, French, Polish, Dutch, Italian and Portuguese) accounted for 81.3% of the bird pageviews.

Of the 10,099 species in our dataset, 9,861 (97.6%) matched to the biogeographic regions in Gill and Donsker (2019). Those that did not match were largely the result of taxonomic differences between Gill and Donsker and eBird/Clements. Species tended to be more popular in

languages with which they overlapped (mean total overlapping views: 61,200 views per species, mean total non-overlapping views: 12,400 views per species) and as a result our inclusion of a weighting for non-range views impacted the relative ranking of species.

While the three methods we compared (equations a, b, and c) all correlated positively with one another, they also generated notable differences in how they ranked species. Only two species (0.03%) appeared amongst the top twenty most popular birds across all three rankings, while 22 (36.7%) appeared in the top twenty for just one of the three. Overall, the two weighted rankings (equations b and c) correlated more strongly with each other ($p < 2.2E-16$, Pearson's r = 0.91) than with the ranking derived from sum pageviews (equation a vs. b: $p < 2.2E-16$, Pearson's r = 0.85; equation a vs. c: $p < 2.2E-16$, Pearson's r = 0.84). The ranking of species derived from the sum pageviews across all languages correlated strongly with a ranking of species from only English language pageviews (equation a vs. English-language pageviews: $p < 2.2E-16$, Pearson's r = 0.96). Results of the R^2 partitioning demonstrated that in equation (b) the views received to non-English language editions was the most significant variable in account for variation in scores (accounting for 40.0% of the total variation) followed by the number of language editions (33.6% of the variation). In equation (c) views received in languages outside of the geographic distribution of a species was the most significant variable (31.0% of variation), closely followed by pageviews in non-English Wikipedias (29.3%) and the total language editions (24.1%). The top 20 most popular bird species according the popularity index that accounts for the both Wikipedia language size and species distribution (equation c) are shown in Fig. 6.

Case study: discussion

We present three examples of data-driven indices for quantifying the popularity of global bird species across more than 250 different languages. By combining a large quantity of data (nearly 757 million pageviews) across a wide range of cultural and geographic contexts with a clear and repeatable methodology, these approaches are an effective way to identify popular species and investigate the drivers of public interest in biodiversity.

Though the results of the three indices correlate strongly with one another (Pearson's $r > 0.84$), they have significant differences. This demonstrates that the method selected plays an important role in how the popularity of species is assessed. In particular, these differences emphasize that both biogeography and linguistic geography influence interest in species in Wikipedia.

Rather than one index being the single best metric, we argue that each has specific strengths and weaknesses. For the purpose of identifying global flagship species with high conservation interest, a method that accounts for both the distribution of species and languages (i.e. equation c) is the most useful. This metric highlights species that attract interest from a wide range of users across many different languages and that receive attention beyond their area of geographic distribution. However, this method underrepresents species that occur in English-speaking regions of the world and tends to emphasize species that appear in the pet trade (e.g. Budgerigar, Cockatiel, Gray Parrot).

These methods represent a 'first-pass' at ranking the popularity of biodiversity features using digital online data. Specific uses of indexing the popularity of species in Wikipedia include: 1) identifying potential flagship species, 2) heightening the profile of projects that include such species, 3) identifying demographics that are likely to support the conservation of a particular species, and 4) understanding what drives human interest in different species. There

are several avenues through which future studies could refine and improve these results. Employing more fine-scale geographic data, for example, could help to refine the rankings of species. Finally, it is important to point out that our list defines interest by number of views a page receives and does not attach any sentiment to that interest. As a result, the list does not distinguish between views that originate because people like a species and those that result from people not liking a species or wanting to catch, hunt or remove it.

Conclusion

Our results demonstrate the significant potential of online data to assess interest in biodiversity across large numbers of biodiversity features, multiple cultural contexts, and vast numbers of users. As we demonstrate with species, Wikipedia provides a powerful tool for making comparisons between different aspects of biodiversity. In addition to species, Wikipedia could also be useful for comparing interest across other types of entities such as protected areas, concepts related to conservation and biodiversity, or groups of organisms at different levels of the taxonomic hierarchy. Additionally, other types of Wikipedia data such as edits could present interesting and relevant avenues for investigation.

Many of the methodological considerations that we highlight here in the context of Wikipedia are also relevant to quantitative comparisons of biodiversity features using other digital big data. When making assessments using Google Trends data, for example, it is equally important to consider features such as temporal variation, taxonomy, the geography of users and the geography of species as it is with Wikipedia. By the same token, since every dataset will have its own particular nuances, there will also be specific points to consider in the context of whichever dataset is used. Given that every data source will have its own particular strengths and

weaknesses, future measures of popularity and human interest in biodiversity will undoubtedly want to consider methods for incorporating multiple data sources. Here we introduce Wikipedia into the larger context of this emerging field.

References

Alexa. (2019). The top 500 sites on the web. Retrieved June 1, 2019, from https://www.alexa.com/topsites

Bollen, J., Mao, H., & Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Bowen-Jones, E., & Entwistle, A. (2002). Identifying appropriate flagship species: the importance of culture and local contexts. *Oryx*, *36*(02), 189–195.

Bunn, A., Korpela, M., Biondi, F., Campelo, F., Merian, P., Qaedan, F., … Wernicke, J. (2018). dplR: Dendrochronology Program Library in R. CRAN.R-project.

Caro, T. (2010). *Conservation by proxy: indicator, umbrella, keystone, flagship, and other surrogate species*. New York, New York, USA: Island Press.

Censorship of Wikipedia. (2018). Retrieved February 7, 2018, from https://en.wikipedia.org/wiki/Censorship_of_Wikipedia

Central Intelligence Agency. (2019). The world factbook. Retrieved August 25, 2017, from https://www.cia.gov/library/publications/the-world-factbook/

Clements, J. F., Schulenberg, T. S., Iliff, M. J., Roberson, D., Fredericks, T. A., Sullivan, B. L., & Wood, C. L. (2018). The eBird/Clements checklist of birds of the world: v2018. Retrieved April 29, 2019, from http://www.birds.cornell.edu/clementschecklist/download/

Correia, R. A., Di Minin, E., Jarić, I., Jepson, P., Ladle, R., Mittermeier, J., … Veríssimo, D.

(2019). Inferring public interest from search engine data requires caution. *Frontiers in Ecology and the Environment*, *17*(5), 254–255.

Correia, R. A., Jepson, P. R., Malhado, A. C. M., & Ladle, R. J. (2016). Familiarity breeds content: assessing bird species popularity with culturomics. *PeerJ*, *4*, 1–15.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2019). *Ethnologue: languages of the World. Twenty-second edition.* Dallas, Texas: SIL International. Retrieved August 1, 2019, from https://www.ethnologue.com/

Funk, S. M., & Rusowsky, D. (2014). The importance of cultural knowledge and scale for analysing internet search data as a proxy for public interest toward the environment. *Biodiversity Conservation*, 3101–3112.

Garibaldi, A., & Turner, N. (2004). Cultural keystone species: Implications for ecological conservation and restoration. *Ecology and Society*, *9*(3).

GBIF: The Global Biodiversity Information Facility. (2019). What is GBIF? Retreived August 1, 2019, from https://www.gbif.org/what-is-gbif

Generous, N., Fairchild, G., Deshpande, A., Valle, S. Y. Del, & Priedhorsky, R. (2014). Global disease monitoring and forecasting with Wikipedia. *PLoS Computational Biology*, *10*(11).

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, *438*(December), 900–901.

Gill, F., & Donsker, D. (Eds.). (2019). *IOC World Bird List (v9.1)*.

Graham, M. (2014). Inequitable Distributions in Internet Geographies: The Global South Is Gaining Access, but Lags in Local Content. *Innovations: Technology, Governance, Globalization*, *9*(3–4), 3–19.

Gromping, U. (2006). Relative importance for linear regression in R: the package realimpo. *Journal of Statistical Software*, *17*(1), 1–27.

Hammarstrom, H., Forkel, R., & Haspelmath, M. (2019). Glottolog 3.4. Retrieved August 25, 2017, from http://glottolog.org/

Hickmann, K. S., Fairchild, G., Priedhorsky, R., Generous, N., Hyman, J. M., Deshpande, A., & Del Valle, S. Y. (2015). Forecasting the 2013–2014 Influenza Season Using Wikipedia. *PLoS Computational Biology*, *11*(5), 1–29.

Iba, T., Nemoto, K., Peters, B., & Gloor, P. A. (2010). Analyzing the creative editing behavior of wikipedia editors through dynamic social network analysis. *Procedia - Social and Behavioral Sciences*, *2*(4), 6441–6456.

Keyes, O., & Lewis, J. (2016). pageviews: an API client for wikimedia traffic data version 0.3.0. Retrieved January 1, 2017, from https://cran.rstudio.com/web/packages/pageviews/index.html

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(June), 1–12.

Ladle, R. J., Correia, R. A., Do, Y., Joo, G. J., Malhado, A. C. M., Proulx, R., … Jepson, P. (2016). Conservation culturomics. *Frontiers in Ecology and the Environment*, *14*(5), 269–275.

Lam, S. T. K., Uduwage, A., Dong, Z., Sen, S., & David, R. (2011). WP: Clubhouse ? An Exploration of Wikipedia's Gender Imbalance. *Proceedings of the 7th international symposium on Wikis and open collaboration*, 1-10.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., … Van Alstyne, M. (2009). Life in the network: the coming age of computational social science. *Science 323*(5915), 721–723.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to Bivariate and*

Multivariate Analysis. Glenview, IL: Scott, Foresman.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2019). cluster: Cluster Analysis Basics and Extensions.

Malcevschi, S., Marchini, A., Savini, D., & Facchinetti, T. (2012). Opportunities for Web-Based Indicators in Environmental Sciences, *7*(8).

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: a revolution that will transform how we live, work and think*. Boston, MA: Houghton Mifflin Harcourt.

Messner, M., & DiStaso, M. W. (2013). Wikipedia versus Encyclopedia Britannica: A Longitudinal Analysis to Identify the Impact of Social Media on the Standards of Knowledge. *Mass Communication and Society*, *16*(February), 465–486.

Miniwatts Marketing Group. (2018). Top 20 countries with the highest number of internet users. Retrieved February 8, 2018, from http://www.internetworldstats.com/top20.htm

Mittermeier, J. C., Roll, U., Matthews, T. J., & Grenyer, R. (2019). A season for all things: phenological imprints in Wikipedia usage and their relevance to conservation. *PLoS Biology*, *17*(3), e3000146.

Mulder, M. B., Schacht, R., Caro, T., Schacht, J., & Caro, B. (2009). Knowledge and attitudes of children of the Rupununi: Implications for conservation in Guyana. *Biological Conservation*, *142*(4), 879–887.

R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Roll, U., Mittermeier, J. C., Diaz, G. I., Novosolov, M., Feldman, A., Itescu, Y., … Grenyer, R. (2016). Using Wikipedia page views to explore the cultural importance of global reptiles. *Biological Conservation*, *204*, 42–50.

Samoilenko, A., & Yasseri, T. (2014). The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics. *EPJ Data Science*, *3*, 1.

Skiena, S., & Ward, C. (2014a). *Who's bigger? Where historical figures really rank*. Cambridge, UK: Cambridge University Press.

Skiena, S., & Ward, C. (2014b). *Who's Bigger? Where historical figures really rank*. New York, New York, USA: Cambridge University Press.

Smith, R. J., Veríssimo, D., Isaac, N. J. B., & Jones, K. E. (2012). Identifying Cinderella species: Uncovering mammals with conservation flagship appeal. *Conservation Letters*, *5*(3), 205–212.

Sutherland, W. J., Butchart, S. H. M., Connor, B., Culshaw, C., Dicks, L. V., Dinsdale, J., … Gleave, R. A. (2018). A 2018 Horizon Scan of Emerging Issues for Global Conservation and Biological Diversity. *Trends in Ecology and Evolution*, *33*(1), 47–58.

Wickham, H. (2019). rvest: Easily Harvest (scrape) Web Pages. Retrieved from https://cran.r-project.org/web/packages/rvest/index.html

Wilson, J. L. (2014). Proceed With Extreme Caution: Citation to Wikipedia in Light of Contributor Demographics and Content Policies. *Vanderbilt Journal of Entertainment & Technology Law*, *16*(4), 857–908.

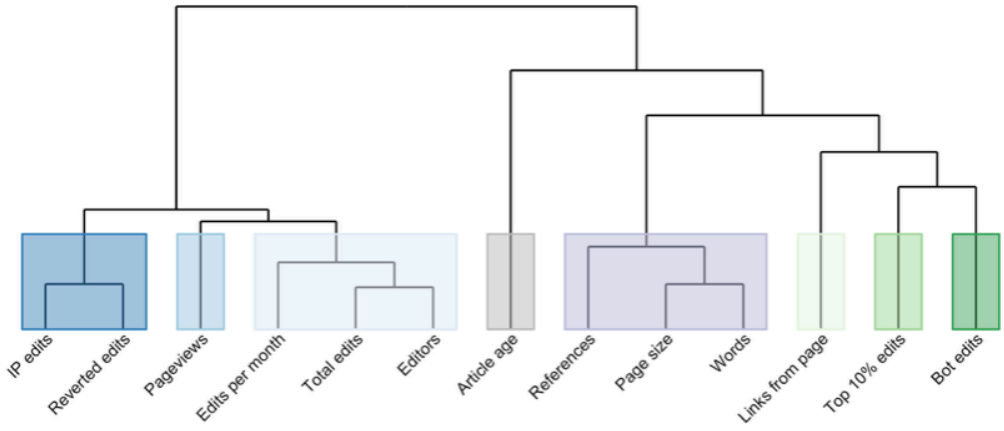
Yasseri, T., Spoerri, A., Graham, M., & Kertész, J. (2014). The most controversial topics in Wikipedia: A multilingual and geographical analysis. *Global Wikipedia: International and Cross-Cultural Issues in Collaboration*, 178.

Yasseri, T., Sumi, R., Rung, A., Kornai, A., & Kertész, J. (2012). Dynamics of conflicts in wikipedia. *PLoS ONE*, *7*(6), 1–12.

Yu, A. Z., Ronen, S., Hu, K. Z., & Hidalgo, C. a. (2016). Pantheon 1.0, a manually verified

dataset of globally fmaous biographies. *Scientific Data*, *2*(150075).

Zachte, E. (2018). Wikimedia Traffic Analysis Report: page views per wikipedia language. Retrieved July 23, 2018, from https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm



	Reverted edits	Pageviews	Total edits	Article age	Words	Edits by top editors
1	Snowy Owl	Dodo	Dodo	Peregrine Falcon	White-tailed Eagle	House Sparrow
2	Canada Goose	Bald Eagle	Common Ostrich	Great Auk	Red-tailed Hawk	Passenger Pigeon
3	Mallard	Peregrine Falcon	Peregrine Falcon	Dodo	Northern Goshawk	Peregrine Falcon
4	Blue Jay	Emu	Bald Eagle	Passenger Pigeon	Passenger Pigeon	Common Ostrich
5	Blue-footed Booby	Golden Eagle	Emperor Penguin	Labrador Duck	Common Buzzard	Northern Bald Ibis
6	Barn Owl	Osprey	Mallard	Carolina Parakeet	Great Horned Owl	Reunion Ibis
7	Emperor Penguin	Common Raven	Canada Goose	Ivory-billed Woodpecker	Martial Eagle	Eurasian Jackdaw
8	Budgerigar	Harpy Eagle	Budgerigar	Common Buzzard	Bonelli's Eagle	Common Raven
9	Red-tailed Hawk	Canada Goose	Passenger Pigeon	Little Penguin	Dodo	Mallard
10	Little Penguin	Emperor Penguin	Snowy Owl	Common Ostrich	Imperial Eagle	Huia

FIGURE 1. Wikipedia contains an array of metadata types and rankings of species depend on the type of metadata used. The dendrogram (top) demonstrates how rankings derived from different types of metadata cluster together based on their rank similarity. The table (bottom) shows the ten highest ranking bird species according to six different metadata types. Darker shading indicates species occurring in more than one of the listed top 10 rankings.

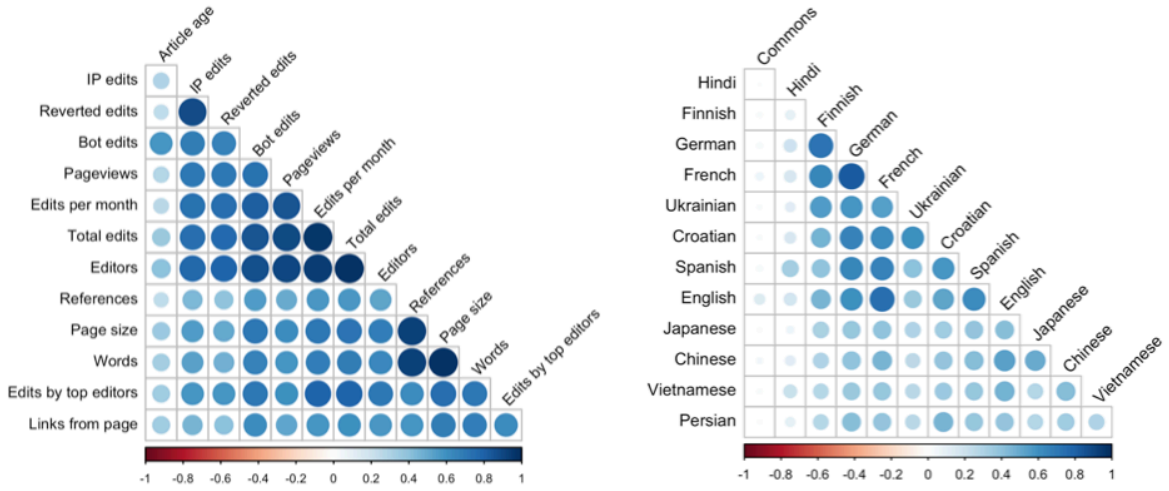


FIGURE 2. Correlation matrices demonstrating the similarity in rankings for popular bird species in Wikipedia according to different metadata types (left) and across different languages (right).

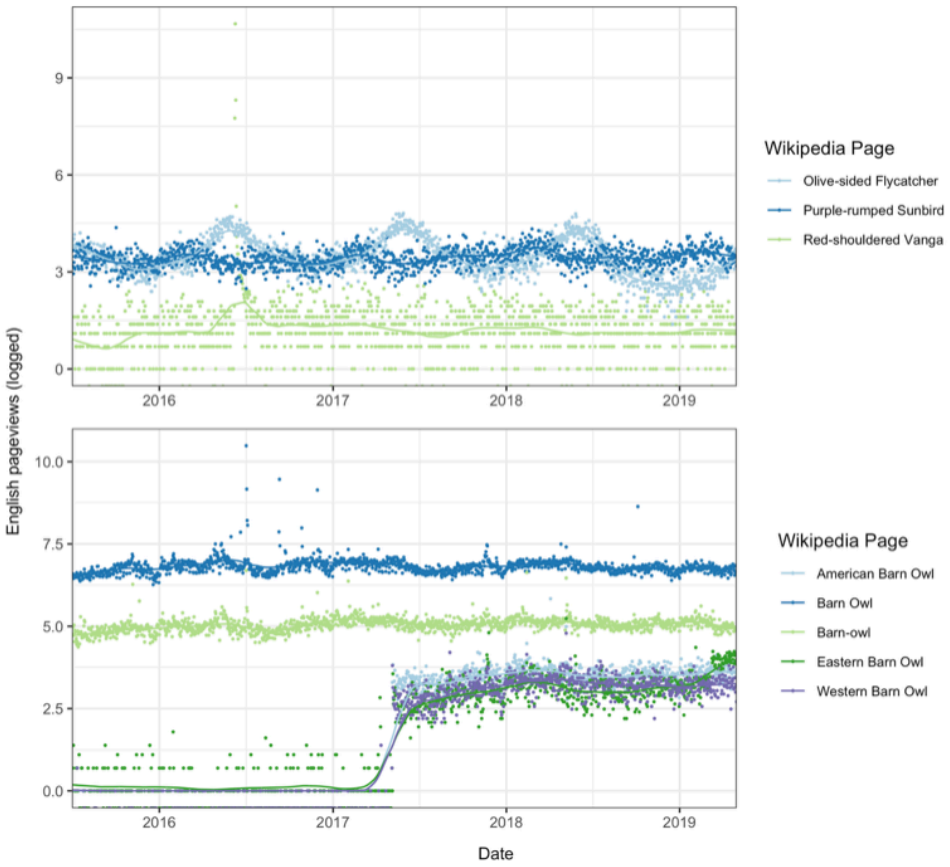


FIGURE 3. Two examples of methodological considerations when comparing pageviews across bird species in Wikipedia. Top: three bird species with similar total pageviews over the period between 2015-2019 achieve their overall total as a result of differing temporal patterns. One species has a relatively consistent interest, another undergoes seasonal fluctuations, while a third receives the majority of its views over a brief four-day period. Bottom: taxonomic differences impact popularity measurements. Neither of three 'split' barn owls (American Barn Owl, Eastern Barn Owl, Western Barn Owl), receive nearly as much attention as the more inclusive Barn Owl species in Wikipedia.

	Finnish	German	Ukrainian	Croatian	Commons
1	Whooper Swan	Eurasian Blackbird	White Stork	White Stork	American Robin
2	White-tailed Eagle	Eurasian Jay	Great Spotted Woodpecker	Eurasian Griffon	Snowy Owl
3	Eurasian Blackbird	Red Kite	Common Cuckoo	Carrion Crow	Emperor Penguin
4	Western Capercaillie	Harpy Eagle	European Starling	Peregrine Falcon	Blue Jay
5	Golden Eagle	Eurasian Eagle-Owl	Eurasian Bullfinch	House Sparrow	European Goldfinch
6	Eurasian Eagle-Owl	Common Kingfisher	Common Raven	Northern Goshawk	Imperial Eagle
7	Osprey	Golden Eagle	Common Ostrich	Common Ostrich	House Sparrow
8	Great Tit	Mallard	Red Crossbill	Common Raven	Hen Harrier
9	Eurasian Hoopoe	European Robin	Golden Eagle	Common Nightingale	Mute Swan
10	Common Crane	Dodo	Peregrine Falcon	European Goldfinch	Mallard

	Vietnamese	Hindi	Persian	Japanese	Spanish
1	Green Peafowl	Indian Peafowl	Budgerigar	Ural Owl	Andean Condor
2	Common Ostrich	House Sparrow	Cockatiel	Shoebill	Indian Peafowl
3	Eurasian Coot	Common Ostrich	Common Ostrich	Bull-headed Shrike	Common Kingfisher
4	Helmeted Guineafowl	Indian Roller	Gray Parrot	Barn Swallow	Harpy Eagle
5	Dodo	Common Hill Myna	Rosy-faced Lovebird	Crested Ibis	Dodo
6	Bald Eagle	Dodo	Bearded Vulture	Eurasian Tree Sparrow	Golden Eagle
7	Mandarin Duck	Chukar	White-eared Bulbul	Japanese Bush Warbler	Common Raven
8	Great Crested Tern	Pied Cuckoo	Rose-ringed Parakeet	Oriental Stork	Common Ostrich
9	Great Hornbill	Baya Weaver	Eurasian Hoopoe	White-cheeked Starling	Scarlet Macaw
10	Eurasian Magpie	Common Myna	Ring-necked Pheasant	Japanese White-eye	Peregrine Falcon

FIGURE 4. Popular species differ across Wikipedia languages. Top panel: the top 10 most viewed bird species across five Wikipedia languages. Darker shading indicates species that occur in more than one of the displayed languages. Bottom panel: a significant aspect of popularity of bird species pages is geographic overlap and across languages the most popular species are often those that occur where a language is primarily spoken. Species shaded green are those that occur in the same biogeographic region as the listed language.

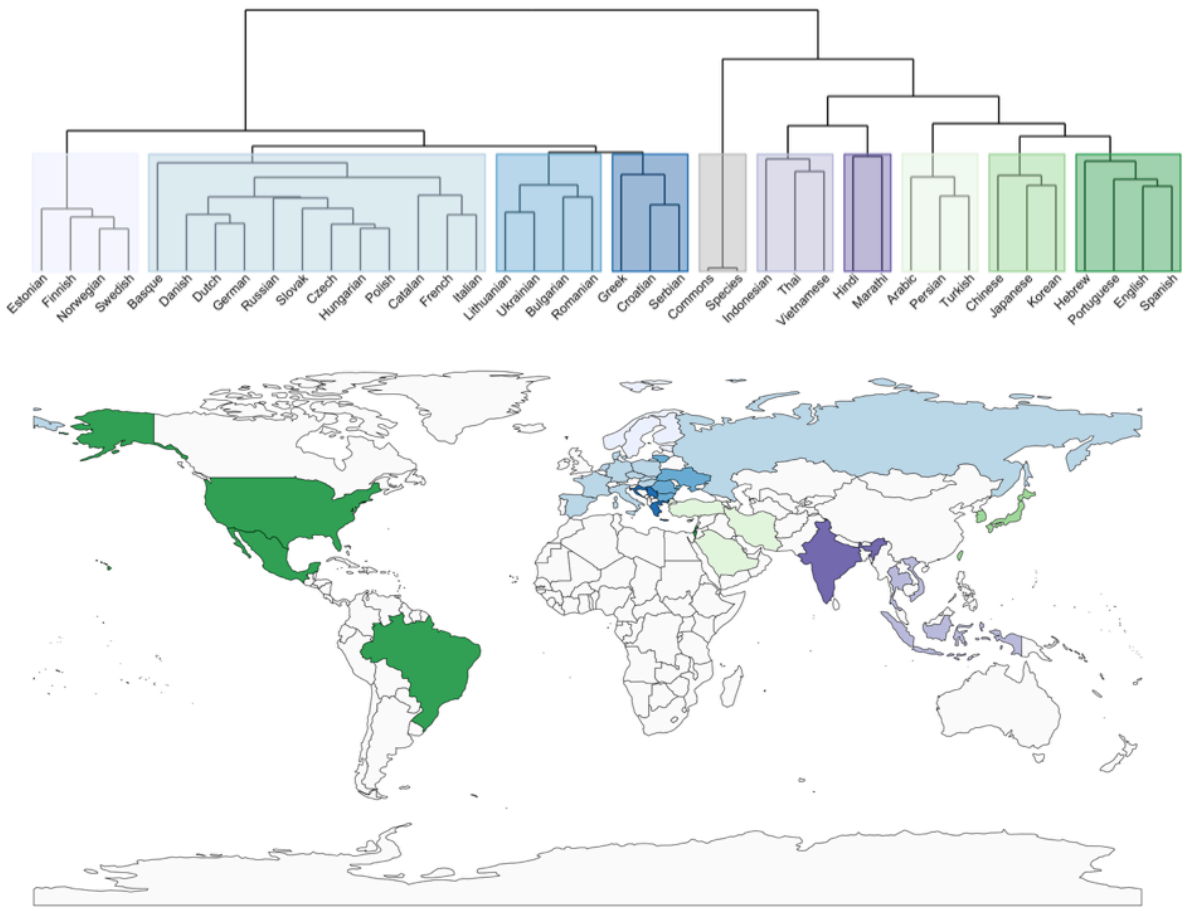


FIGURE 5. Popularity rankings of bird species in Wikipedia cluster geographically. Similarity amongst language rankings are shown in the dendrogram (top) while the geographic distribution of the languages are mapped below.

English overlap

	Daily views all	Popularity index	Species popularity index
1	Dodo	Common Ostrich	Common Ostrich
2	Golden Eagle	Eurasian Magpie	Dodo
3	Peregrine Falcon	Eurasian Hoopoe	Budgerigar
4	Bald Eagle	Eurasian Blackbird	Indian Peafowl
5	Common Raven	White Stork	Cockatiel
6	Common Ostrich	House Sparrow	Emperor Penguin
7	Budgerigar	Common Raven	Gray Parrot
8	Eurasian Blackbird	Golden Eagle	Andean Condor
9	House Sparrow	Common Cuckoo	Island Canary
10	Emu	Budgerigar	Rose-ringed Parakeet

European species

	Daily views all	Popularity index	Species popularity index
1	Dodo	Common Ostrich	Common Ostrich
2	Golden Eagle	Eurasian Magpie	Dodo
3	Peregrine Falcon	Eurasian Hoopoe	Budgerigar
4	Bald Eagle	Eurasian Blackbird	Indian Peafowl
5	Common Raven	White Stork	Cockatiel
6	Common Ostrich	House Sparrow	Emperor Penguin
7	Budgerigar	Common Raven	Gray Parrot
8	Eurasian Blackbird	Golden Eagle	Andean Condor
9	House Sparrow	Common Cuckoo	Island Canary
10	Emu	Budgerigar	Rose-ringed Parakeet

Pet trade species

	Daily views all	Popularity index	Species popularity index
1	Dodo	Common Ostrich	Common Ostrich
2	Golden Eagle	Eurasian Magpie	Dodo
3	Peregrine Falcon	Eurasian Hoopoe	Budgerigar
4	Bald Eagle	Eurasian Blackbird	Indian Peafowl
5	Common Raven	White Stork	Cockatiel
6	Common Ostrich	House Sparrow	Emperor Penguin
7	Budgerigar	Common Raven	Gray Parrot
8	Eurasian Blackbird	Golden Eagle	Andean Condor
9	House Sparrow	Common Cuckoo	Island Canary
10	Emu	Budgerigar	Rose-ringed Parakeet

	Species popularity index
1	Common Ostrich
2	Dodo
3	Budgerigar
4	Indian Peafowl
5	Cockatiel
6	Emperor Penguin
7	Gray Parrot
8	Andean Condor
9	Island Canary
10	Rose-ringed Parakeet
11	Bald Eagle
12	Emu
13	King Penguin
14	Zebra Finch
15	Shoebill
16	Helmeted Guineafowl
17	Harpy Eagle
18	Wild Turkey
19	Blue-and-yellow Macaw
20	Wandering Albatross
21	Rosy-faced Lovebird
22	Toco Toucan
23	Kakapo
24	Common Hill Myna
25	Black Swan
26	Golden Eagle
27	Muscovy Duck
28	Secretarybird
29	Canada Goose
30	Greater Rhea

FIGURE 6. Considerations for rankings of interest in species across multiple languages. Left: three different ranking measures highlight different aspects of interest. A ranking based on unadjusted raw sums (daily views all) has strong overlap with English, Wikipedia's largest language edition. Meanwhile a ranking that accounts for language edition size (Popularity index) highlights European species due to the euro-centric distribution of Wikipedia's language editions. Finally, a ranking that accounts for both language edition size and the geographic distribution of species and languages (Species Popularity Index) highlight species that appear in the pet trade. Right: the top 30 most popular bird species worldwide according to a Species Popularity Index.

Methods 2

As I discussed in Methods 1, interpreting big data can be challenging. This is especially true in situations where it is impossible to access to raw data, as is the case in many online platforms such as Google, and Facebook. This brief communication is a response to a paper that used Google Trends to make inferences about the extent of public interest in conservation (Burivalova et al. 2018). We point out that the authors failed to account for one of Google's methods for scaling search results across topics, and that this omission influences the results of their study. In the context of the thesis, I include this paper as an example of the challenges associated interpreting online data. Additionally, this paper provides further justification for my use Wikipedia, with its open access data policies, as a source throughout the thesis.

Submission status: Published. Correia, R.A., Di Minin, E., Jaric, I., Jepson, P., Ladle, R., Mittermeier, J.C., Roll, U., Soriano-Redondo, A., and D. Verissimo. 2019.* *Inferring public interest in conservation from search engine data requires caution.* Frontiers in Ecology and the Environment 17(5) 254-255. *all authors contributed equally to the paper.

Personal contribution: Joint lead author. The contribution is a multi-author piece in which all authors contributed equally. I helped with editing the manuscript as well as providing context based on my research.

Inferring public interest from search engine data requires caution

In a recent communication, Burivalova *et al.* (2018) analyzed Google Trends data with respect to interest in conservation and made two important claims: (1) that public interest in conservation is rising since 2004 and (2) that conservation and climate change–related topics have similar levels of public interest. Their assertions are based on a proposed new method to back-adjust Google Trends data from relative to absolute search volume. Their results contradict those of earlier studies using similar data to claim that public interest in conservation is waning (McCallum and Bury 2013; Troumbis 2017). However, after reproducing Burivalova *et al.*'s analysis and correcting an error in their algorithm, we find that their claims may not hold under scrutiny.

We applaud the effort by Burivalova and colleagues to develop new ways to explore Google Trends data, and we see great potential in their proposed method. However, their assertions rest on implicit assumptions that (1) an observed growth in absolute search volume reflects an increase in public interest and that (2) their method correctly reflects differences in public interest across topics. We reproduced Burivalova *et al.*'s Figure 3 using the same methodology (Figure 1a) and found that their metric was adjusted in relation to the maximum search volume (highest a_i) observed within each topic (Equation 8 and step 6 in their WebFigure 4). This is valid only when analyzing topics individually, because it does not preserve differences in search volume across topics. To preserve such differences using Equation 8, the data must be scaled using a value of a_i that reflects the maximum search volume observed across *all* topics. Doing so, we find that search volume for climate-change topics is approximately double that observed for conservation topics in recent years (apart from "Extinction"; Figure 1b).

Furthermore, the rapid growth of search engine use over time means that the absolute number of searches is likely to have increased for any topic, independently of public interest. Given this, a more reasonable assessment of how public interest for conservation topics has changed might be achieved by comparing the rate of change across topics (Nghiem *et al.* 2016). Calculating the ratio between searches for conservation topics and searches for the topic "climate change", we find results are not constant over time. The ratio decreases for all terms between 2004 and 2007, and again after 2014, indicating that searches for "climate change" took prominence in relation to conservation topics during these periods (Figure 1c). Repeating the analysis with "global warming" produces similar results, which concur with reports of an increase in news media attention toward climate-change topics during these periods (Legagneux *et al.* 2018). Overall, our results suggest that temporal changes in public interest toward conservation are different and more nuanced than those presented in Burivalova *et al.* (2018).

In our experience, temporal dynamics of search engine usage are complex, and inferences on changes in public interest derived from such data should be approached with caution. Accounting (or not) for multiple confounding factors such as the growth in internet access, search engine usage, time spent online, and the changing nature of internet usage (eg work versus leisure; see Ficetola [2013]) is likely to produce markedly different results. Furthermore, as Burivalova and colleagues rightly point out, Google

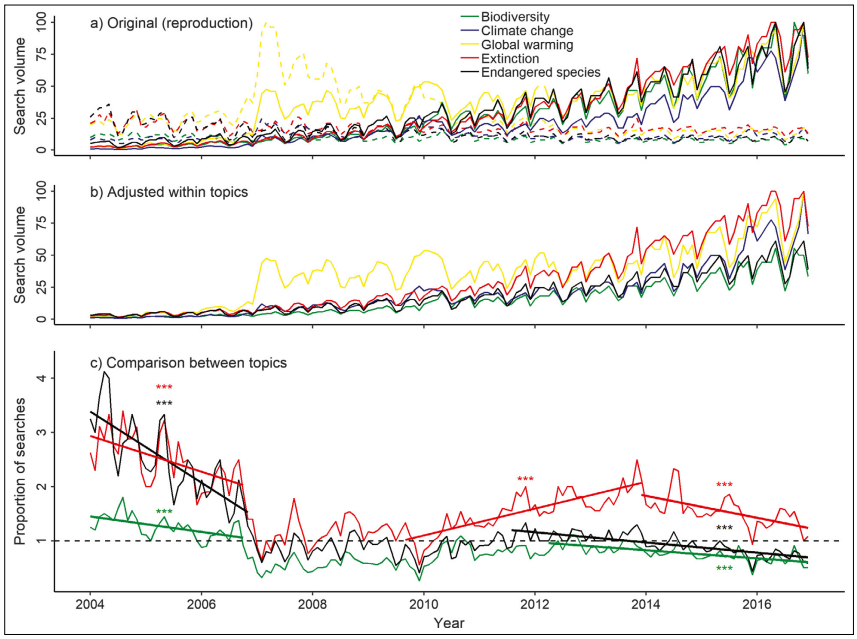


Figure 1. Changes in search volume for conservation and climate-change topics over time. (a) Reproduction of Figure 3 in Burivalova *et al.* (2018), where dashed lines represent original search volume obtained from Google Trends and solid lines represent adjusted search volume using the method proposed by Burivalova *et al.* (2018). (b) Search volume data adjusted according to our proposed method, which accounts for differences in search volume across topics. (c) The monthly proportion of absolute searches for conservation topics (color key same as in [a]) in relation to climate change, indicated by the thin solid lines, and underlying significant temporal trends, indicated by thick solid lines. The dashed line indicates a similar search volume for conservation topics and climate change; values above the dashed line indicate higher volume of searches for conservation topics and vice versa. Points of change in the temporal trends were identified using package "strucchange" and statistically non-zero trends (***, $P < 0.001$) were estimated using generalized linear models in R software v.3.4.2 (R Core Team 2017).

Trends data present an additional challenge in this context because the raw data are unavailable due to proprietary constraints. Moreover, considering only a single data source may produce a biased view of changes in public interest toward any topic. We believe that combining results from multiple sources (Veríssimo *et al.* 2014; Cooper *et al.* 2019; Jarić *et al.* 2019) – such as different search engines, social media platforms, online news media, Wikipedia page views, internet blogs and forums – is more likely to provide meaningful insights on social and cultural trends.

The exploration of conservation-related topics using digital data sources provides new opportunities for conservation science and practice (Sutherland *et al.* 2018). We are advocates of such potential (Di Minin *et al.* 2015; Ladle *et al.* 2016; Soriano-Redondo *et al.* 2017) and actively encourage efforts to positively engage with culturomics methods for the benefit of conservation. We have also faced challenges and limitations associated with these methods including, for example, issues of semantic complexity, language dynamics, and data collection and curation (Ladle *et al.* 2016), and have been careful to elucidate them. Our research has aimed to offer practical solutions to some of these challenges (Jarić *et al.* 2016; Correia *et al.* 2018; Roll *et al.* 2018), and we have recently established a Conservation Culturomics working group within the Society for Conservation Biology. This group embodies our belief in open collaboration, and we hope it will facilitate knowledge-sharing and collaborative efforts to overcome challenges through a welcoming, supportive, and stimulating environment. We encourage all interested parties to join this endeavor toward advancing digital methods for conservation.

Ricardo A Correia[1,2*†], Enrico Di Minin[3,4,5†], Ivan Jarić[6,7†], Paul Jepson[8†], Richard Ladle[2†], John Mittermeier[8†], Uri Roll[9†], Andrea Soriano-Redondo[10,11,12†], and Diogo Veríssimo[13,14,15†]

*[1]DBIO & CESAM-Centre for Environmental and Marine Studies, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal; [2]Institute of Biological and Health Sciences, Federal University of Alagoas, Maceió, Brazil *(rahc85@gmail.com); [3]Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland; [4]Helsinki Institute of Sustainability Science, University of Helsinki, Helsinki, Finland; [5]School of Life Sciences, University of KwaZulu-Natal, Durban, South Africa; [6]Biology Centre of the Czech Academy of Sciences, Institute of Hydrobiology, České Budějovice, Czech Republic; [7]University of South Bohemia, Faculty of Science, Department of Ecosystem Biology, České Budějovice, Czech Republic; [8]School of Geography and the Environment, University of Oxford, Oxford, UK; [9]Mitrani Department of Desert Ecology, The Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Midreshet Ben-Gurion, Israel; [10]Centre for Ecology and Conservation, College of Life and Environmental Sciences, University of Exeter, Cornwall Campus, UK; [11]CIBIO/InBio, Centro de Investigação em Biodiversidade e Recursos Genéticos, Laboratório Associado, Universidade do Porto, Portugal; [12]CIBIO/InBio, Centro de Investigação em Biodiversidade e Recursos Genéticos, Laboratório Associado, Instituto Superior de Agronomia, Universidade de Lisboa, Portugal; [13]Department of Zoology, University of Oxford, Oxford, UK; [14]Oxford Martin School, University of Oxford, Oxford, UK; [15]San Diego Zoo Institute for Conservation Research, Escondido, CA; [†]all authors contributed equally to this work*

Burivalova Z, Butler RA, and Wilcove DS. 2018. Analyzing Google search data to debunk myths about the public's interest in conservation. *Front Ecol Environ* **16**: 509–14.

Cooper MW, Di Minin E, Hausmann A, *et al.* 2019. Developing a global indicator for Aichi Target 1 by merging online data sources to measure biodiversity awareness and engagement. *Biol Conserv* **230**: 29–36.

Correia RA, Jarić I, Jepson P, *et al.* 2018. Nomenclature instability in species culturomic assessments: why synonyms matter. *Ecol Indic* **90**: 74–78.

Di Minin E, Tenkanen H, and Toivonen T. 2015. Prospects and challenges for social media data in conservation science. *Front Environ Sci* **3**: 63.

Ficetola GF. 2013. Is interest toward the environment really declining? The complexity of analysing trends using internet search data. *Biodivers Conserv* **22**: 2983–88.

Jarić I, Correia RA, Roberts DL, *et al.* 2019. On the overlap between scientific and societal taxonomic attentions – insights for conservation. *Sci Tot Environ* **648**: 772–78.

Jarić I, Courchamp F, Gessner J, *et al.* 2016. Data mining in conservation research using Latin and vernacular species names. *PeerJ* **4**: e2202.

Ladle RJ, Correia RA, Do Y, *et al.* 2016. Conservation culturomics. *Front Ecol Environ* **14**: 270–76.

Legagneux P, Casajus N, Cazelles K, *et al.* 2018. Our house is burning: discrepancy in climate change vs biodiversity coverage in the media as compared to scientific literature. *Front Ecol Evol* **5**: 175.

McCallum ML and Bury GW. 2013. Google search patterns suggest declining interest in the environment. *Biodivers Conserv* **22**: 1355–67.

Nghiem LTP, Papworth SK, Lim FKS, *et al.* 2016. Analysis of the capacity of Google Trends to measure interest in conservation topics and the role of online news. *PLoS ONE* **11**: e0152802.

R Core Team. 2017. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Roll U, Correia RA, and Berger-Tal O. 2018. Using machine learning to disentangle homonyms in large text corpora. *Conserv Biol* **32**: 716–24.

Soriano-Redondo A, Bearhop S, Lock L, *et al.* 2017. Internet-based monitoring of public perception of conservation. *Biol Conserv* **206**: 304–09.

Sutherland WJ, Butchart SHM, Connor B, *et al.* 2018. A 2018 horizon scan of emerging issues for global conservation and biological diversity. *Trends Ecol Evol* **33**: 47–58.

Troumbis AY. 2017. Declining Google Trends of public interest in biodiversity: semantics, statistics or traceability of changing priorities? *Biodivers Conserv* **26**: 1495–505.

Veríssimo D, MacMillan DC, Smith RJ, *et al.* 2014. Has climate change taken prominence over biodiversity conservation? *BioScience* **64**: 625–29.

Chapter 2: Patterns

In Chapter 2, I compare metrics of online interest with biological and ecological data from an array of biodiversity databases to investigate the drivers of public interest in biodiversity. This chapter contains three primary papers which complement each other by varying in their taxonomic scope, the biological datasets that they use, and the variables that they explore. Patterns 1 collects online data for ca. 10,000 species of reptiles and compares their popularity in Wikipedia to biological attributes in the Global Assessment of Reptile Distributions (GARD). Patterns 2 uses pageview data from all of the nearly 32,000 IUCN-listed species present in Wikipedia and examines the impact of temporal patterns, specifically seasonality, in influencing online interest. Finally, Patterns 3 explores how the abundance of a species in landscape impact the attention it receives online. It uses pageview data for over 10,000 bird species in Wikipedia and correlates them with information from the citizen science database eBird. The research focus and data source of each of these three papers is outlined in Table 2.

	Taxonomic scope	Number of taxa	Reference Dataset	Drivers investigated
1. Patterns 1	Reptile species	10,000	GARD	Body size, venomousness, IUCN status
2. Patterns 2	IUCN listed species	32,175	IUCN	Seasonality
3. Patterns 3	Bird species	10,000	eBird	Abundance

TABLE 2. Analytical papers in Chapter 2 with the taxonomic scope of each paper, the number of taxa, the biological reference dataset, and the drivers investigated in the study.

Patterns 1

This paper explores the role of biological traits in influencing patterns of online interest across species. Using a database of global reptile distributions (GARD), we compare how Wikipedia pageviews are distributed amongst reptile species and investigate how the distribution of these pageviews correlates with the phylogeny, threat status and biological attributes of species. In the case of reptiles, we find that being large, venomous, and endangered are all important in contributing to higher interest. We also find that some species consistently attract high levels of interest across a wide range of languages. These high interest species could be candidates for global flagship species (see Applications 2). Finally, we provide some insight into spatial variations in interest. This paper was the first of the thesis to be published, and the first to use Wikipedia pageviews for measuring interest in species. As such, it provides a proof-of-concept for other publications in the thesis.

Submission status: Published. Roll*, U., J.C. Mittermeier*, G.I. Diaz*, M. Novosolov, A. Feldman, Y. Itescu, S. Meiri and R. Grenyer. 2016. *Using Wikipedia page views to explore the cultural importance of global reptiles.* Biological Conservation 204(A): 42-50. *These authors contributed equally to this paper.

Personal contribution: Joint lead author. This paper included roughly equal contributions from myself, Uri Roll, and Gonzalo Diaz. I came up with the initial concept of the paper with Uri, downloaded and curated the data with Gonzalo, devised the methodology with Uri, contributed to the initial draft of the text with Uri, and contributed to editing the text with all the authors involved.

Biological Conservation 204 (2016) 42–50

Contents lists available at ScienceDirect

Biological Conservation

journal homepage: www.elsevier.com/locate/bioc

Using Wikipedia page views to explore the cultural importance of global reptiles

Uri Roll [a,b,*,1], John C. Mittermeier [b,1], Gonzalo I. Diaz [c,1], Maria Novosolov [d], Anat Feldman [d], Yuval Itescu [d], Shai Meiri [d], Richard Grenyer [b]

[a] Department of Zoology, University of Oxford, Tinbergen Building, South Parks Road, Oxford OX1 3PS, UK
[b] School of Geography and the Environment, University of Oxford, South Parks Road, Oxford OX1 3QY, UK
[c] Department of Computer Science, University of Oxford, Wolfson Building, Parks Road, Oxford OX1 3QD, UK
[d] Department of Zoology, Tel-Aviv University, Tel-Aviv 6997801, Israel

ARTICLE INFO

Article history:
Received 30 November 2015
Received in revised form 26 February 2016
Accepted 30 March 2016
Available online 4 May 2016

Keywords:
Big data
Conservation
Culture
Endangered
Language
Flagship species

ABSTRACT

Modern conservation operates at the nexus of biological and social influences. While the importance of social and cultural factors is often mentioned, defining, measuring and comparing these factors remains a significant challenge. Here, we explore a novel method to quantify cultural interest in all extant reptile species using Wikipedia — a large, open-access online encyclopaedia. We analysed all page views of reptile species viewed during 2014 in all of Wikipedia's language editions. We compared species' page view numbers across languages and in relationship to their spatial distribution, phylogeny, threat status and various other biological attributes. We found that the three species with most page views are shared across major language editions, beyond these, page view ranks of species tend to be specific to particular language editions. Interest within a language is mostly focused on reptiles found in the regions where the language is spoken. Overall, interest is greater for reptiles that are venomous, endangered, widely distributed, larger and that have been described earlier. However, within individual reptile families not all the above factors predict page views. Most families contain at least one species in the top 5% of page views, but 29 families (with 1,450 species) have no 'high interest species' in them. Overall, our analyses elucidate novel patterns of human interests in nature over large geographical, cultural and taxonomic spectra using big-data techniques. Such approaches hold much promise for incorporating social perceptions in future conservation practices.

1. Introduction

Various cultural elements exert a powerful influence on how conservation attention and resources are allocated. Challenges facing species conservation projects may even be primarily social rather than biological (Kellert, 1985; Tisdell, 2014). Thus, in order to secure better outcomes for conservation management schemes – in addition to biological attributes – social and cultural variables should also be incorporated in decision making (Ladle and Jepson, 2008). Nevertheless, these attributes are often neglected in the conservation decision-making process (Gunnthorsdottir, 2001; Kellert, 1985; Stokes, 2007).

Most global and regional conservation prioritization schemes rely on quantifiable differences in the geographic distribution, population size, ecological role, and evolutionary distinctness of species (Faith, 1992; IUCN, 2014; Mills et al., 1993; Vane-Wright et al., 1991). However, species are also unequal in their contributions to human culture — in how they are perceived by, and attract attention from, humans. While a few authors have addressed this point (Cristancho and Vining, 2004; Garibaldi and Turner, 2004), the extent to which species vary in their cultural importance or impact remains very poorly studied and how these potentially affect conservation practices is mostly unknown. Nevertheless, in order for conservation actions to be fruitful they need to incorporate both traditional conservation parameters and cultural values in local to global scales of the different actors and interventions attempted.

As with other human practices, conservation may suffer from biases due to the non-randomness in human interests and affections. For example we are more interested in the well-being and prolonged persistence of big, 'fluffy', attractive animals (Gunnthorsdottir, 2001; Johnson et al., 2010; Ward et al., 1998), those with large, forward facing eyes (Macdonald et al., 2015), those who are more brightly coloured (Prokop and Fančovičová, 2013; Stokes, 2007) and preferably more phylogenetically (and thus morphologically) close to us (Gunnthorsdottir, 2001).

Reptiles as a group are usually less in the public eye when compared to the other groups of tetrapods, due to several potential biases and knowledge deficiencies. This may have great ramifications for their prolonged conservation. Reptiles comprise about 30% of all extant land vertebrate species (Meiri and Chapple, 2016, this issue), and are likely

* Corresponding author at: Department of Zoology, Tinbergen Building, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.
E-mail address: uri.roll@zoo.ox.ac.uk (U. Roll).
[1] Contributed equally to the paper.

http://dx.doi.org/10.1016/j.biocon.2016.03.037
0006-3207/

to have an even greater representation amongst threatened species (IUCN, 2014). Nevertheless, their representation in targeted species conservation schemes is usually much lower (Clucas et al., 2008). Here we list reptiles' representation in targeted species programs of a few global conservation NGOs, acknowledging that local conservation schemes may have different representations of reptiles. Of the World Wildlife Fund's 36 priority species or species groups, only sea turtle and 'Asian tortoises and freshwater turtles' are reptilian (http://wwf.panda.org). Of the 1,031 projects supported by the Mohamed bin Zayed Species Conservation Fund which incorporate tetrapods, only 17% include reptiles (http://www.speciesconservation.org). None of the African Wildlife Foundation's projects target reptiles (http://www.awf.org). Reptiles comprise 16% of the specific species of interest listed by the Defenders of Wildlife organization, but only 6.5% of their species up for adoption (http://www.defenders.org). While 13 of the 36 species (36%) under management by the Durrell Wildlife Conservation Trust are reptiles, only one of the 14 species (7%) up for adoption on their website is a reptile (http://www.durrell.org). Furthermore, as compared with mammals and birds, the scientific knowledge of basic biological attributes of reptiles is much lower, and thus so is our ability to develop sound conservation practices addressing their prolonged survival (Böhm et al., 2013; Meiri and Chapple, 2016, this issue). For example, while the distributions of all other groups of tetrapods has been known for a decade now (Grenyer et al., 2006; Orme et al., 2005), only recently has a parallel effort been completed for reptiles (http://www.gardinitiative.org).

Within the ~10,300 recognized species of reptiles (Uetz and Hošek, 2015) there are great differences between species in the cultural representations (i.e. appearance at all in the public sphere) and importance in various cultural roles they play. Some reptile species (e.g., venomous snakes, geckos, tortoises, crocodiles) have potent roles across an array of cultural mediums — in the pet trade, as food objects, as fictional characters, as objects of fear or aspiration, etc. (Alves et al., 2009; Alves et al., 2008; Campbell, 2009; Klemens and Thorbjarnarson, 1995). Nevertheless most species remain unknown beyond a few herpetology specialists. As such, there are potentially great differences in the contributions of individual reptile species to the various domains of human culture. If conservation hopes to preserve features such as the 'aesthetic, historical, and recreational values' of species (Millennium Ecosystems Assessment, 2005), then identifying which species contribute to those values is of fundamental importance. Previous studies have examined cultural attitudes towards particular reptile species within local contexts (Ceríaco, 2012; Ceríaco et al., 2011; Deb and Malhotra, 2001; Jones et al., 2008; Ramstad et al., 2007), yet there have been no global efforts to compare the cultural significance of reptiles. Since many conservation policies and frameworks operate globally, considering cultural value at a global scale is potentially very useful.

'Culture' is one of the most widely used terms in the English language (Taras et al., 2009). In the context of conservation, 'cultural value' is frequently applied to defining ways in which humans assign value to different species. Though useful in the abstract, it creates challenges in measuring exactly what it means and creates confusion through the various meanings of value. Here we explore page view statistics (elaborated below) extracted from the Wikipedia online digital text archive for all extant reptiles in all language editions as a measure of the prominence of an entity or idea within a given cultural context (Yu et al., 2015).

Digital text archives are an increasingly significant resource for the study of human culture and enable questions and scales of investigation that were unfeasible until recently (Aiden and Michel, 2013; Lazer et al., 2009; Schich et al., 2014). The use of these resources for studying cultural patterns relevant to conservation is beginning to be recognized but remains low (Arts et al., 2015; Correia et al., 2016). The cultural salience of reptile species could theoretically be studied in a variety of digital archives. Within this context, Wikipedia is particularly appealing for several reasons: 1) it is huge (>35 million articles in English to date); 2) it is multilingual (287 languages including 12 with >1 million articles); 3) it is open access and free to download; 4) it follows a standardized structure that groups information on a species together and thus avoids many of the challenges of unstructured text databases; and 5) a growing body of academic literature addresses aspects of Wikipedia's coverage (Giles, 2005; Halavais and Lackaff, 2008; Messner and DiStaso, 2013; Samoilenko and Yasseri, 2014), credibility (Brown, 2011; Miller and Murray, 2010; Wilson, 2014), contributor demographics (Wilson, 2014) and user dynamics (Yasseri et al., 2012; Yasseri et al., 2014).

Wikipedia also has important limitations in the results it can produce and biases in whose cultural information it reflects. Unsurprisingly, Wikipedia skews heavily towards the Global North with respect to both content generation and usage, and New Guinean, as well as, African languages are poorly represented (Graham et al., 2014). Wikipedia contributors also tend to be a highly skewed demographic from within the Global North: English-language Wikipedia contributors, for example, are primarily male, and mostly under 29 years old (Wilson, 2014). As of 2013, 4.3 million registered users made at least one edit to all of Wikipedia, but only about 130,000 registered users made more than 100 edits (Wilson, 2014). Another significant challenge in analysing Wikipedia from a cultural standpoint is that some of its contributors are not human. A proportion of Wikipedia articles are created or edited by specialized programs called 'bots'. As an example, one of the most active bots, called 'Lsjbot', has contributed various types of information to over 2.7 million articles. Results obtained from Wikipedia therefore need to be considered within this context. We therefore want to emphasize that Wikipedia should not be seen as reflecting universal values nor representing the voices of groups such as indigenous people or individuals with limited internet access.

Wikipedia provides several potential referential metrics of cultural interest or saliency of different objects, each with potential benefits and flaws. Each Wikipedia page has been created at a particular date, been edited several times by a different number of editors, has a particular length, is linked to and from other pages (within and outside Wikipedia), appears in a set of different language editions, has been viewed a particular number of times, etc. Some of these metrics are potentially very information rich. Unfortunately, many of these metrics may suffer from inherent biases due to bot activity. Therefore, for our initial exploration of these data as a source for cultural attitudes towards nature, we limited our scope of reference only to the number of page views in different language editions of Wikipedia reptile pages. We suggest that page views within a given language measure the general interest that a page attracts from the public speaking that language (with the above biases in mind). We acknowledge that page views are recorded in a way that cannot account for page queries made by bots. Nevertheless, as most page views are made by humans (http://stats.wikimedia.org/archive/squid_reports/2014-12/SquidReportCrawlers.htm) we posit that they can provide some insight as to which reptiles attract more interest in the public sphere globally.

Here, we provide a novel approach to quantify and compare one aspect of the cultural interest associated with global reptile species: the number of times individual reptile pages are viewed, in a large, user-generated, multi-lingual, online encyclopaedia. We explore patterns at the species level, as many consider species the fundamental unit of biodiversity (Wilson, 1992) and many conservation actions are designated towards individual species (Brooks, 2010). This enables us to explore i) those species that may have greater conservation value because of their higher cultural interest, and ii) cross-cultural differences in interests towards reptile species, a key attribute in unravelling many conservation challenges. We address three questions relevant to the investigations of culture and conservation: 1) which reptile species attract most cultural interest globally, 2) what biological traits characterize those species, and 3) how does the relative cultural interest in species vary across languages.

2. Materials and methods

We obtained cultural data on reptile species from two related sources: (i) DBpedia (http://wiki.dbpedia.org, version "Dataset 2014"), a repository of structured data, extracted and curated from Wikipedia, and (ii) Wikidata (http://www.wikidata.org, version 2015–07), a publicly editable repository of structured data, which aims to gather structured data from diverse sources including DBpedia, the Integrated Taxonomic Information System (ITIS - http://www.itis.gov), and many others. For both Wikidata and DBpedia, the full datasets were downloaded. For data processing scripts see the supplementary information.

To extract species-level entities within Wikidata, we used the fact that the global taxonomy of life via ITIS is fully integrated into this database. We therefore queried Wikidata for all entities marked as (i) having a 'taxon rank' property (https://www.wikidata.org/wiki/Property:P105) set to the value species (https://www.wikidata.org/wiki/Q7432), or (ii) having the property 'taxon name' (https://www.wikidata.org/wiki/Property:P225) set to some value (as opposed to no value). Our definition of a species was therefore anything with either a binomial or a 'species' label. Each species in the resulting list 'Wikidata all species' corresponded to a unique URL within the Wikidata database. We identified reptiles in this list by matching them to Uetz and Hošek (2015) which served as the backbone taxonomy for this work. To obtain information on language editions and page views across languages, we cross-referenced our 'Wikidata reptiles' with DBpedia (data currently not found in Wikidata). DBpedia only includes a language edition for a species if a page for that species exists in a given language. The resulting list 'Wikipedia reptile URLs' contained every page title, in any language, for a species in Wikidata reptiles. We limited our analyses only to those pages that have been viewed at least once as those that have not been viewed at all are most likely bot-generated pages.

Wikipedia page views and article traffic statistics are stored and made publically available at https://wikitech.wikimedia.org/wiki/Analytics/Data/Pagecounts-raw (a third party visualization tool which can be found at http://stats.grok.se). This dataset consists of files collated on an hourly basis for page views to all Wikipedia articles across all language editions. To extract page views for reptiles, we downloaded page view files for the calendar year 2014 (collected per hour), and then matched page titles and their corresponding view counts to Wikipedia reptile URLs. Hourly view counts for each language edition of a species were summed to count total views per species. Altogether we identified 10,002 reptile species in 'Wikidata all species' that were viewed in 2014.

In order to examine patterns of page view activity across reptiles, we assembled various traits per species. Year of description was obtained from Uetz and Hošek (2015). Range sizes of the species as well as global gridded distribution maps on a 1° Behrmann equal area projection were obtained from the GARD initiative (http://www.gardinitiative.org), as was data on the presence of venom. Threat status, for assessed species, was obtained from the IUCN red list (http://www.iucnredlist.org). Body-size measurements for lepidosaurs were taken from Feldman et al. (2016), for crocodiles and turtles these were collected by one of the authors (YI). Any species assessed as VU or above by the IUCN was recorded as threatened. Any known venomousness of a species was recorded as 'yes' in a binomial venomousness predictor. Species that were not known to be venomous or threatened with extinction were recorded as non-venomous and non-threatened, respectively. All the variables were used as a predictor set for a model of page views across all language editions, and separately for the English language edition. Subset models of the total page views for several taxonomic groups were also explored.

We modelled page views using a negative binomial GLM, with the theta parameter estimated from the data by maximum likelihood (a starting value from a Poisson error model showed problematic over-

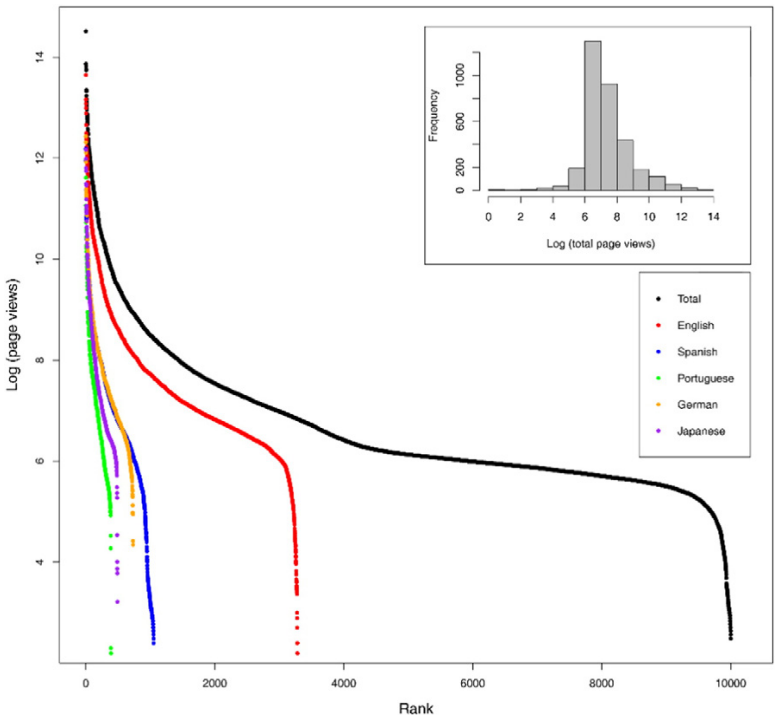


Fig. 1. The abundance and frequency distributions of page views. The main pane shows the ranked abundance distribution for ln transformed total page views and views of five main families. The inset shows the frequency distribution of log transformed total page views.

Table 1
Species rank for total page views across all languages and for five key language editions. The top 20 species for all of Wikipedia and five key language editions are shown and ordered from top to bottom, with page views given in parenthesis. Species appearing in more than one language edition are colour-coded according to the column on the right.

Total	English	Spanish	Portuguese	German	Japanese	Binomial	# top
Varanus komodoensis (2014932)	Varanus komodoensis (845265)	Iguana iguana (196312)	Varanus komodoensis (110791)	Natrix natrix (251174)	Gloydius blomhoffii (191748)	Crocodylus porosus	5
Vipera berus (1059665)	Dendroaspis polylepis (520406)	Varanus komodoensis (155033)	Chelonoidis carbonaria (64113)	Vipera berus (223389)	Protobothrops flavoviridis (160635)	Dendroaspis polylepis	5
Crocodylus porosus (1055428)	Crocodylus porosus (478207)	Boa constrictor (97573)	Caretta caretta (59071)	Anguis fragilis (181409)	Elaphe climacophora (134282)	Varanus komodoensis	5
Dendroaspis polylepis (1042072)	Ophiophagus hannah (439853)	Eunectes murinus (86817)	Boa constrictor (45784)	Eunectes murinus (151228)	Gekko japonicus (127422)	Chelonoidis nigra	4
Ophiophagus hannah (1008676)	Heloderma suspectum (396522)	Crocodylus porosus (60603)	Caiman latirostris (39123)	Varanus komodoensis (130003)	Rhabdophis tigrinus (126388)	Eunectes murinus	4
Natrix natrix (949592)	Agkistrodon piscivorus (315207)	Dermochelys coriacea (57041)	Dendroaspis polylepis (37993)	Zamenis longissimus (88508)	Takydromus tachydromoides (97362)	Boa constrictor	3
Eunectes murinus (929057)	Alligator mississippiensis (266741)	Eretmochelys imbricata (52099)	Crocodylus porosus (34337)	Lacerta agilis (84408)	Varanus komodoensis (97124)	Crocodylus niloticus	3
Boa constrictor (629112)	Dermochelys coriacea (254229)	Dendroaspis polylepis (51405)	Bothrops jararaca (28880)	Testudo hermanni (80563)	Pelodiscus sinensis (94847)	Dermochelys coriacea	3
Anguis fragilis (616326)	Crocodylus niloticus (240528)	Caiman yacare (51215)	Bothrops alternatus (27771)	Dendroaspis polylepis (79740)	Trachemys scripta (72495)	Ophiophagus hannah	3
Crocodylus niloticus (613623)	Boa constrictor (240469)	Caretta caretta (49466)	Python regius (27680)	Crocodylus porosus (76318)	Elaphe quadrivirgata (63705)	Caiman yacare	2
Dermochelys coriacea (559746)	Eunectes murinus (233751)	Crocodylus acutus (47570)	Lachesis muta (27530)	Ophiophagus hannah (62606)	Ophiophagus hannah (61424)	Caretta caretta	2
Heloderma suspectum (521818)	Agkistrodon contortrix (225881)	Chelonia mydas (47417)	Bothrops insularis (25673)	Oxyuranus microlepidotus (62434)	Plestiodon japonicus (57898)	Chelonia mydas	2
Iguana iguana (498330)	Macrochelys temminckii (204320)	Chelonoidis carbonaria (41247)	Caiman yacare (23431)	Vipera aspis (55498)	Chelydra serpentina (55670)	Chelonoidis carbonaria	2
Caretta caretta (476772)	Crocodylus acutus (201540)	Caiman crocodilus (33843)	Spilotes pullatus (23058)	Chelonoidis nigra (55451)	Crocodylus porosus (55111)	Chelydra serpentina	2
Chelonoidis nigra (471396)	Chelonoidis nigra (200239)	Bothrops asper (32822)	Hemidactylus mabouia (21975)	Python molurus (52781)	Mauremys reevesii (47096)	Crocodylus acutus	2
Chelonia mydas (458579)	Gavialis gangeticus (199622)	Bothrops atrox (32577)	Bothrops jararacussu (20316)	Coronella austriaca (46730)	Dendroaspis polylepis (41840)	Eretmochelys imbricata	2
Malayopython reticulatus (432497)	Python bivittatus (198632)	Chelonoidis nigra (30041)	Dermochelys coriacea (19032)	Python regius (40647)	Macrochelys temminckii (40402)	Macrochelys temminckii	2
Alligator mississippiensis (425631)	Chelydra serpentina (190934)	Tarentola mauritanica (28872)	Melanosuchus niger (19027)	Crocodylus niloticus (40644)	Mauremys japonica (37894)	Oxyuranus microlepidotus	2
Gavialis gangeticus (393183)	Chelonia mydas (182412)	Vipera aspis (27377)	Eretmochelys imbricata (18853)	Oxyuranus scutellatus (39910)	Eunectes murinus (35703)	Python regius	2
Agkistrodon piscivorus (391239)	Oxyuranus microlepidotus (173869)	Crocodylus niloticus (27351)	Chelonoidis nigra (18212)	Emys orbicularis (37246)	Malayopython reticulatus (34727)	Vipera aspis	2

dispersion). Continuous variables were paired with a quadratic term. We restricted our analyses to those species with complete cases — i.e. without missing values in any of the data columns (for sample sizes see Table 2). Analyses were conducted in R (R-Core-Team, 2015) using the glm.nb function in the MASS library (Venables and Ripley, 2002). Model averaging was carried out using the MuMIn library (Barton, 2015) by all-subsets searches of the complete model (models with only the quadratic term for continuous variables, and not the main term, were excluded). We restricted our analysis to those models within the top 4 AIC units of the best model (Burnham and Anderson, 2002). We present coefficients, significance probabilities, and variable relative importance from the AICc weighted average model assuming a coefficient of zero for variables with no evidence weight in individual models (the "full" coefficient averages in MuMIn).

Initially we plotted the median value of the total page views for all the species in each 1° grid-cell. We then explored the global distribution patterns of page views in five large Wikipedia language editions which are not known to have extensive bot edit histories, and are dominant in the countries where they are spoken (Graham et al., 2014) — English, Spanish, Portuguese, German and Japanese. For each language we calculated the total number of page views for each species. We then assigned to each grid cell all the page-views of the species that reside in it and divided this value by the total number of species in that cell with Wikipedia pages in that language. This gave us a measure of the relative visibility in Wikipedia, for each cell, correcting for global trends in species richness. For each reptile family we noted whether it included species found in the top 5 percentile of page views. We then indicated on a phylogenetic tree of reptile families based on Reeder et al. (2015) and Pyron and Burbrink (2014) those that do and do not have such 'high interest species'.

3. Results

Extracting page views for the year 2014 resulted in 67,062 pages of Wikipedia reptile URLs with at least a single view (138 pages or 0.2% had only a single view); reptile pages were viewed a total of 55.5 million times in 2014. There were 146 different language editions of Wikipedia with reptile pages in them. Median total views per species is 828.3, and mean value is 5,553.3 giving a very skewed distribution of page views

Table 2
Modelling page views with various traits. The results of modelling page views with negative binomial GLMs and quadratic terms for continuous variables. Models are for all page views for all species and English page view for all species. Models for snakes, lizards (includes *Sphenodon*) and Scincidae are for total page views. Results are for the global models of these groups which includes all terms (see text). coeff. denotes coefficients. Asterisks denote p values — ** <0.01, *** <0.001.

	All reptiles		All reptiles (English)		Snakes		Lizards		Scincidae	
	coeff.	P	coeff.	P	coeff.	P	coeff.	P	coeff.	P
Venomousness	0.346	***	0.397	***	0.667	***	1.685	***	n/a	
Threat	0.733	***	0.288	***	0.643	***	0.739	***	0.340	***
Body mass	−0.145	***	−0.092	***	−0.501	***	−0.072	***	0.061	
Body mass^2	0.052	***	0.036	***	0.088	***	0.051	***	0.021	**
Description year	−0.214	***	−0.116	***	−0.250	***	−0.205	***	−0.177	***
Description year^2	5.3E-05	***	2.9E-05	***	6.3E-05	***	5.1E-05	***	4.4E-05	***
Area	−0.117	***	−0.170	***	−0.111	***	−0.152	***	−0.105	***
Area^2	0.009	***	0.013	***	0.009	***	0.012	***	0.006	***
n	9701		3115		3353		5932		1557	
Adjusted D^2	0.671		0.579		0.670		0.623		0.470	

with respect to the species of reptile in question (Fig. 1). Eighty-two (0.8%) species received over 50% of total views, and the top five species received 11.1% of all the views. The English version has many more page views than the other language editions and comprises 39.4% of all reptile page views. However, while in English there are about 1,850 species with over 1,000 page views, there are 3,150 species that receive over 1,000 page views when all languages editions' page views are combined (Fig. 1). Furthermore, 67% of species with page views in other languages do not even have a Wikipedia page in English. For total page views, and to lesser degree also for English and Spanish, there is a set of several hundreds of species (at the tail end of the distribution) that receive very few views in Wikipedia. Table S1 in the supplement gives the total page view values for all species and for the five main language editions explored.

Table 1 displays the species with the most page views for all of Wikipedia combined and for the five chosen Wikipedia language editions. Only three species of reptiles are found in the top 20 page views for all the five languages, *Varanus komodoensis* — Komodo dragon (top species in overall page views), *Crocodylus porosus* — salt-water crocodile (third overall) and *Dendroaspis polylepis* — the black mamba (fourth overall). These three species are also the three most visible pages in the English version of Wikipedia. Two more species: *Eunectes murinus* — the green anaconda (7th overall) and *Chelonoidis nigra* — the Galapagos tortoise (15th overall) are found in the top 20 of four of the five languages. *Vipera berus* — the common European adder, while being second in total page views is only found in the top 20 of page views of the German edition of Wikipedia (out of these five languages). Of the 63 species found in the top 20 of these five language editions, only 20 species are shared between more than one language and the rest are unique to a single language.

Our modelling procedure for all reptiles combined, and for reptile groups that have more than 1,500 species, highlighted a single model — the full model (with all the parameters included) as having all of the information (over 99% of the AIC weights). Thus for these groups we report only the results of this model (Table 2). For various reptilian sub-groups/sub-clades between 2 and 14 models contained most of the information Table 3 (for details on the contributing models to each groups' average see Table S2 in the Supplementary Information). Our modelling procedure was able to account for around 60% of the deviance in page views for all reptile page views in Wikipedia as well as just for the English version of Wikipedia (Table 2). Models for turtles and lizard families had around 10% less explanatory power (Tables 2 and 3). None of the chosen predictors explained important variation in page views of Amphisbaenia. For the analyses of all reptiles, as well as for all lizards, all snakes, and all reptiles in the English version, all the terms we tested in our model proved significant (Table 2). For other subsets, we see that different predictors are highlighted as significant and important (Tables 2, and 3). The year of the description of the species is an important predictor for all groups, with earlier described species being more visible. Threatened species attract more page views for many groups. Beyond these being venomous is important globally. Body mass is an important positive predictor globally, and it is also important for skinks, agamids, chameleons, colubrids and elapids. Geographic range size of the species is positively related to page view numbers for geckos, agamids, colubrids and vipers. It is important to note that the positive relationship between range size and page views is in the opposite direction to the relationship between threat status and page views, suggesting that the threat status relationship is not driven by the small range size of threatened species.

Overall the species of interest to Wikipedia users are found predominantly in North America, Europe and Japan (Fig. 2A). However, for individual language editions different patterns arise (Fig. 2B–F). English language Wikipedia users predominantly view reptiles living in North America, northern Europe as well as Indonesia and Eastern Africa (Fig. 2B). The Spanish edition's page views highlights species in South America, southern Europe and Southeast Asia (Fig. 2C). Portuguese

Table 3
Modelling total page views of reptile groups and key families with various traits. Results are for the top models in each group within 4 AIC units. coeff. are the averaged coefficients for each model and each term imp. are the importance values of the terms from the averaged modelled. Asterisks denote p values — * < 0.05, ** < 0.01, *** < 0.001.

	Turtles			*Amphisbaenia*			*Gekkonidae*			*Agamidae*			*Chamaeleonidae*			*Colubridae*			*Viperidae*			*Elapidae*		
	coeff.	P	imp.	coeff.	P	imp.	coeff.	P	imp.	coeff.	P	imp.	coeff.	P	imp.	coeff.	P	imp.	coeff.	P	imp.	coeff.	P	imp.
Venomous	n/a			n/a			n/a			n/a			n/a			3.6E-06		0.91	n/a			n/a		
Threat	5.8E-06	***	1.00	−1.0E-05		0.37	2.5E-05	***	1.00	1.7E-06		0.47	2.1E-07		0.21	3.8E-07	***	0.26	4.1E-06	***	1.00	8.7E-07		0.68
Body mass	1.3E-06		0.81	4.6E-05		0.62	6.2E-06		1.00	−2.9E-05	*	1.00	−8.0E-05	***	1.00	−1.0E-05		1.00	−2.9E-06		1.00	−1.3E-05	**	1.00
Body mass^2	5.9E-07		0.19	−6.8E-05		0.68	9.7E-06		0.63	5.7E-05	***	1.00	9.3E-05	***	1.00	3.0E-05	***	1.00	7.3E-06		0.73	2.6E-05	***	1.00
Description year	−2.4E-04	***	1.00	−1.7E-03		1.00	−1.1E-03	***	1.00	−8.7E-04	***	1.00	−6.7E-04	**	1.00	−1.0E-03		1.00	−2.9E-05		1.00	−7.0E-05		1.00
Description year^2	2.2E-04	**	1.00	1.6E-03		0.70	1.1E-03	***	1.00	8.5E-04	***	1.00	6.4E-04	**	1.00	1.0E-03	***	1.00	1.7E-05		0.37	6.3E-05		0.78
Area	−1.4E-05		0.81	4.7E-06		0.20	−2.9E-05	*	1.00	−2.9E-05	*	1.00	−2.3E-07		0.20	−9.6E-06		1.00	−1.3E-05	***	1.00	−5.3E-06		0.77
Area^2	1.6E-05		0.73	−3.2E-06		0.04	5.0E-05	***	1.00	5.5E-05	***	1.00	n/a			1.8E-05		0.87	1.7E-05	***	1.00	5.0E-06		0.64
n	206			187			1004			436			197			812			322			282		
Num. top models	6			14			2			2			3			4			4			8		
Adjusted D^2	0.532			0.293			0.528			0.542			0.584			0.633			0.576			0.735		

Wikipedia users view species residing in South and Central America, Sub-Saharan Africa and Southeast Asia (Fig. 2D). German Wikipedia users mostly view north Palearctic lizards (Fig. 2E). Japanese language Wikipedia highlights reptiles from east and Southeast Asia as well as southeast North America, several other regions in eastern South America, the Nile Valley, eastern India and western Southern Africa (Fig. 2F).

Several patterns arise when looking at the phylogeny highlighting families without representatives in the top five percentile of page views (Fig. 3). Twenty-nine of the 88 families do not have a single species in the top fifth percentile. Furthermore, several of the unrepresented families — such as Liolaemidae, Gymnophthalmidae and Sphaerodactylidae are speciose (with 286, 243, and 208 species respectively). Altogether about 1,450 species are found in sections of the tree without representation. All the families of crocodiles and turtles have at least one highly viewed species in them. However the tuatara (*Sphenodon punctatus*) is not found in the top five percentile of species' page views. There are three other small clades without any representatives in in the top five percentile. Nevertheless, we note that if we were to choose 5% of reptile species at random from our sample that would leave on average 33.8 families unsampled (standard deviation = 2.9, 10,000 randomizations).

4. Discussion

In recent years there has been a growing interest in incorporating people's attitudes towards nature while setting conservation priorities. In most cases, individual surveys were used to gain insight into people's perceptions, preferences and choices about nature (Macdonald et al., 2015; Taras et al., 2009). This approach is labour intensive, and usually limits the scope of the study. Here we use, for the first time, an online repository of user-generated content to gain insight into people's interests about an entire class – reptiles – over the entire globe and across many languages. We find interest is greater for reptiles that are venomous, endangered, widely distributed, large, and that were described earlier. Furthermore, we show clearly that page views within a language edition increases for species found where that language is spoken. This approach holds much promise for the future in elucidating general trends in people's attitudes towards nature and conservation.

The first thing we were able to highlight was those species ranked top overall and top in the different languages (Table 1). It seems that, unsurprisingly, large, venomous and potentially dangerous animals dominate the top spots: big fierce animals may be rare, but at least in reptiles they also receive disproportionally high internet interest (Fig. 1). These are led by the Komodo dragon which alone attracts 3.6% of total page views, followed by salt-water crocodile and the black mamba. The potential—however overstated—for fatal interaction with people, and the associated folklore and cultural salience, is clearly a large determinant of page view activity. The same could also be true for the green anaconda (a top 20 species in 4 languages). However, this narrative is clearly not true for the Galapagos giant tortoise which shares prominence with the anaconda. Beyond the shared superstars, language-specific priorities emerge, however they are still driven strongly by venom and the potential for harm. Thirty-five of the 63 top ranked species in the five languages we highlight could potentially be fatal. Most top ranked species are also of larger body size than the average reptile. Of the 61 top ranked species in the five languages (with body mass data), 42 (68.8%) are found in the top 5% of body

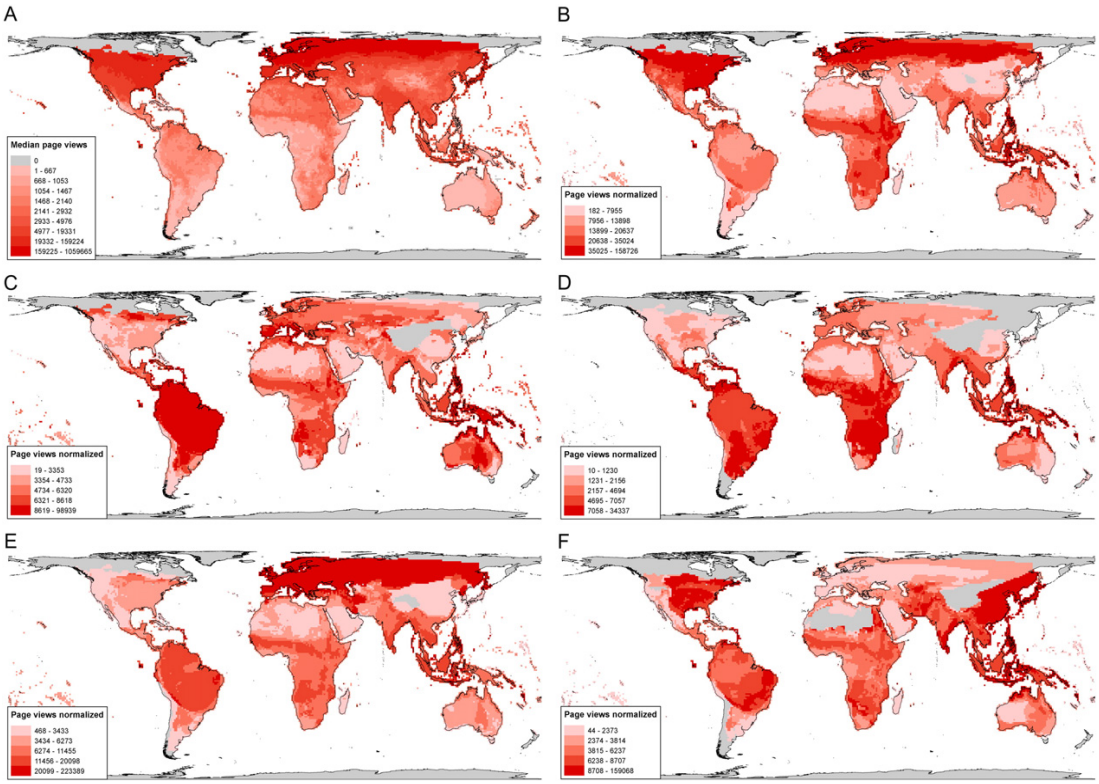


Fig. 2. Global distribution maps of page views of reptiles. Pane A displays the median value of the total page views for all the species, calculated per grid cell. Panes B–F show patterns of page views in five main Wikipedia language — English (B), Spanish (C), Portuguese (D), German (E) and Japanese (F). Each of these panes shows total number of page views per grid cell in that language divided by the number of species in that cell with Wikipedia pages in that particular language.

sizes of all reptiles. Nevertheless, as 82 top ranked species are generating more than half the total page views in our dataset, the vast majority of reptiles which are small, harmless and do not have a conservation narrative remain alarmingly invisible.

Greater interest or visibility in an online encyclopaedia is nevertheless no guarantee for greater support for targeted conservation of these species — especially for dangerous or venomous animals. However, greater interest could be harnessed to attract support by using both Wikipedia, other online tools and more traditional sources of information and campaigns to acknowledge the ecological roles and importance of these interesting, yet potentially dangerous species. Thus greater cultural interest could turn into greater cultural value. However, this will require a shift in our ideas of what makes a good flagship species — not just large, fluffy, big-eyed animals but potentially also mambas, cobras, and dragons.

The overall 'heat-map' of reptile page views for all Wikipedia language editions combined resembles that of content generation in Wikipedia in general (Graham et al., 2014). We find that more affluent societies with better internet connectivity, which are more represented in Wikipedia activity, are interested predominantly in the species that reside in their surrounding regions (Fig. 2A). While the classical latitudinal gradients of species richness highlights tropical richness (Rosenzweig, 1995), species of Wikipedia interest mostly inhabit higher latitudes. We present perhaps the first global map of the frequency of interactions with nature of a significant number of people. It is interesting to see its similarity with the general pattern of the Wallacean shortfall in other groups (Whittaker et al., 2005) and perhaps to speculate that they share a mechanism.

Apart from the absolute top 3–5 species, different language editions of Wikipedia highlight different top species of great interest. The local imprint of language becomes very evident when exploring the distributions of the species that receive most page views in the different language editions of Wikipedia (Fig. 2B–F). By combining the knowledge of where a language is spoken with the unique page views of different language editions of Wikipedia, our approach enabled us to show that the reptile species people are most interested in, are those which are found where they live. While this phenomenon has been shown before at local scales, for few species (Campos et al., 2012; Lindemann-Matthies, 2005; Shwartz et al., 2013), here we show it for an entire class of organism at a global scale. These findings further support the importance of regional conservation management plans which target the species considered important by those people most affected by and involved in their conservation (Miller, 2005).

Our models were able to explain much of the variance in the interest people have in different species using biological and other traits of the species (Table 2). The importance of description year perhaps highlights that similar mechanisms were at play for the selection of species to describe by the early reptile taxonomists as they are today for the general internet-using public. People are also more interested in large and venomous reptiles (Gunnthorsdottir, 2001; Johnson et al., 2010; Ward et al., 1998; Woods, 2000). Importantly, we find that species listed by the IUCN as threatened with extinction attract more interest (see also

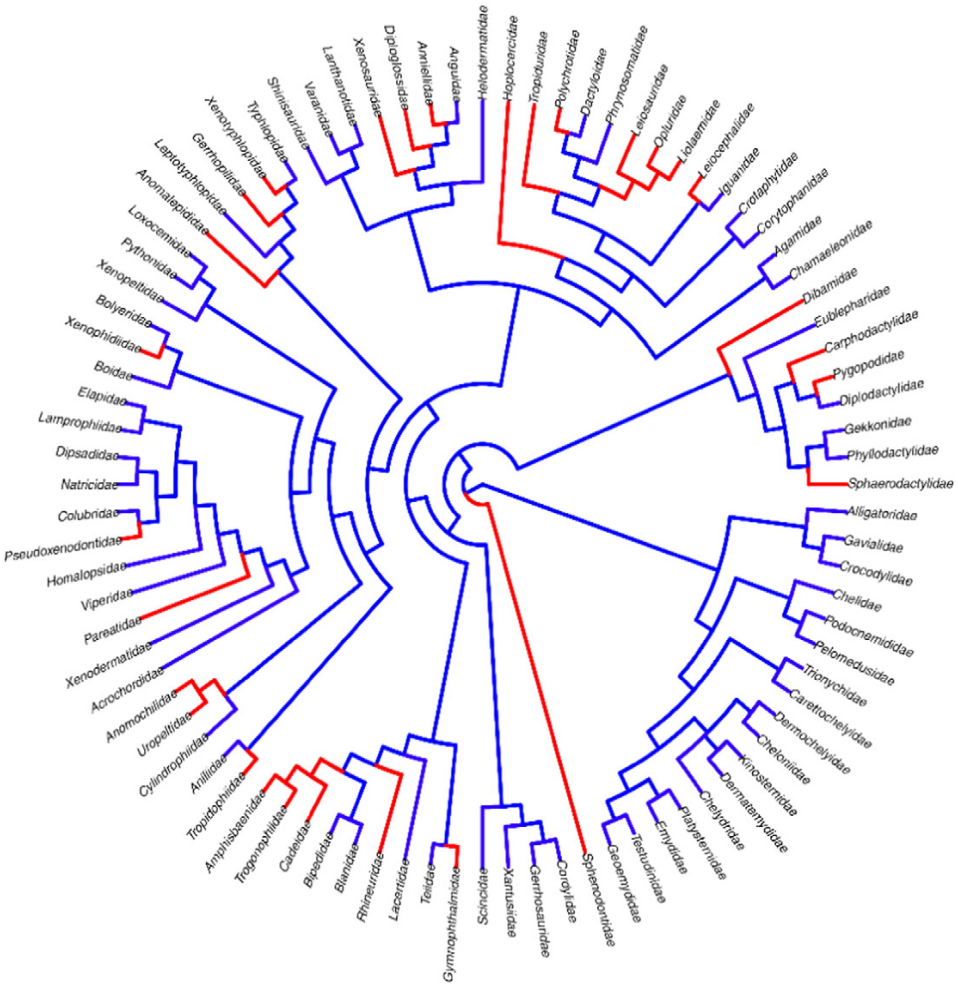


Fig. 3. Phylogeny of reptile families. Family branches coloured red are those without a single species represented in the top 5 percentile of total page views.

Johnson et al., 2010), irrespective of their body size, distribution or venomous status. This finding suggests that the IUCN red-listing process has intrinsic cultural impact, at least for reptiles (Ceballos and Ehrlich, 2002). Models for selected reptile families and major groups show group-specific differences in the importance and significance of particular variables (but are always congruent in sign with each other and with the overall reptile models). Consequently, interest between particular reptile groups is likely to be influenced by different factors. This finding could be of value when searching for effective flagship species for conservation (Barua et al., 2011; Verissimo et al., 2011).

Following the notion of protecting unique evolutionary linages or phylogenetic diversity (PD) we plotted on a family-level tree of reptiles those families that have at least one representative species which is highly visible in Wikipedia (Fig. 3). We find species in the top 5% of page views to be distributed widely across the phylogeny, leaving 33% of the 88 reptile families but only four distinct clades without a species of high interest. How interest, as measured by page views, relates to protection of phylogenetic diversity of course depends upon how we think interest influences conservation action. One conservative interpretation would simply be that a set of high-interest species exists which as passive recipients of conservation action, might effectively sample the phylogeny. At the other extreme, we could argue for direct use of page views as a measure of conservation importance. Page views in an online encyclopaedia are a quantifiable, omnipresent and easily obtainable metric of cultural interest, and could have obvious pragmatic benefits. Perhaps adopting such a metric together with other common conservation measures (threat, PD, function diversity etc.) could bring about a more holistic suite of parameters for designating species for conservation.

Using large online repositories and big-data approaches holds much promise for conservation biology (Correia et al., 2016). We present an initial exploration of reptile species viewed in different language editions of Wikipedia. Interpreting these results should be done with caution as there are several known biases inherent to Wikipedia (Brown, 2011; Graham et al., 2014; Miller and Murray, 2010; Wilson, 2014). As Arts et al. (2015) state, new technologies in conservation show "a need for rigorous evaluation [and] more comprehensive consideration of social exclusion". Wikipedia page views, if applied uncritically as measures of conservation priority, would directly exclude the cultural values of the majority of humanity. Nevertheless, as an increasing amount of human activity is represented online and more tools for analysing this activity are being developed and tested, approaches such as ours become more useful and comparable. Exploring patterns of other metrics within Wikipedia, as well as other digital text corpuses with perhaps either more inclusive, or more targeted cultural salience, could be very useful. Trying to match these broad online survey techniques with more traditional surveys could prove useful, as theories and methods for the latter are much more robust. As challenges of protecting biodiversity are increasing, we need to develop new tools, approaches and mind-sets to tackle it (Sharman and Mlambo, 2012); here we provide such an example.

Conflict of interest

All Authors claim no conflict of interest.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.biocon.2016.03.037.

Acknowledgments

We would like to thank all the members of the GARD consortium for providing data on reptile distributions. We would also like to thank David Chapple, Ricardo Correia, and two anonymous reviewers for insightful comments on earlier versions of this paper. JCM is supported by the Wilfrid Knapp Scholarship at St Catherine's College, Oxford. RG acknowledges the John Fell Fund of the University of Oxford for support.

References

Aiden, E., Michel, J.-B., 2013. Uncharted: Big Data as a Lens on Human Culture. Riverhead Books, New York, New York, USA.
Alves, R.R.N., Vieira, W.L.S., Santana, G.G., 2008. Reptiles used in traditional folk medicine: conservation implications. Biodivers. Conserv. 17, 2037–2049.
Alves, R.R.N., Léo Neto, N.A., Santana, G.G., Vieira, W.L.S., Almeida, W.O., 2009. Reptiles used for medicinal and magic religious purposes in Brazil. Appl. Herpetol. 6, 257–274.
Arts, K., van der Wal, R., Adams, W.M., 2015. Digital technology and the conservation of nature. Ambio 44, 661–673.
Barton, K., 2015. MuMIn: Multi-Model Inference, R package version 1.15.1. http://CRAN.R-project.org/package=MuMIn.
Barua, M., Root-Bernstein, M., Ladle, R., Jepson, P., 2011. Defining flagship uses is critical for flagship selection: a critique of the IUCN climate change flagship fleet. Ambio 40, 431–435.
Böhm, M., Collen, B., et al., 2013. The conservation status of the world's reptiles. Biol. Conserv. 157, 372–385.
Brooks, T., 2010. Conservation planning and priorities. In: Sodhi, N.S., Ehrlich, P.R. (Eds.), In Conservation Biology for All. Oxford University Press, Oxford, UK, pp. 199–219.
Brown, A.R., 2011. Wikipedia as a data source for political scientists: accuracy and completeness of coverage. Pol. Sci. Politics 44, 339–343.
Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. second ed. Springer-Verlag, New York.
Campbell, M., 2009. Repositioning zoogeography within the nature-culture borderlands: an animal geography of reptiles in southern Ghana. Appl. Geogr. 29, 260–268.
Campos, C.M., Greco, S., Ciarlante, J.J., Balangione, M., Bender, J.B., Nates, J., Lindemann-Matthies, P., 2012. Students' familiarity and initial contact with species in the Monte desert (Mendoza, Argentina). J. Arid Environ. 82, 98–105.
Ceballos, G., Ehrlich, P.R., 2002. Mammal population losses and the extinction crisis. Science 296, 904–907.
Ceríaco, L.M.P., 2012. Human attitudes towards herpetofauna: the influence of folklore and negative values on the conservation of amphibians and reptiles in Portugal. J. Ethnobiol. Ethnomed. 8, 8.
Ceríaco, L.M.P., Marques, M.P., Madeira, N.C., Vila-Viçosa, C.M., Mendes, P., 2011. Folklore and traditional ecological knowledge of geckos in southern Portugal: implications for conservation and science. J. Ethnobiol. Ethnomed. 7, 1–10.
Clucas, B., McHugh, K., Caro, T., 2008. Flagship species on covers of US conservation and nature magazines. Biodivers. Conserv. 17, 1517–1528.
Correia, R.A., Jepson, P.R., Malhado, A.C.M., Ladle, R.J., 2016. Familiarity breeds content: assessing bird species popularity with culturomics. PeerJ 4, e1728.
Cristancho, S., Vining, J., 2004. Culturally defined keystone species. Hum. Ecol. Rev. 11, 153–164.
Deb, D., Malhotra, K.C., 2001. Conservation ethos in local traditions: the West Bengal heritage. Soc. Nat. Resour. 14, 711–724.
Faith, D.P., 1992. Conservation evaluation and phylogenetic diversity. Biol. Conserv. 61, 1–10.
Feldman, A., Sabath, N., Pyron, R.A., Mayrose, I., Meiri, S., 2016. Body-Sizes and Diversification Rates of Lizards, Snakes, Amphisbaenians and the Tuatara. Glob. Ecol. Biogeo. 25, 179–187.
Garibaldi, A., Turner, N., 2004. Cultural keystone species: implications for ecological conservation and restoration. Ecol. Soc. 9, 1.
Giles, J., 2005. Internet encyclopaedias go head to head. Nature 438, 900–901.
Graham, M., Hogan, B., Straumann, R.K., Medhat, A., 2014. Uneven geographies of user-generated information: patterns of increasing informational poverty. Ann. Assoc. Am. Geogr. 104, 746–764.
Grenyer, R., Orme, C.D.L., Jackson, S.F., Thomas, G.H., Davies, R.G., Davies, T.J., Jones, K.E., Olson, V.A., Ridgely, R.S., Rasmussen, P.C., Ding, T.S., Bennett, P.M., Blackburn, T.M., Gaston, K.J., Gittleman, J.L., Owens, I.P.F., 2006. Global distribution and conservation of rare and threatened vertebrates. Nature 444, 93–96.
Gunnthorsdottir, A., 2001. Physical attractiveness of an animal species as a decision factor for its preservation. Anthrozoos 14, 204–215.
Halavais, A., Lackaff, D., 2008. An analysis of topical coverage of Wikipedia. J. Comput.-Mediat. Commun. 13, 429–440.
IUCN, 2014. The IUCN red list of threatened species Version 2014.3.
Johnson, P.J., Kansky, R., Loveridge, A.J., Macdonald, D.W., 2010. Size, rarity and charisma: valuing African wildlife trophies. PLoS One 5, e12866.
Jones, J.P.G., Andriamarovololona, M.M., Hockley, N., 2008. The importance of taboos and social norms to conservation in Madagascar. Conserv. Biol. 22, 976–986.
Kellert, S.R., 1985. Social and perceptual factors in endangered species management. J. Wildl. Manag. 49, 528–536.
Klemens, M.W., Thorbjarnarson, J.B., 1995. Reptiles as a food resource. Biodivers. Conserv. 4, 281–298.
Ladle, R.J., Jepson, P., 2008. Toward a biocultural theory of avoided extinction. Conserv. Lett. 1, 111–118.
Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., 2009. Life in the network: the coming age of computational social science. Science 323, 721–723.
Lindemann-Matthies, P., 2005. 'Loveable' mammals and 'lifeless' plants: how children's interest in common local organisms can be enhanced through observation of nature. Int. J. Sci. Educ. 27, 655–677.
Macdonald, E.A., Burnham, D., Hinks, A.E., Dickman, A.J., Malhi, Y., Macdonald, D.W., 2015. Conservation inequality and the charismatic cat: Felis felicis. Glob. Ecol. Conserv. 3, 851–866.
Meiri, S., Chapple, D.G., 2016. Biases in the current knowledge of threat status in lizards, and bridging the 'assessment gap'. Biol. Conserv. 204, 6–15.

Messner, M., DiStaso, M.W., 2013. Wikipedia versus Encyclopedia Britannica: a longitudinal analysis to identify the impact of social media on the standards of knowledge. Mass Commun. Soc. 16, 465–486.
Millennium Ecosystems Assessment, 2005. Ecosystems and Human Well-Being: Biodiversity Synthesis. World Resources Institute, Washington, D.C.
Miller, J.R., 2005. Biodiversity conservation and the extinction of experience. Trends Ecol. Evol. 20, 430–434.
Miller, J.C., Murray, H.B., 2010. Wikipedia in court: when and how citing Wikipedia and other consensus websites is appropriate. St. John's Law Review 84, 633–656.
Mills, L.S., Soulé, M.E., Doak, D.F., 1993. The keystone-species concept in ecology and conservation. Bioscience 43, 219–224.
Orme, C.D.L., Davies, R.G., Burgess, M., Eigenbrod, F., Pickup, N., Olson, V.A., Webster, A.J., Ding, T.S., Rasmussen, P.C., Ridgely, R.S., Stattersfield, A.J., Bennett, P.M., Blackburn, T.M., Gaston, K.J., Owens, I.P.F., 2005. Global hotspots of species richness are not congruent with endemism or threat. Nature 436, 1016–1019.
Prokop, P., Fančovičová, J., 2013. Does colour matter? The influence of animal warning coloration on human emotions and willingness to protect them. Anim. Conserv. 16, 458–466.
Pyron, R.A., Burbrink, F.T., 2014. Early origin of viviparity and multiple reversions to oviparity in squamate reptiles. Ecol. Lett. 17, 13–21.
R-Core-Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Ramstad, K.M., Nelson, N.J., Paine, G., Beech, D., Paul, A., Paul, P., Allendorf, F.W., Daugherty, C.H., 2007. Species and cultural conservation in New Zealand: maori traditional ecological knowledge of tuatara. Conserv. Biol. 21, 455–464.
Reeder, T.W., Townsend, T.M., Mulcahy, D.G., Noonan, B.P., Wood Jr., P.L., Sites Jr., J.W., Wiens, J.J., 2015. Integrated analyses resolve conflicts over squamate reptile phylogeny and reveal unexpected placements for fossil taxa. PLoS One 10, e0118199.
Rosenzweig, M.L., 1995. Species Diversity in Space and Time. Cambridge University Press, Cambridge, UK.
Samoilenko, A., Yasseri, T., 2014. The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics. EPJ Data Science 3, 1–11.
Schich, M., Song, C., Ahn, Y.-Y., Mirsky, A., Martino, M., Barabási, A.-L., Helbing, D., 2014. A network framework of cultural history. Science 345, 558–562.
Sharman, M., Mlambo, M.C., 2012. Wicked: the problem of biodiversity loss. GAIA 21, 274–277.
Shwartz, A., Cheval, H., Simon, L., Julliard, R., 2013. Virtual garden computer program for use in exploring the elements of biodiversity people want in cities. Conserv. Biol. 27, 876–886.
Stokes, D.L., 2007. Things we like: human preferences among similar organisms and implications for conservation. Hum. Ecol. 35, 361–369.
Taras, V., Rowney, J., Steel, P., 2009. Half a century of measuring culture: review of approaches, challenges, and limitations based on the analysis of 121 instruments for quantifying culture. J. Int. Manag. 15, 357–373.
Tisdell, C.A., 2014. Human Values and Biodiversity Conservation: The Survival of Wild Species. Edward Elgar Publishing, Cheltenham, UK.
Uetz, P., Hošek, J., 2015. The Reptile Database. http://reptile-database.org/.
Vane-Wright, R.I., Humphries, C.J., Williams, P.H., 1991. What to protect? — Systematics and the agony of choice. Biol. Conserv. 55, 235–254.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. fourth ed. Springer, New York, USA.
Verissimo, D., MacMillan, D.C., Smith, R.J., 2011. Toward a systematic approach for identifying conservation flagships. Conserv. Lett. 4, 1–8.
Ward, P.I., Mosberger, N., Kistler, C., Fischer, O., 1998. The relationship between popularity and body size in zoo animals. Conserv. Biol. 12, 1408–1411.
Whittaker, R.J., Araujo, M.B., Jepson, P., Ladle, R.J., Watson, J.E., Willis, K.J., 2005. Conservation biogeography: assessment and prospect. Divers. Distrib. 11, 3–23.
Wilson, E.O., 1992. The Diversity of Life. Cambridge, Massachusetts, Belknap.
Wilson, J.L., 2014. Proceed with extreme caution: citation to Wikipedia in light of contributor demographics and content policies. Vanderbilt J. Entertain. Technol. Law 16, 857–908.
Woods, B., 2000. Beauty and the beast: preferences for animals in Australia. J. Tour. Stud. 11, 25–35.
Yasseri, T., Sumi, R., Rung, A., Kornai, A., Kertész, J., 2012. Dynamics of conflicts in Wikipedia. PLoS One 7, 1–12.
Yasseri, T., Spoerri, A., Graham, M., Kertész, J., 2014. The most controversial topics in Wikipedia: A multilingual and geographical analysis. In: P., F., N., H. (Eds.), Global Wikipedia: International and Cross-cultural Issues in Online Collaboration. Scarecrow Press.
Yu, A.Z., Ronen, S., Hu, K., Lu, T., Hidalgo, C.A., 2015. Pantheon: a dataset for the study of global cultural production arXiv:1502.07310v1[physics.soc-ph].

Patterns 2

While Patterns 1 investigated biological traits, Patterns 2 focuses on how temporal patterns impact online interest. Using a larger subset of species than before (nearly 32,000 IUCN-listed species with pages in Wikipedia) and nearly three years of daily pageviews for over 126,000 Wikipedia pages, I found that seasonal patterns in pageviews are prevalent amongst pages for species. Furthermore, the prevalence of seasonality amongst species-pages corresponds in interesting ways with geography (seasonality tends to be more prevalent amongst languages primarily spoken at higher latitudes) and with biology (seasonality in pageviews tends to be more prevalent amongst taxonomic clades that undergo conspicuous phenological cycles such as flowering plants and insects).

Like Patterns 1, this paper contributes to understanding why people prefer certain species (specifically as seasonal markers). In a theme that remerges in Patterns 3, the findings here suggest that direct interactions with species in the wild may be a driver of online interest. In terms of data-driven approaches to identifying high interest species, Patterns 2 contributes the methodological insight that temporal variations need to be considered when comparing interest in species. As shown in Fig. 2, which species attract the most online attention can depend in part on timing. Finally, Patterns 2 demonstrates that interest in species undergoes temporal fluctuations and that online data can be used to monitor these fluctuations.

Submission status: Published. Mittermeier, J.C., Roll, U., Matthews, T.J., and R. Grenyer. 2019. *A season for all things: phenological imprints in Wikipedia usage and their relevance to conservation*. PLoS Biology doi.org/10.1371/journal.pbio.3000146.

Personal contribution*:* Lead author. I devised the initial concept of the paper, downloaded and curated the data, conducted the formal analysis, did the visualizations, and wrote the initial draft of the text. Co-authors assisted with discussing the initial concept, funding, reviewing the methodology, and editing drafts.

PLOS BIOLOGY

SHORT REPORTS

A season for all things: Phenological imprints in Wikipedia usage and their relevance to conservation

John C. Mittermeier[1]*, Uri Roll[2], Thomas J. Matthews[3,4], Richard Grenyer[1]

1 School of Geography and the Environment, University of Oxford, Oxford, United Kingdom, **2** Mitrani Department of Desert Ecology, The Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Midreshet Ben-Gurion, Israel, **3** School of Geography, Earth and Environmental Sciences, and Birmingham Institute of Forest Research, University of Birmingham, Edgbaston Birmingham, United Kingdom, **4** CE3C –Centre for Ecology, Evolution and Environmental Changes/Azorean Biodiversity Group and Univ. dos Açores–Depto de Ciências e Engenharia do Ambiente, Angra do Heroísmo, Açores, Portugal

* john.mittermeier@gmail.com

OPEN ACCESS

Citation: Mittermeier JC, Roll U, Matthews TJ, Grenyer R (2019) A season for all things: Phenological imprints in Wikipedia usage and their relevance to conservation. PLoS Biol 17(3): e3000146. https://doi.org/10.1371/journal.pbio.3000146

Academic Editor: Pedro Jordano, Estacion Biologica de Doñana CSIC, SPAIN

Received: October 26, 2018

Accepted: January 29, 2019

Published: March 5, 2019

Data Availability Statement: All data Wikipedia page-view data is open-access and publicly available from https://dumps.wikimedia.org. eBird data can be publicly accessed through eBird.org and the Cornell Lab of Ornithology. Data used to produce the figures in the paper have been submitted together with the manuscript.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Phenology plays an important role in many human–nature interactions, but these seasonal patterns are often overlooked in conservation. Here, we provide the first broad exploration of seasonal patterns of interest in nature across many species and cultures. Using data from Wikipedia, a large online encyclopedia, we analyzed 2.33 billion pageviews to articles for 31,751 species across 245 languages. We show that seasonality plays an important role in how and when people interact with plants and animals online. In total, over 25% of species in our data set exhibited a seasonal pattern in at least one of their language-edition pages, and seasonality is significantly more prevalent in pages for plants and animals than it is in a random selection of Wikipedia articles. Pageview seasonality varies across taxonomic clades in ways that reflect observable patterns in phenology, with groups such as insects and flowering plants having higher seasonality than mammals. Differences between Wikipedia language editions are significant; pages in languages spoken at higher latitudes exhibit greater seasonality overall, and species seldom show the same pattern across multiple language editions. These results have relevance to conservation policy formulation and to improving our understanding of what drives human interest in biodiversity.

Author summary

Digital information archives offer novel opportunities to study human attitudes towards nature and to better understand how people interact with other species of animals and plants. The insights gained from such studies may be able to inform conservation efforts. Our study uses time-series of views to pages in the online encyclopedia Wikipedia to look at how human interest in other species varies seasonally across a wide range of different languages. In total, we extracted pageviews for 31,751 species of plants and animals across 245 Wikipedia language editions. Spanning nearly three years, our data set comprises 2.33 billion pageviews across 126,697 pages. We tested each time-series in our data set to see

Abbreviations: AICc, Akaike Information Criterion corrected; IUCN, International Union for Conservation of Nature; NGO, nongovernmental organization.

how well it fit a seasonal pattern and in doing so found several interesting patterns. First, seasonality is a significant factor in when people view information for many plants and animals online; over 20% of all of our species pages met our criteria for seasonality. Second, the prevalence of seasonality varies across different biological classes and also across languages. These variations appear to reflect differences in the life history of species and in the geographic distribution of languages and can correspond to phenological patterns in nature. Our results are relevant to conservationists seeking to understand how interest in various plants and animals may fluctuate over time.

Introduction

Many aspects of human society respond to seasonal patterns. Seasonality can influence the foods people consume [1], the prevalence of diseases [2], and people's mental health [3], among much else. Human interactions with biodiversity can also be seasonal; people may await the blooming of a flowering plant or expect the return of migrating birds at a particular season, and these phenological patterns can be an essential component of the "cultural value" of a species [4]. Despite this, seasonal patterns in human interest in biodiversity have received little attention in the conservation literature. Here, we demonstrate that identifying which species people interact with on a seasonal cycle and assessing whether those patterns are consistent across human geography and cultural groups is relevant to understanding what drives interest in biodiversity and can help promote improved conservation actions.

Public attitudes towards species can have a profound impact on conservation, with popular preferences for particular plants and animals potentially affecting the allocation of conservation resources by both governments and nongovernmental organizations (NGOs) [5,6] and determining the success and effectiveness of conservation initiatives [7,8]. Accordingly, it is important to understand what attributes of a species contribute to increased human interest [6,9], and how an organism's popularity may vary as a result of factors such as a person's nationality [10,11], socioeconomic status and age [12], education levels, and gender [13,14]. Seasonal changes in human interest in plants and animals can have an important role in conservation in at least three ways: (a) by identifying species for which phenology forms a component of their "value," (b) by helping to reveal differences or similarities in how species are valued across cultural groups, and (c) by providing temporal awareness to help maximize the effectiveness of conservation marketing campaigns.

We analyze seasonal patterns in nearly three years of Wikipedia pageviews across 245 Wikipedia language editions for a large and taxonomically diverse group of species in order to explore (a) the prevalence of seasonal patterns in species pageviews, (b) differences in seasonal patterns across languages, (c) the effect of seasonality on the relative popularity of a species, and (d) the relationship between seasonal patterns in pageviews and phenology. As language is linked to human cultural identity and also has a geographic component, analyzing differences across languages provides insight into how seasonal patterns may vary culturally and geographically. Meanwhile, assessing the relationship between pageview seasonality and phenology helps in understanding the extent to which digital patterns parallel biological ones.

Large online data sets provide novel opportunities to examine interest in biodiversity at scales and resolutions that were previously inconceivable [15], and Wikipedia, the open-access encyclopedia, offers a unique resource that is well suited to making temporal comparisons in interest across large numbers of plants and animals. Wikipedia is currently the fifth most-visited website on the internet and, as of 2018, receives 14 to 16 billion pageviews per month across over 300

language editions [16,17]. Furthermore, Wikipedia's structure, in which each page corresponds to a specific entity, facilitates the comparison of topics by avoiding semantic challenges such as homonyms [15,18]. Previous research has demonstrated that studying temporal trends in Wikipedia pageviews can provide insight into real-world phenomena [3,19,20]. Within conservation, Wikipedia pageviews have been used to compare human interest in reptile species [21]. As with any online data source, it is important to note that Wikipedia pageviews are not representative of all conservation stakeholders (people with limited access to the internet or those who live in countries where Wikipedia is blocked or unpopular are not represented), but pageview data do reflect the interests of a large and growing demographic that is of significant relevance to conservation.

Results

Prevalence of seasonality in online views

We identified seasonal patterns in Wikipedia pageviews in species pages (126,697 pages for 31,751 species) and a large sample of randomly selected nonspecies pages (121,638 pages for 9,158 entities). We assessed seasonality by comparing the fit (using adjusted R^2 values) of detrended pageview data to sinusoidal models with either one or two annual peaks. A large proportion (20.2%) of pages in our species data set exhibit seasonal variation according to our criteria. This is significantly higher than among the randomly selected nonspecies pages, of which only 6.51% met our seasonality criteria ($\chi^2_1 = 10{,}083$, $p < 0.001$). Adjusted R^2 values for both single annual peak and double annual peak seasonal models were also significantly higher for species pages than for random pages. For the single annual peak model, for example, species-page mean adjusted $R^2 = 0.275$ (standard error = 7.45×10^{-4}), and random-page mean adjusted $R^2 = 0.131$ (standard error = 5.56×10^{-4}; Student $t_{230,730} = -155$, $p < 0.001$). In both species pages and random pages, the majority of seasonal pages fit a single annual peak (89.6% species, 82.5% random), with the remainder having two annual peaks. Aggregated at the species level, 25.2% of the species in our data set (8,015 of 31,751) show seasonality in at least one language edition. In many cases, seasonal patterns are striking and clearly correspond with phenological patterns (e.g., bird migration or breeding; Fig 1); however, seasonality also arises as a result of repeated cultural events, such as annual holidays (Fig 2).

For taxonomic classes with at least 100 species in our data set ($n = 20$), the proportion of seasonal pages ranged from approximately 5% (Cycadopsida [cycads], Anthozoa [sea anemones and corals], Cephalopoda [squid and octopus]) to over 30% (Liliopsida [monocotyledons], Insecta [insects], Equisetopsida [horsetails]; Fig 3). For classes with greater than 1,000 species ($n = 10$), the most seasonal classes were insects (34.9% of pages) and monocotyledons (30.7% of pages), whereas the least seasonal were mammals (14.8%) and Elasmobranchii (sharks and rays, 9.31%).

Variation across languages

Our data set of species pages includes pages from 245 Wikipedia language editions, and over half of the species in the data set (53.8%) appear in more than one language edition (mean pages per species = 3.99). Species tend not to show seasonality in all of the language editions in which they appear, even if they show strong seasonality in some, indicating that biogeographic and cultural complexity likely lies behind many interactions. For species that show seasonality in at least one language edition ($n = 8{,}015$), the mean percentage of seasonal pages per species is 40.7% (standard deviation = 21.9%), indicating that seasonal patterns are usually not consistent across languages. In only 126 cases (1.57%) is the same seasonal pattern (one or two peak) present across all language editions that a species occurs in, and in only 5 instances does this occur for a species that exists in more than five language editions (a migratory bird, two insects, and two flowering plants, all of them native to Europe).

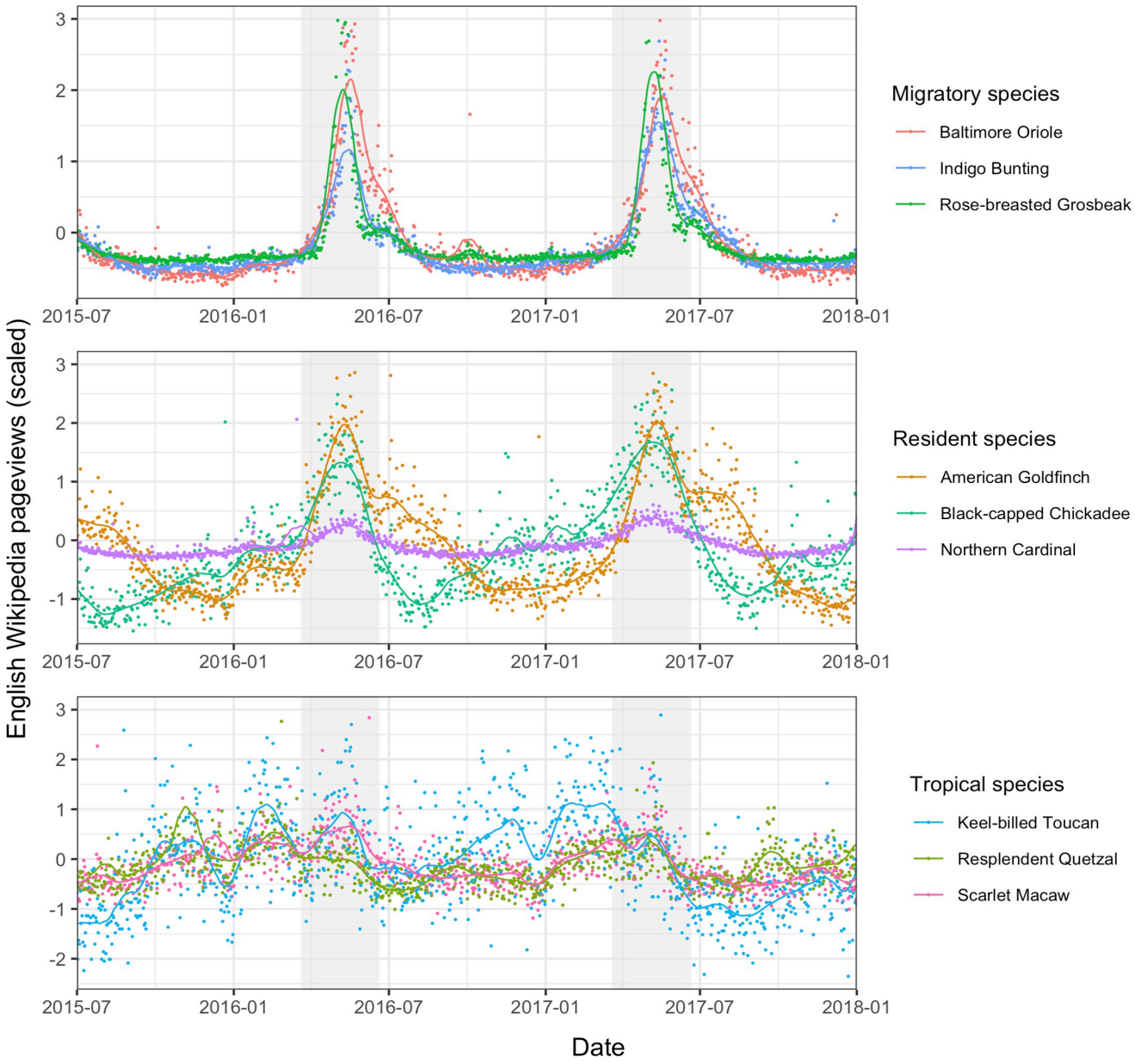


Fig 1. Daily pageviews in English-language Wikipedia for nine bird species. Pageviews for three migratory birds (top panel) show a strong seasonal peak coinciding with the bird's arrival on breeding grounds in the United States. Pageviews for three North American resident birds show more variable patterns (second panel), and pageviews for three tropical species (bottom panel) that do not occur in the US show fluctuations over the course of the year but no consistent seasonality. Investigating the drivers of these patterns for individual species could be a rich area for further study, in particular as to whether changes in pageviews can be related to trends in population abundance and location or to particular human activities. Background gray shading indicates the dates of spring in the Northern Hemisphere. Pageviews for each species are scaled by subtracting the mean and dividing by the standard deviation (for data used in plots, see S1 Data).

https://doi.org/10.1371/journal.pbio.3000146.g001

For languages with at least 100 species pages ($n = 60$), the number of seasonal pages is uncorrelated with the total number of pages ($p = 0.215$, Pearson's $r = 9.4 \times 10^{-3}$), indicating

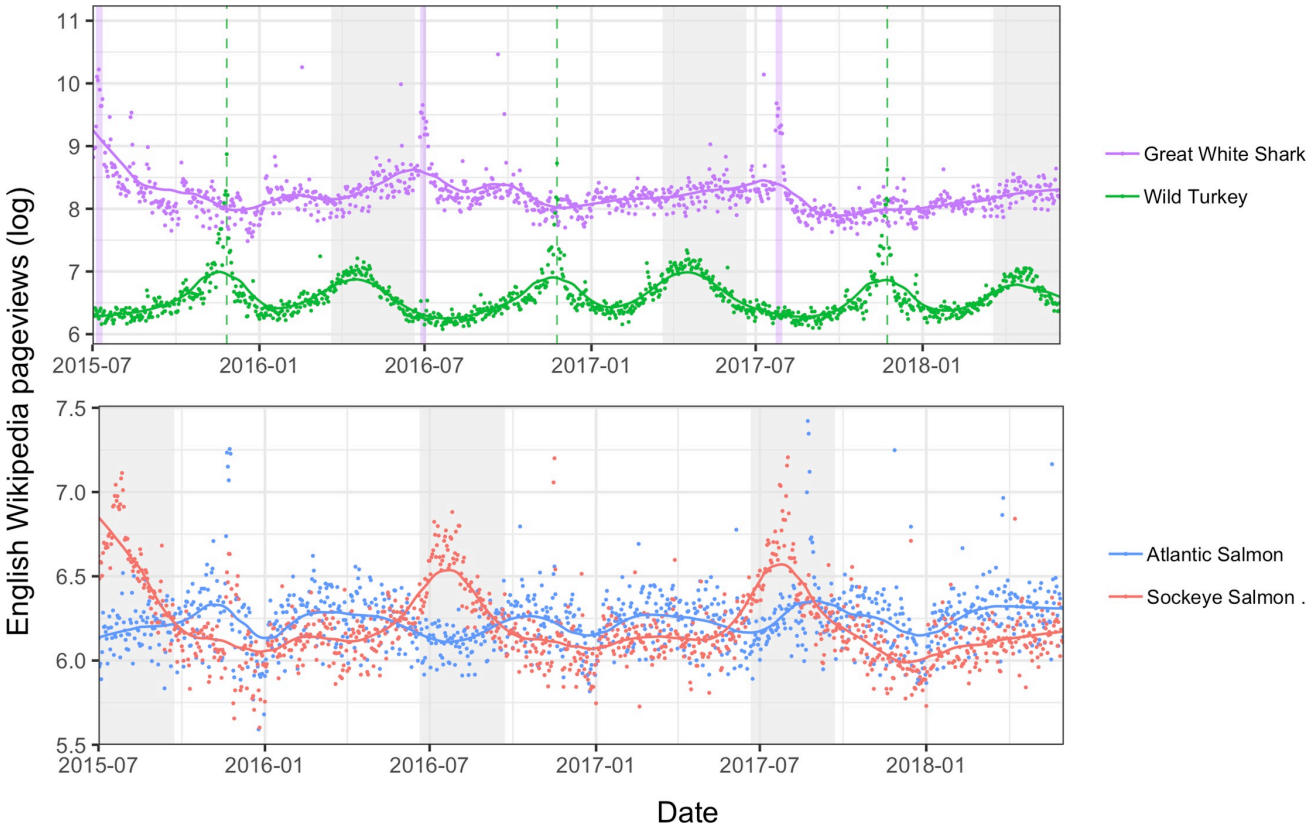


Fig 2. Effects of culture and phenology on the seasonality of interest in species. (Top panel) Patterns in Wikipedia pageviews respond to cultural influences as well as biological ones. English-language pageviews (logged) for great white shark *Carcharodon carcharias* (purple) are relatively stable throughout the year but show a brief spike during the days when "Shark Week" was aired on television by the Discovery Channel (days highlighted in purple). Pageviews for wild turkey *Meleagris gallopavo* (green) show a seasonal peak in the spring and a sharp peak during the Thanksgiving holiday in the US (date marked with a dashed line). The spring peak roughly coincides with the spring hunting season for wild turkey in many US states. (Bottom panel) The popularity of species relative to one another, as measured in Wikipedia pageviews, can vary as a result of seasonal fluctuations. Sockeye salmon *Oncorhynchus nerka* (red) and Atlantic salmon *Salmo salar* (blue) alternate in relative popularity depending on the time of year. Sockeye, which has a pronounced seasonal pattern, are more popular than Atlantic salmon in the Northern Hemisphere summer (dates shaded gray; for data used in plots, see S2 Data).

https://doi.org/10.1371/journal.pbio.3000146.g002

that smaller language editions are just as likely to have seasonality in their pageviews as larger ones. However, seasonality does show a significant positive relationship with capital city latitude (Fig 4; S1 Text), both when measured as the percentage of seasonal pages in a language (percent seasonal pages = $0.63 \times |\text{latitude}| + 0.24$, $p < 0.001$, adjusted $R^2 = 0.44$) and when calculated as the mean seasonality of all pages fitting a single seasonal peak in a language (mean seasonality = $4.5 \times 10^{-3} \times |\text{latitude}| + 0.13$, $p < 0.001$, adjusted $R^2 = 0.48$).

Impact of seasonality on relative popularity

Monthly rankings in the total pageviews for a species varied over the temporal period of our dataset (mean standard deviation = 1,850 ranks), with the average species shifting in its popularity (as measured by summed pageviews) relative to other species by 5.8%. As is to be expected statistically, these fluctuations were more pronounced for species with fewer views overall (mean standard deviation lower pageview quartile = 2,660 ranks; upper quartile = 520 ranks), indicating that relative popularity is more stable for the species that receive more

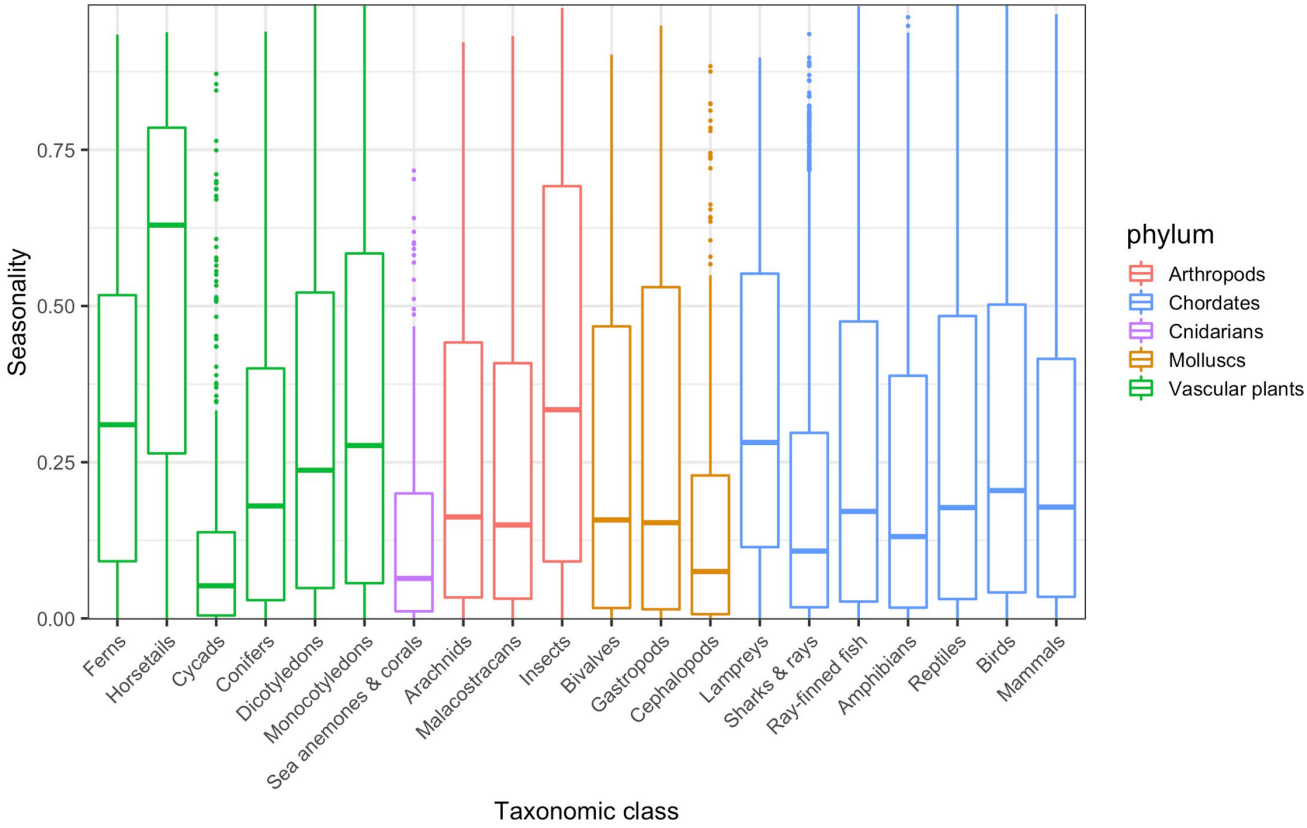


Fig 3. Prevalence of seasonality in Wikipedia pageviews varies across clades. Horizontal thick lines show median seasonality for species in 20 taxonomic classes in our data set that include over 100 species (n = 126,058 pages for 31,495 species). Bars indicate interquartile range, whiskers extend to 1.5× the interquartile range beyond the nearest quartile, and individual points are outliers beyond this. Seasonality is measured by the amount of variation in pageviews explained by a sinusoidal model with a single annual peak (via adjusted R^2; for data used in plots, see S3 Data).

https://doi.org/10.1371/journal.pbio.3000146.g003

views. Nevertheless, temporal fluctuations shifted the relative popularity of species over the course of a year even for some of the most viewed species overall.

Correspondence between pageview seasonality and eBird frequency

We used bird frequency records from four countries (Italy, Germany, Sweden, and the US), extracted from the citizen science database eBird.org, to correlate annual variations in the frequency with which a bird is recorded in a country with seasonal patterns in Wikipedia pageviews in that country's national language. For the 862 pages that we tested across our four target languages, 54.1% (466 pages) showed a significant relationship between monthly Wikipedia pageviews and monthly eBird frequency ($p < 0.05$), and 87.6% of these (408 pages; 47.3% of all pages) showed a significant positive relationship between the two (mean scaled coefficient = 0.640, mean adjusted R^2 = 0.42; S1 Fig). The proportion of bird pages in our data set with a significant positive relationship between monthly eBird frequency and monthly pageviews was relatively consistent across the four countries—49.1% of species in the US, 48.9% in Germany, 45.4% in Sweden, and 39.8% in Italy. For species that occurred in more than one of the four languages and/or countries evaluated (n = 146), only 34.2% (n = 50) showed a significant positive relationship between eBird frequency and pageviews across

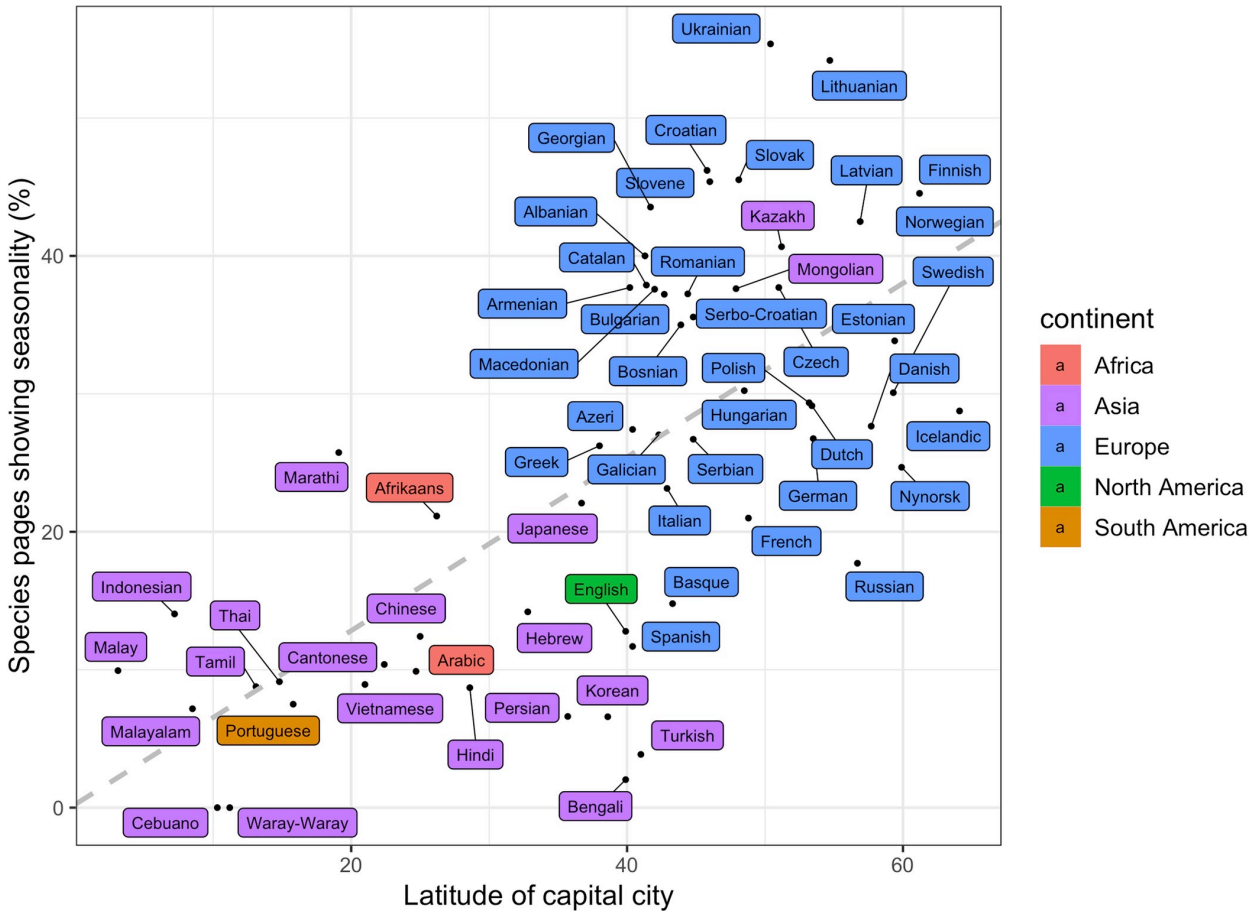


Fig 4. Seasonality and latitude of Wikipedia language editions. The percentage of seasonal pages for the 60 Wikipedia language editions in our data set that have over 100 species pages plotted against the absolute value of the latitude of the capital city of the country or province that accounts for the highest proportion of Wikipedia pageviews in that language (for data used in plots, see S4 Data).

https://doi.org/10.1371/journal.pbio.3000146.g004

multiple languages. Species that were more seasonal in their pageviews (i.e., had a better fit to the seasonal model) were more likely to have a positive relationship between eBird frequency and pageviews (coefficient = 2.09, standard error = 0.497, $p < 0.001$). In contrast, the total number of views that a page received was not a good predictor of whether there was a positive relationship between pageviews and eBird frequency (coefficient = 1.13×10^{-6}, standard error = 4.65×10^{-7}, $p = 0.015$).

Discussion

Using a large, online encyclopedia, we provide the first broad exploration of seasonal patterns of interest in nature across many species and cultures. We show that seasonal patterns play a widespread and significant role in online interest in plants and animals, with a quarter of the 31,751 species in our data set exhibiting seasonal patterns in at least one language. Wikipedia, and internet usage in general, may respond to seasonal fluctuations; however, the fact that seasonality is significantly more prevalent in species pages than among a random selection of Wikipedia articles suggests that human interactions with nature are particularly likely to be

seasonal. The drivers of these seasonal patterns are complex and variable (e.g., Figs 1 and 2). As evidenced by the bird species in our eBird sample, however, seasonality in online interest can often have a positive relationship with phenological patterns, such as variations in the local presence and abundance of a species. The variation in pageview seasonality across taxonomic clades also suggests links to phenology. Insects and flowering plants—for example, clades whose species frequently have conspicuous phenological variations—display higher seasonality than two clades of plants (cycads and conifers) and several groups of marine organisms (sharks and rays, sea anemones and corals, squid and octopus), in which phenological variations or reproductive events are likely to be less visible to the nonspecialist (Fig 3).

These broad patterns in seasonality can have direct relevance for conservation practice. Sharp and predictable temporal spikes in interest create a clear opportunity for NGOs seeking to maximize the impact of their fundraising campaigns. Exploring the existence of temporal trends in interest in a specific target species and investigating what drives those trends should be an initial first step when setting up fundraising campaigns. For this purpose, the methodology we present here could aid in effectively identifying and targeting flagship species [22]. Furthermore, the seasonal variations in relative popularity across species is a relevant methodological consideration for researchers comparing interest across sets of species, particularly when using online behavior as a proxy for interest or popularity. As we demonstrate, online views for some species can fluctuate significantly on an annual cycle, and these variations could be a confounding variable in investigating other drivers of interest. For example, the online popularity of two salmon species depends, in part, on when the data are collected (Fig 2). These fluctuations are particularly pronounced for less popular species but also impact highly viewed ones.

We also demonstrate the high importance of geographic and cultural differences (in our case, exemplified in language editions) in human interactions with other plants and animals. In our data, this is reflected in the fact that species seldom show the same seasonal pattern across Wikipedia language editions. It is important to note, however, that the potential drivers of these differences are complex. Although language-level differences in pageviews can result from biological or cultural factors, they can also arise from differences in the structure of Wikipedia. Language editions vary vastly in size; a minority have millions of articles and users, whereas most have considerably fewer. The fact that we were able to detect predictable variation in seasonality between languages despite these differences (Fig 4) underscores the significance of seasonal patterns. Being aware of these seasonal patterns can benefit conservation campaigns, efforts to raise awareness, and educational endeavors. With the addition of other more precisely geolocated digital tools, it would be possible to identify (and manage) areas where people go to observe particular phenological phenomena (e.g., [23]).

Overall, we highlight the utility of culturomic tools in broad explorations of the seasonal aspect of human–nature interactions and demonstrate how phenology, geography, and culture can play intricate roles when exploring the seasonality of interest in nature. There is fertile ground for further study of some of the patterns that appear in this initial exploration. As always, drawing broad conclusions when using these tools should be done with caution, particularly if generalizing beyond the specific user base of the online resource [15]. Nevertheless, we feel that these methods hold much promise in elucidating the seasonal component of human–nature interactions, which has, to date, been understudied and underappreciated.

Materials and methods

Data

Our data set comprises 2.33 billion pageviews for 126,697 pages across 245 Wikipedia languages with each page having a corresponding time-series of daily pageviews over 1,067 days.

The data set contained pages for 31,751 species representing 52 taxonomic classes and 1,611 families. Our list of species and subspecies, representing a wide range of taxa, was compiled from Wikidata, a document-oriented database that provides data to projects run by the Wikimedia foundation. Using the Wikidata Query Service (https://query.wikidata.org), we extracted items that had both a Global Biodiversity Information Facility ID (identifier: P846) and an International Union for Conservation of Nature (IUCN) conservation status (statement: P141) on 05 June 2018. This two-part validation helped to control against items tagged erroneously. Additional taxonomic data (scientific name, higher taxa information) for each of these was compiled using the package "rgbif"[24] in R [25], and corresponding Wikipedia pages were extracted using "rvest" [26]. Pageviews for the period between 01 July 2015 and 02 June 2018 were summarized with "pageviews" [27]. These dates were selected to maximize the time frame of our data set; 01 July 2015 is the date that Wikipedia began its current system for archiving pageview data, and thus the earliest date for which we could extract pageviews with these methods, and 02 June 2018 was the date of our data extraction. To assess seasonality, we restricted our analyses to pages that received, on average, at least one view per day (using Tukey's biweight mean; package "dplR"[28]). We used these same methods to extract a comparative data set of 2.45 billion pageviews for 121,638 pages for 9,158 randomly selected non-species Wikidata entities across 290 Wikipedia languages. Random pages were selected by generating 10,000 random integers between 0 and 1 million (sample.int function in R) and then converting these into Wikidata entity IDs. The resulting Wikidata entities were cross-referenced against our list of species, and species pages were removed from the random sample.

Determination of seasonality

For each series of pageviews (corresponding to one Wikipedia page) in our data sets, we fitted a locally estimated scatterplot smoothing (LOESS) model to the logged daily views using the R package 'fANCOVA" [29]. Smoothing parameters for the LOESS were chosen using the Akaike Information Criterion corrected for small sample size (AICc [30]). We then used this model's fitted values to test for seasonality by (a) identifying annual consistency in the time-series by testing the fit between the data from the first year (days 1–365) and the second year (days 366–730) and (b) fitting sinusoidal models with one or two annual peaks to the detrended (residuals of a linear model) daily views of each time-series. The first two years of the data were used for testing the model, because these were complete cycles of 365 days (the third year was approximately one month short of a full year due to the timing of our study). We used a linear model rather than a simple null model in order to account for overall growth in the size and use of Wikipedia over the sample period. We used a log-likelihood test [31] to select between the two seasonal models and set thresholds based on adjusted R^2 to classify whether a time-series was consistent (year 1 pageviews predicted year 2 pageviews with adjusted $R^2 > 0.5$) and whether it fit the sinusoidal model (adjusted $R^2 > 0.5$ for time-series fitting a single annual peak, and > 0.3 for a double annual peak). The selected thresholds for adjusted R^2 were derived by manually reviewing the data and comparing results to well-established seasonal patterns, such as the annual variation in day length. Sensitivity testing with lower (adjusted $R^2 > 0.3$) and higher thresholds (adjusted $R^2 > 0.7$) varied the results in a predictable manner but did not significantly alter the relative proportion of seasonal patterns relative to each other within species pages or between species pages and random pages (S1 Text). Initial tests that used sinusoidal models with three and four annual peaks were never selected by the log-likelihood test, and these models were excluded from the final analysis. A time-series met our criteria for being seasonal if it passed both the thresholds for annual consistency and the best-fitting sinusoidal model (with either one or two peaks).

Linguistic variation in Wikipedia pageviews

Wikipedia language editions vary greatly in their total pages, and as expected, our data included a large range in the number of species pages per language (range 1–28,393; mean 517). For languages that contained more than 100 species pages ($n = 60$), we correlated pageview seasonality with language edition size (measured by the total number of species pages in the language) and modeled the relationship between pageview seasonality and the absolute value of the latitude of the capital city of the country that accounts for the highest percentage of Wikipedia pageviews in that language (e.g., Portuguese → Brazil → Brasilia → 15.8˚ S; Wikipedia views by country from [32]; see S1 Text for details). We explored temporal trends in the rank popularity of species by summarizing the monthly views across all of the pages for a species and calculating its monthly ranking against other species in the data set.

Concordance between Wikipedia seasonality and eBird frequency data

We extracted bird frequency records for Italy, Germany, Sweden, and the US from eBird.org [33] using "rebird" [34] (see S1 Text for details). For bird pages in the relevant language for each of the four countries (Italian, German, Swedish, and English, respectively), we extracted pages that (a) met our criteria for pageview seasonality and (b) covered a species that occurred in the country in question and therefore had an eBird frequency distribution (species per country: US, $n = 472$; Italy, $n = 108$; Germany, $n = 141$; Sweden, $n = 141$). Accordingly, each of the 862 pages in our data set corresponds to a specific species and country, and species that occur in more than one country are represented by separate pages for each country. For each page, we used a linear model to assess the relationship between monthly eBird frequency and monthly pageview totals. Significance values were adjusted using the false discovery rate for multiple comparisons [35].

Supporting information

S1 Text. Additional information on methods to establish thresholds for seasonality, assign latitude to Wikipedia languages, and compare eBird frequency and Wikipedia pageview data.
(DOCX)

S1 Fig. Histogram of the distribution of significant correlation coefficients (FDR-adjusted $p < 0.05$) between monthly eBird frequency and monthly pageviews to seasonal Wikipedia pages for bird species that occur in Germany, Sweden, Italy, and the US. The distribution of values indicates that monthly increases in pageviews often correspond temporally with increased observations of a bird species in a given country. FDR, false discovery rate.
(TIF)

S1 Data. Data used in Fig 1.
(CSV)

S2 Data. Data used in Fig 2.
(CSV)

S3 Data. Data used in Fig 3.
(CSV)

S4 Data. Data used in Fig 4.
(CSV)

Acknowledgments

JCM is supported by a Wilfrid Knapp Scholarship at St Catherine's College. All authors acknowledge the efforts of Wikipedia and the Wikimedia Foundation in collecting and making available via public license the content, metadata, and pageview data on which this study depends.

Author Contributions

Conceptualization: John C. Mittermeier, Uri Roll, Thomas J. Matthews, Richard Grenyer.

Data curation: John C. Mittermeier.

Formal analysis: John C. Mittermeier, Thomas J. Matthews.

Funding acquisition: Uri Roll.

Investigation: John C. Mittermeier.

Methodology: John C. Mittermeier, Uri Roll, Thomas J. Matthews, Richard Grenyer.

Resources: John C. Mittermeier.

Software: John C. Mittermeier, Thomas J. Matthews.

Validation: John C. Mittermeier.

Visualization: John C. Mittermeier.

Writing – original draft: John C. Mittermeier.

Writing – review & editing: John C. Mittermeier, Uri Roll, Thomas J. Matthews, Richard Grenyer.

References

1. Rossato SL, Olinto MTA, Henn RL, Moreira LB, Camey SA, Anjos LA, et al. Seasonal variation in food intake and the interaction effects of sex and age among adults in southern Brazil. Eur J Clin Nutr. 2015; 69(9): 1015–22. https://doi.org/10.1038/ejcn.2015.22 PMID: 25828623
2. Carneiro HA, Mylonakis E. Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. Clin Infect Dis. 2009; 49(10): 1557–64. https://doi.org/10.1086/630200 PMID: 19845471
3. Dzogang F, Lansdall-Welfare T, Cristianini N. Seasonal Fluctuations in Collective Mood Revealed by Wikipedia Searches and Twitter Posts. IEEE Int Conf Data Min Work ICDMW. 2017; 931–7.
4. Garibaldi A, Turner N. Cultural keystone species: Implications for ecological conservation and restoration. Ecol Soc. 2004; 9(3): 1.
5. Martín-Forés I, Martín-López B, Montes C. Anthropomorphic Factors Influencing Spanish Conservation Policies of Vertebrates. Int J Biodivers. 2013. https://doi.org/10.1155/2013/142670
6. Macdonald EA, Burnham D, Hinks AE, Dickman AJ, Malhi Y, Macdonald DW. Conservation inequality and the charismatic cat: Felis felicis. Glob Ecol Conserv. 2015; 3: 851–66.
7. Kellert SR. Factors in Endangered Social and Perceptual Species Management. J Wildl Manage. 1982; 49(2): 528–36.
8. Stokes DL. Things we like: human preferences among similar organisms and implications for conservation. Hum Ecol. 2007; 35: 361–9.
9. Verissimo D, Pongilluppi T, Cintia M, Santos M, Develey PF, Fraser I, et al. Using a Systematic Approach to Select Flagship Species for Bird Conservation. Conserv Biol. 2013; 28(1): 269–77. https://doi.org/10.1111/cobi.12142 PMID: 24033848
10. Bowen-Jones E, Entwistle A. Identifying appropriate flagship species: the importance of culture and local contexts. Oryx. 2002; 36(2): 189–95.
11. Özel M, Prokop P, Uşak M. Cross-Cultural Comparison of Student Attitudes toward Snakes. Soc Anim. 2009; 17(3): 224–40.

12. Kellert SR. Attitudes toward animals: Age-related development among. J Environ Educ. 1985; 16(3): 29–39.
13. Pinheiro LT, Fabrício J, Rodrigues M, Borges-nojosa DM. Formal education, previous interaction and perception influence the attitudes of people toward the conservation of snakes in a large urban center of northeastern Brazil. J Ethnobiol Ethnomed. 2016. https://doi.org/10.1186/s13002-016-0096-9 PMID: 27324788
14. Ceríaco LM. Human attitudes towards herpetofauna: the influence of folklore and negative values on the conservation of amphibians and reptiles in Portugal. J Ethnobiol Ethnomed. 2012; 8(1): 8.
15. Ladle RJ, Correia RA, Do Y, Joo GJ, Malhado ACM, Proulx R, et al. Conservation culturomics. Front Ecol Environ. 2016; 14(5): 269–75.
16. Wikimedia Stats. Page Views for Wikipedia [Internet]. 2018 [cited 2018 Jul 16]. Available from: https://stats.wikimedia.org/EN/TablesPageViewsMonthlyCombined.htm.
17. Wikimedia. List of Wikipedias [Internet]. 2018 [cited 2018 Jul 16]. Available from: https://meta.wikimedia.org/wiki/List_of_Wikipedias.
18. Roll U, Correia RA, Berger-Tal O. Using machine learning to disentangle homonyms in large text corpora. Conserv Biol. 2018; 32(3): 716–24. https://doi.org/10.1111/cobi.13044 PMID: 29086438
19. Vilain P, Larrieu S, Cossin S, Caserio-Schonemann C, Filleul L. Wikipedia: a tool to monitor seasonal diseases trends? Online J Public Health Inform. 2017; 9(1): 1004239.
20. Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, et al. Forecasting the 2013–2014 Influenza Season Using Wikipedia. PLoS Comput Biol. 2015; 11(5): 1–29.
21. Roll U, Mittermeier JC, Diaz GI, Novosolov M, Feldman A, Itescu Y, et al. Using Wikipedia page views to explore the cultural importance of global reptiles. Biol Conserv. 2016; 204: 42–50.
22. Verissimo D, MacMillan DC, Smith RJ. Toward a systematic approach for identifying conservation flagships. Conserv Lett. 2011; 4: 1–8.
23. Hausmann A, Toivonen T, Heikinheimo V, Tenkanen H, Slotow R, Minin E DI. Social media reveal that charismatic species are not the main attractor of ecotourists to sub-Saharan protected areas. Sci Rep. Springer US; 2017; 7(1): 1–9. https://doi.org/10.1038/s41598-016-0028-x
24. Chamberlain S, Barve V, McGlinn D, Oldoni D, Geffert L, Ram K. rgbif: interface to the Global "Biodiversity" Information Facility API. 2018. Available from: https://cran.r-project.org/web/packages/rgbif/index.html. [cited 2018 Jun 02].
25. R Core Team. R: A language and environment for statistical computing, v. 3.5.0 [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available from: https://www.r-project.org/. [cited 2018 Jun 02].
26. Wickham H. rvest: easily harvest (scrape) web pages [Internet]. 2016 [cited 2017 May 1]. Available from: https://cran.r-project.org/web/packages/rvest/index.html.
27. Keyes O, Lewis J. pageviews: an API client for wikimedia traffic data version 0.3.0 [Internet]. 2016 [cited 2017 Jan 1]. Available from: https://cran.rstudio.com/web/packages/pageviews/index.html.
28. Bunn A, Korpela M, Biondi F, Campelo F, Merian P, Qaedan F, et al. dplR: dendrochronology program library in R [Internet]. 2017 [cited 2017 May 1]. Available from: https://cran.r-project.org/web/packages/dplR/index.html.
29. Wang X-F. fANCOVA: nonparametric analysis of covariance [Internet]. 2010 [cited 2018 May 1]. Available from: https://cran.r-project.org/web/packages/fANCOVA/index.html.
30. Hurvich CM, Simonoff JS, Tsai C-L. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J R Stat Soc Ser B Statistical Methodol. 1998; 60(2): 271–93.
31. Hothorn T, Zeileis A, Farebrother RW, Cummins C, Millo G, Mitchell D. lmtest: Testing Linear Regression Models [Internet]. 2018 [cited 2018 Jul 20]. Available from: https://cran.r-project.org/web/packages/lmtest/index.html.
32. Zachte E. Wikimedia Traffic Analysis Report: page views per wikipedia language [Internet]. 2018 [cited 2018 Jul 23]. Available from: https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm.
33. Sullivan BL, Wood CL, Iliff MJ, Bonney RE, Fink D, Kelling S. eBird: A citizen-based bird observation network in the biological sciences. Biol Conserv. 2009; 142(10): 2282–92.
34. Maia R, Chamberlain S, Teucher A, Pardo S. rebird: R client for the eBird Database of Bird Observations [Internet]. 2018 [cited 2018 Aug 01]. Available from: https://cran.r-project.org/web/packages/rebird/index.html.
35. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach for mulitple testing. J Roy Stat Soc Series B. 1995; 57: 289–300.

Patterns 3

The relationship between the physical presence of a species in a region and its online interest emerged as a theme in Patterns 2. In that paper, I did preliminary investigations into the relationship between these two variables by testing whether birds that were more seasonal in their occurrence in a region were also more seasonal in their pageviews. Patterns 3 expands on this question and assesses it across a broader geographic context. Rather than investigating temporal variations, however, Patterns 3 focuses on the spatial component. It looks at whether the frequency with which a bird is observed in a region correlates with its online popularity. The results show conclusively that it does. In some languages, the frequency with which species are recorded in a region can account for up to 50% of the variance in Wikipedia pageviews. This result contributes to understanding why some species attract more interest from people than others. For birds at least, interest in species is generated by the opportunity to interact with them in the wild.

Patterns 3 also provides insights into the other major conclusions of the thesis. In using big data to select high interest species, it is important to take geography into account. Otherwise, which species rank highest will be influenced by the geographic distribution of people represented by the dataset (Methods 1 touches on this as well). Furthermore, the fact that interest correlates with the presence of species suggests that online data such as pageviews could potentially be used to track fluctuations in the distribution and abundance of species.

Submission status: Submitted to Conservation Biology 6 October 2019. Mittermeier, J.C., Roll, U., Matthews, T.J., Correia, R., and R. Grenyer. *It is good to be common: birds that are more frequently encountered in the wild generate higher online interest.*

Personal contribution*:* Lead author. I devised the initial concept of the paper and the methodology, downloaded and curated the data, conducted the formal analysis, and wrote the initial draft of the text. Co-authors assisted with reviewing the methodology and editing drafts.

It is good to be common: birds that are more frequently encountered in the wild generate higher interest online

John C. Mittermeier[1], Uri Roll[2], Thomas J. Matthews[3,4], Ricardo Correia[5,6], Rich Grenyer[1]

[1] School of Geography and the Environment, University of Oxford, South Parks Road, Oxford, OX1 3QY, UK

[2] Mitrani Department of Desert Ecology, The Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Midreshet Ben-Gurion 8499000, Israel

[3] School of Geography, Earth and Environmental Sciences, and Birmingham Institute of Forest Research, University of Birmingham, Edgbaston Birmingham, UK

[4] CE3C – Centre for Ecology, Evolution and Environmental Changes/Azorean Biodiversity Group and Univ. dos Açores – Depto de Ciências e Engenharia do Ambiente, Angra do Heroísmo, Açores, Portugal

[5] DBIO & CESAM-Centre for Environmental and Marine Studies, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

[6] Institute of Biological and Health Sciences, Federal University of Alagoas, Maceió, Brazil

Keywords: Wikipedia, conservation culturomics, eBird, birds

Abstract

Many once common and widespread species are declining in their abundance and extent of distribution. In addition to the ecological and evolutionary consequences of this global

‘defaunation,’ it is important to understand how the declining populations and regional extinctions of species impact public engagement with biodiversity. Here we explore how the regional occurrence of bird species and the frequency with which people encounter them in the wild corresponds to the extent of interest that those species receive online. We evaluate data from a large sample of human-wildlife interactions (367 million Wikipedia pageviews) across a range of cultural contexts (25 different languages and geographic regions) and compare it to distributional and occurrence data derived from eBird.org. Across languages, we find that the internet-using public tends to be more interested in birds that occur in their local region than those that occur outside of it. Amongst those species that occur in a region, the frequency with which people encounter them in the wild predicts the attention they receive online. This effect is strong, with the local encounter rate of species predicting up to 50% of the variance in pageviews in some languages. The most commonly encountered birds play an outsize role in generating online interest. Across languages, the twenty most common birds account for an average 16% of the total pageviews for bird species while only consisting of 2.6% of the total pages. We find that the local abundance of a bird is a significantly stronger predictor of online interest than its body size, a trait often associated with high public interest. Our results help to provide insight into what drives public interest in species. Most importantly, we quantify the vital role that common species play in generating public awareness of biodiversity and the significance of facilitating opportunities for people to engage with these common species in the wild.

Introduction

Populations of many common species are declining at alarming rates (Dirzo et al. 2014; Inger et al. 2015; Rosenberg et al. 2019), and there are now an estimated 50% fewer individual

wild animals than there were just over a century ago (Ceballos et al. 2017). This global 'defaunation' has profound ecological and evolutionary consequences (Gaston & Fuller 2008; Oliver et al. 2015; Young et al. 2016). It also impacts people how and where people interact with and experience nature. Fewer animals mean fewer opportunities for people to encounter species in the wild, trends already exacerbated by urbanization and greater proportions of time spent indoors (Miller 2005; Soga & Gaston 2016). Likewise, receding populations and regional extinctions mean that many species will vanish from people's local regions (Ceballos et al. 2017). While the prevalence of these defaunation trends has received attention (e.g. Dirzo et al. 2014; Young et al. 2016), the extent to which they will influence public attitudes towards and engagement with declining species remains less well understood.

Exposure to natural world is important to generating pro-environmental attitudes and increasing people's likelihood to support conservation actions and policies (Wells & Lekies 2006; Soga & Gaston 2016). The importance of direct interactions with wild animals in generating interest in specific species, however, is not as clear. On the one hand, public interest in species may have little to with their local occurrence and abundance. Most of the species that rank highest in their 'charisma' and popularity are large-bodied animals that the majority of people never encounter in the wild (Colléony et al. 2017). Furthermore, many people have limited awareness of the distribution and abundance of local species particularly in urbanized societies (Bebbington 2005; Pilgrim et al. 2008), and interactions with biodiversity increasingly occur virtually on television or computer screens where the population size and distribution of an organism is not a factor. Species that have a higher extinction risk according to the IUCN Red List, and thus a smaller population, may actually be more popular in some cases (Roll et al. 2016), suggesting that population size can have a negative impact on public interest. In other

situations, however, the abundance of species as well as their occurrence in people's local regions have been shown to be important to increasing public interest in species. For butterflies in the UK and birds in Poland, the commonness of a species, and thus the likelihood of people having a direct encounter with it in the wild, is positively correlated with the attention that species receives online (Zmihorski et al. 2013). Likewise, in Brazil, species that overlap with areas of higher human population density and are therefore more familiar are the focus of greater public interest (Correia et al. 2016). In the context of planet's accelerating defaunation, understanding the prevalence and magnitude of this relationship between direct encounters with wild species and the extent of public interest they receive is important. If the frequency with which people encounter an animal in the wild significant driver of public attention across many cultural contexts, then local extinctions and population declines have the potential to result in an 'extinction of interest' in many species.

Online 'big data' provide the potential to examine patterns of human attention across vast numbers of people in the internet-using public and over broad geographic and cultural contexts. The data generated by people's online actions can provide insight into a range of 'real-world' human preferences and behaviors (e.g. McIver & Brownstein, 2014; Paul & Dredze, 2011). In conservation, recent studies have used internet data to investigate questions around visitation to protected areas (Hausmann et al. 2018), seasonal fluctuations in interest in biodiversity (Mittermeier et al. 2019) and monitoring of public perceptions towards species (Soriano-Redondo et al. 2017). More broadly, the potential to apply online data to conservation has been recognized as a new research area called 'conservation culturomics' (Ladle et al. 2016).

Here we combine information from two online data sources, eBird and Wikipedia, to investigate the relationship between the regional occurrence and abundance of a bird species and

the extent of online interest that it receives. eBird is a citizen-science database that provides global occurrence data on the distribution and abundance of bird species (Sullivan et al. 2009; Wood et al. 2011). The eBird dataset contains over 560 million records, making it is the largest contributor of biodiversity records to the Global Biodiversity Information Facility (GBIF) and the largest source of occurrence and abundance data for global bird species. Previous studies have used eBird data to monitor the migration of birds (Fink et al. 2013; Walker & Taylor 2017) and examine patterns in avian distribution and abundance (Clark 2017). In areas with high sampling effort, eBird data produce comparable estimates of species diversity and abundance as traditional ornithological survey methods (Callaghan & Gawlik 2015).

Wikipedia is the fifth most-trafficked website on the internet (Alexa 2019) and widely used as a source of public information and knowledge acquisition with resources in over 300 different languages (Wikimedia 2019a). Comparative assessments of Wikipedia pageviews have been used to quantify the popularity of famous people (Skiena & Ward 2014; Yu et al. 2016), predict stock market fluctuations and box office revenues (Mestyán et al. 2013; Moat et al. 2013), and monitor outbreaks of infectious diseases (Generous et al. 2014). In the conservation context, Wikipedia pageviews have been used to assess interest in reptile species (Roll et al. 2016) and compare seasonal patterns in biodiversity awareness (Mittermeier et al. 2019). The extent of knowledge that people have about a species has been shown to correlate positively with their interest in conserving it (Lindemann-Matthies 2005; Veríssimo et al. 2017), and as one of the world's most-widely used information resources, Wikipedia provides insight into which species people seek to increase their knowledge of.

By combining information on avian distributions and occurrences from eBird with pageview data from Wikipedia, we examine the relationship between the regional occurrence

and abundance of a bird species and its level of public interest across over a vast number of human-wildlife interactions. We address three questions with these data. First, we test whether Wikipedia users tend to be more interested in the species that occur in their local region as opposed to those that occur outside of it. Second, we assess whether the abundance of a species in a region (measured as the frequency with which it is reported in eBird) impacts its online interest. Finally, we compare the importance of a bird's regional abundance to its body size as a predictor of its Wikipedia interest. Large body size is often cited as the preeminent factor in driving public interest in species (Leader-Williams & Dublin 2000; Small 2012; Albert et al. 2018), and thus provides a comparative measure to contextualize the importance of abundance.

Methods

Data selection and extraction

Following previous studies, we use the number of pageviews that a page receives in Wikipedia as a measure of interest in that page (e.g. Skiena & Ward 2014; Roll et al. 2016). Wikipedia does not include location information with each of its pageviews. However, summary data with the percentage of views a language edition receives from a given country are available (Zachte 2018). Thus, we use Wikipedia's language editions as a coarse proxy for geography by pairing the views received in a language with the geographic region that accounts for the majority of that language's pageviews. This method follows previously published studies (Mittermeier et al. 2019). For our dataset, we selected language-region pairs where: a) the language has a large and robust Wikipedia edition; b) the pageviews to that language are strongly associated with a single geographic region; and c) the region has a high number of unique sampling events in eBird.

Wikipedia—We identified Wikipedia editions for non-artificial languages that as of 22 June 2019 had over 100,000 articles, a Wikipedia editing depth higher than 10 and over 50% of the language's pageviews originating from a single country (Zachte 2018; Wikimedia 2019a). Wikipedia editing depth provides a measure of the language edition's quality (Wikimedia 2019b). We identified pages for bird species in Wikipedia by using the Wikidata Query Service (https://query.wikidata.org/) to extract a list of entities in Wikidata tagged with an eBird taxon ID (Wikidata property: P3444) as of 23 June 2019. Wikidata is a secondary database that collects structured data for Wikimedia projects, including all Wikipedia language editions (https://www.wikidata.org). We cross-referenced our list of entities from Wikidata with the eBird/Clements global avian taxonomy (Clements et al. 2018) in order to ensure that non-bird pages were not erroneously included in our analyses. We obtained page links for Wikipedia pages in our target languages and downloaded pageviews for each page for the period between 01 July 2015—22 June 2019 (n = 1,453 days) using 'pageviews' in R (Keyes & Lewis 2016).

eBird—For each of the regions responsible for more than half of the pageviews in one of our Wikipedia language editions, we downloaded the eBird Basic Dataset (version April 2019) with records for all species and all years (eBird 2019). In most cases we defined a region as a single country, however, in three instances where the distribution of a language corresponds strongly to a specific subnational region, we downloaded data for that region rather than the entire country (Catalan - Catalonia, Basque - Basque Country, Tamil - Tamil Nadu). We limited our analyses to regions with a minimum of 10,000 unique sampling events submitted to eBird. We selected species that appeared in over 10 unique sampling events in the region's eBird dataset. This minimum threshold excluded rare instances where an erroneously listed species had not yet been removed via eBird's review process (Wood et al. 2011). Bird pages in each

Wikipedia language edition were assigned as either "local" if they occurred on the eBird list for the associated region or "global" if they did not (eBird 2019). For local species, we calculated the observation frequency of a species as the total observations of that species in the region divided by the total sampling events in the region (Sullivan et al. 2009). For body size data, we used the bird's body mass per species, obtained from Dunning (Dunning 2008).

Data analysis

We examined four aspects of online interest in birds using our combination of Wikipedia pageviews, eBird frequency, and body size data: a) pageviews for local vs. global species across regions; b) the relationship between eBird frequency and Wikipedia pageviews for local species; c) the relationship between body size and Wikipedia pageviews for local and global species species; and d) the relationship between body size, eBird frequency and Wikipedia pageviews for local species. For all regression models, we checked the model residuals for approximate normality and homoscedasticity. Analyses were undertaken using R (version 3.6.0, R Core Team, 2019).

Interest in local vs. global species—To assess online interest in local vs. global species we compared the mean pageviews per species in both groups using a Wilcoxon rank-sum test. For each language, we calculated the effect size of the difference between mean pageviews for local and global species (effect sizes standardized to a mean of 0 and stdev of 1). We then used a linear model to assess the influence of six aspects of the Wikipedia language edition and the regional eBird dataset on this effect size. Predictors in our model included: a) the total number of articles in the Wikipedia language edition; b) the number of pages for bird species in the Wikipedia language; c) the total pageviews for bird-pages in the Wikipedia language; d) the

diversity of eBird species in the region; e) the number of eBird sampling events in the region; and f) the proportion of the region's eBird species with a Wikipedia page in the corresponding language. Variables were tested for multi-collinearity using a variance inflation factor with a threshold of 10 (Fox & Weisberg 2011). Data on the overall size of Wikipedia languages were obtained from https://meta.wikimedia.org (Wikimedia 2019a). We used manual F-test-based backward selection with a cutoff of 0.05 to identify significant inputs to the model and variation partitioning using partial linear regression to compare the influence of significant explanatory attributes (Oksanen et al., 2017).

Frequency as a predictor of interest—For local species with pageviews in our dataset, we tested the Pearson correlation between eBird frequency (logit transformed) and Wikipedia pageviews (log transformed). Since language editions varied substantially in their average views, we assessed this relationship for each of the 25 languages pairs in our dataset separately. For each language, we modeled the fit of a linear as opposed to a polynomial relationship, both based on the normal distribution, with frequency as a predictor and pageviews as a response. We compared the linear vs. polynomial model for each language using a likelihood-ratio test (Hothorn et al. 2018).

Body size as a predictor of interest—We assessed the influence of body size on the popularity of both local and global species across each of the language in our dataset using a linear model. We compared patterns in this relationship between the two group (local vs. global) using analysis of covariance (Heiberger 2018).

Frequency vs. body size —Finally, we compared eBird frequency and body size as predictors of Wikipedia pageviews for local species. For each language, we used a linear model with both frequency and body size as predictors of pageviews. We used variation partitioning

based on partial linear regression to compare the influence of each of the two predictors in their ability to explain variation in pageviews (Oksanen et al. 2017).

Results

Data selection and extraction

Thirty-nine Wikipedia language editions met our criteria for overall size, depth, and geographic distribution. Of these, 25 had more than 10,000 eBird sampling events in the region responsible for the majority of the language's Wikipedia pageviews. Thus, our final dataset included these 25 language-region pairs (S1 Data). Europe (15 regions) and Asia (nine regions) accounted for all of the regions meeting our criteria except for one (Portuguese - Brazil). Africa, North America, and the Pacific were not represented in our dataset.

Our Wikipedia dataset included 78,415 pages for birds across the 25 languages. In total 10,174 bird species had at least one Wikipedia page in our dataset (96.1% of the total bird species listed in the eBird/Clements global taxonomy; Clements et al. 2018). Numbers of pages (i.e. species) per language varied considerably from 103 (Hindi) to ca. 10,000 (9,821 in Basque and 10,103 in Dutch; mean across all languages 3,137 pages; S2 Data). Pages received a total of 367 million views over the sampling period (views per languages 417,000-70.9 million, mean 14.7 million).

The eBird dataset for regions associated with our Wikipedia languages included 2.3 million unique sampling events with records of 4,340 bird species. Unique sampling events per region ranged from 10,600 (South Korea) to 817,000 (India; mean overall 91,900), and eBird species diversity per region ranged from 313 species (Czech Republic) to 1,716 (Brazil; mean 615 species; S2 Data). The proportion of a region's eBird species with Wikipedia pages in the

associated language ranged between 7.62% of species (Hindi - India) to >98% of species in eleven language-regions pairs (mean overall 82.4%). We obtained body size data for 92.3% of the species with Wikipedia pages in our dataset (9,406 species).

Interest in local vs. global species

Our dataset contained many more pages that we classified as global than local (70,250 global vs. 8,165 local). This is expected given that only species that occurred in the wild in one of our 25 regions could qualify as local, and many local species overlapped between regions. Despite having fewer overall pages, the pages for local species received more views overall than those for global species (218 million vs. 149 million pageviews; Fig. 1), and the mean pageview for local species was higher than the mean pageviews for global species (mean local = 26,700, mean global 2,120). In each language, the mean pageviews for local species was always higher than the mean pageviews for global species. This difference was statistically significant in all but two languages (Wilcoxon rank-sum $p < 0.05$; mean pageviews local 723 – 131,000; mean pageviews global 106-16,600; Fig. 2, Table 1).

For the majority of languages in our dataset, whether a species of was classified as local or global had either large or moderate effect size on the pageviews that species received (large = > 0.5, moderate = 0.3 – 0.5; effect size values normalized to mean of 0 and stdev of 1). Of the variables included in the final model from our stepwise regression model selection, only a) the number of Wikipedia bird pages in a language, and b) the proportion of a region's eBird species with pages in the corresponding language were significant in explaining variation in the effect size of global vs. local in a language. The former had a slight negative impact on effect size and the latter a positive impact (Wikipedia page diversity: β = -2.43E-5, std error = 7.32E-6, p =

3.12E-3; proportion of species coverage: β = 5.72E-5, std error = 8.67E-2, p = 1.22E-6; all variables standardized to a mean of 0 and stdev of 1). Together, these two factors explained more than 60% of the variance in effect size between languages (adj. $R^2 = 0.64$). Of the two variables, the proportion of eBird species with a Wikipedia page explained substantially more variance than the overall number Wikipedia bird pages in a language (partition of variance using partial linear regression: adj. R^2 diversity | proportion = 0.16; adj. R^2 proportion | diversity = 0.67).

Frequency as a predictor of interest

We assessed the relationship between the eBird frequency of a species in a region and its pageviews in the corresponding Wikipedia language for all species that qualified as local in our dataset (8,165 pages and 2,634 species). In all 25 languages, frequency and pageviews were positively correlated ($p < .001$, Pearson's adjusted r^2 range 0.27-0.70, mean = 0.49, stdev = 0.13; Fig. 3). In all languages except for two, the polynomial model [log(pageviews) = logit (eBird frequency) + logit(eBird frequency)^2] provided a better fit to the data than the linear model based on the likelihood ratio test (χ^2 (1) $p < 0.05$). In the remaining two languages (Hindi and Greek), the models were statistically indistinguishable. Across languages, the frequency model described between 9.1% - 50.0% of the variance in pageviews ($p < 0.01$, adj. R^2 = 0.09-0.50, mean = 0.31, stdev = 0.13; Fig. 3, Table 2). As expected based on the fit of the polynomial relationship, the most frequently recorded species in each language received a high percentage of the total pageviews. Amongst local species, the twenty most frequently recorded species in eBird accounted for 21.2 – 54.3% of the pageviews for local bird species, while making up 3.14 – 20.1% of the total local species-pages (mean proportion of views = 40.0%, stdev = 8.24%; mean proportion of species = 13.1% , stdev = 5.08%). This pattern was also evident when considering

all the bird pages in a language (both local and global) where the twenty most commonly-observed local species in a region's eBird dataset accounted 3.16 - 19.6% of the total views for all bird pages in the associated language while only consisting of 0.198 - 4.20% of the species pages (mean proportion of views = 13.35%, stdev = 4.66%; mean proportion of species = 1.29%, stdev = 1.14%; Fig. 4).

Frequency vs. body size

In each language, we assessed the relationship between a bird's body size, as measured by its mass, and its Wikipedia pageviews for both local and global species using a linear regression model (species per language = 99 – 9,387; across all languages 74,823 pages for 9,395 species). Considering local and global species together, body size was a significant explanatory variable for the pageviews a species received in all but four language editions ($p < 0.01$ in 84% of the 25 languages; S3 Data). In all languages where the relationship was significant, body mass had a positive effect on pageviews and described between 2.1% and 24.5% of the variance in pageviews ($p < 0.01$, adj. R^2 = 0.02-0.25, mean = 0.11, stdev = 0.08).

ANCOVA analysis indicated that the interaction between a bird's body mass and whether it was local or global was statistically significant when all languages were grouped together ($F_{1, 74822} = 79$, $p < 0.001$). Body mass had a stronger positive effect for global species than for local ones (estimates national = 0.25, international = 0.34). For individual languages, the interaction between local-global and body size was significant in 12 languages ($p < 0.01$; 48%). With one exception (Portuguese – Brazil), the positive relationship between body mass and pageviews was stronger for international species than for national ones (national species mean coefficient =

0.18, stdev = 0.11, international species mean coefficient = 0.34, stdev = 0.05; for full results see S4 Data, S4 Fig.).

For local species receiving at least one pageview for which we had body mass data (8,013 pages for 2,543 species), a model that included both eBird frequency and body mass was always significant in explaining variance in pageviews ($p < 0.01$). For all languages except one (Portuguese – Brazil), eBird frequency explained more of the variance in pageviews than body mass (partition of variance using partial linear regression: adj. R^2 frequency | mass = 0.07-0.52, mean = 0.28, stdev = 0.13; adj. R^2 mass | frequency = 0.01-0.26, mean = 0.08, stdev = 0.06; Table 3). Often the difference between the two was substantial; in 14 languages eBird frequency explained over three times more variance than body mass.

Discussion

Conservation requires public support and engagement to be effective, thus it is important to understand how the planet's accelerating defaunation may influence public attention and awareness. Our results highlight three patterns regarding the relationship between the distribution and abundance of bird species and the extent of interest they receive online. First, the internet-using public tends to be more interested in birds that occur in their local region than those that occur outside of it (Fig. 1, 2). This pattern is widespread across languages and significant. The difference between a species being local or global had a moderate to large effect size on the mean pageviews per species in the majority of languages in our dataset (Table 1). Furthermore, the effect size was large in all languages where >99% of the local species had a Wikipedia page, suggesting that lower effect sizes in some languages may result from variations in Wikipedia usage rather than in the underlying relationship.

Second, for bird species that occur locally, their abundance in a region (as measured by the frequency they are reported in eBird) is an important predictor of their online interest. Abundance predicted a mean of 28.7% of the variance in Wikipedia pageviews across all languages in our dataset, and in nearly a third of the languages this single variable predicted more than 40% of the variance in Wikipedia pageviews amongst species (Table 2). This relationship is especially important for the most common species overall, which despite making up only a small percentage of the total number of bird pages in a language, receive a disproportionate amount of the total pageviews (Fig. 4). In striking examples, such as Finnish and German, the twenty most frequently encountered species in each country account for 15.2% and 16.3%, respectively, of the pageviews while consisting of only 0.3% and 0.5% of total bird pages.

Third, for local species, a bird's abundance is a significantly better predictor of its online interest than its body size. When both variables were modeled together, abundance absent body size explained a mean of 27.8% of the variation in Wikipedia pageviews in each language, while body size absent abundance explained a mean of 8.0% (Table 3). In half the languages, abundance explained three or more times the amount of variance as body size. This result follows the finding of Correia et al. (2016) for birds in Brazil as does the conclusion that body size is more important for global species than local ones. These patterns demonstrate the importance of geographic scale in determining interest in species. At local scales, where people have the potential to encounter species in the wild, commonness and familiarity are most important. At global scales, where species that are less likely to be directly encountered, body size takes on a more important role for. Given the widely acknowledged role of body size as one of the most significant drivers of interest in species (Small 2012; Albert et al. 2018), abundance

is undoubtedly among the most important, if not the most important, features in determining online interest amongst locally occurring bird species.

On the surface, our conclusions seem to contradict the established wisdom that large-bodied, often rarely encountered in the wild, species are the most important attractors of public interest. Part of the reason for this methodological. We examined broad-scale patterns of interest across large numbers of species, while many previous studies of species popularity have considered smaller subsets of highly appealing species (e.g. Albert et al. 2018). While the mean views for global species was lower than for local species in our dataset, global species often contained extreme outliers with a few species featuring amongst the most viewed pages in each language (see box plot outliers in Fig. 2). The importance of these global outliers is emphasized when aggregating views across languages. By having appeal outside of their area of distribution, these species feature in prominently in many languages and thus rank higher when pageviews from multiple languages are added together. For example, if views from all languages in our dataset are summed by species, the Dodo, a large-bodied bird and traditional flagship (and one that for obvious reasons people do not encounter in the wild), receives the most pageviews. Also, Wikipedia pageviews reflect patterns of people's attention and their searches for additional knowledge, but not necessarily their preferences. Even if someone frequently looks for information about a common bird online, they may be unlikely to describe it as their favorite species. A final reason for this difference may have to do with our choice of taxonomic group. Birds generate unique forms of human-nature interaction through popular activities such as birdwatching and bird-feeding and they are more easily observed and identified in the wild than many other types of organisms (Cocker et al. 2013). As a result of the specific modes of

interaction that birds enable, direct encounters in the wild could be more significant in determining interest in birds than for other taxonomic groups.

Our method did not assess the mechanisms underlying the high online interest in local, common birds. A variety of potential mechanisms can by hypothesized—people may be more likely to travel within their own country or region, leading to increased awareness of species that occur within those borders, for example—and testing these is of interest for future research. Importantly, however, the mechanisms underlying these patterns may vary culturally and geographically. As is often the case with 'big data' analyses, identifying causation at very large analytical scales can be challenging or even impossible. At these big data scales, however, patterns themselves are frequently sufficient to provide important insights (Kitchin 2014).

The prevalence of local distribution and abundance in predicting online interest in bird species suggests two conclusions with significant relevance to conservation. First, these patterns help to understand why some species command greater public interest than others, a question that has long been of interest to conservationists (Kellert 1982; Lorimer 2007; Ducarme et al. 2013). Our results show that in addition to physical features such as bright colors or perceived attractiveness (Gunnthorsdottir 2001; Stokes 2007; Frynta et al. 2010), the distribution of species and the opportunity to encounter them in the wild are important in determining public attention. Among other things, this relationship may help to explain why preferences for some species vary regionally. This result concurs with previous studies (Zmihorski et al. 2013; Correia et al. 2016), but we demonstrate its significance across a broad range of cultural and geographic scales.

Second, our results highlight the role of common species as a mechanism for engaging public interest in the natural world. Few if any of the twenty most common birds in the regions in our dataset would qualify as 'charismatic megafauna' under most criteria, and yet they

account for between 10% and 20% of the total bird pageviews in most languages. These common birds are undoubtedly important entry points for people's interactions with biodiversity and likely play a role in spurring general interest in nature. Since interest levels for many species correlate with how frequently people encounter them in the wild, the declining populations of these species has the potential to result in lower levels of interest. Thus, in addition to an 'extinction of experience' for people, there may be an 'extinction of interest' with increased apathy towards specific species of wildlife as encounters with them decline. Large scale measures such as habitat protection and reducing agricultural intensification are needed maintain healthy populations of common species and combat defaunation (Young et al. 2016; Rosenberg et al. 2019). Our results show that facilitating opportunities for people to interact with common species in the wild also needs to be an important priority in these efforts.

References

Albert C, Luque GM, Courchamp F. 2018. The twenty most charismatic species. PLoS ONE **13**:e0199149.

Alexa. 2019. The top 500 sites on the web. Available from https://www.alexa.com/topsites (accessed June 1, 2019).

Bebbington A. 2005. The ability of A-level students to name plants. Journal of Biological Education **39**:63–67.

Callaghan CT, Gawlik DE. 2015. Efficacy of eBird data as an aid in conservation planning and monitoring. Journal of Field Ornithology **86**:298–304.

Ceballos G, Ehrlich PR, Dirzo R. 2017. Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. Proceedings of the National Academy of Sciences of the United States of America **114**:E6089–E6096.

Clark CJ. 2017. eBird records show substantial growth of the Allen's Hummingbird (Selasphorus sasin sedentarius) population in urban Southern California. The Condor **119**:122–130.

Clements JF, Schulenberg TS, Iliff MJ, Roberson D, Fredericks TA, Sullivan BL, Wood CL. 2018. The eBird/Clements checklist of birds of the world: v2018. Available from http://www.birds.cornell.edu/clementschecklist/download/ (accessed April 29, 2019).

Cocker M, Tipling D, Elphick J, Fanshawe J. 2013. Birds and people. Jonathan Cape, London, UK.

Colléony A, Clayton S, Couvet D, Saint Jalme M, Prévot AC. 2017. Human preferences for species conservation: animal charisma trumps endangered status. Biological Conservation **206**:263–269.

Correia RA, Jepson PR, Malhado ACM, Ladle RJ. 2016. Familiarity breeds content: assessing

bird species popularity with culturomics. PeerJ **4**:1–15.

Dirzo R, Young HS, Galetti M, Ceballos G, Isaac NJB, Collen B. 2014. Defaunation in the Anthropocene. Science **345**:401–406.

Ducarme F, Luque GM, Courchamp F. 2013. What are "charismatic species" for conservation biologists? Biosciences Master Reviews:1–8.

Dunning JB. 2008. CRC Handbook of Avian Body Masses, second edition. CRC Press, Boca Raton, FL.

eBird. 2019. eBird Basic Dataset. Version: EBD_relApril-2019. Cornell Lab of Ornithology, Ithaca, NY.

Fink D, Damoulas T, Dave J. 2013. Adaptive Spatio-Temporal Exploratory Models: Hemisphere-Wide Species Distributions from Massively Crowdsourced eBird Data. Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence:1284–1290.

Fox J, Weisberg S. 2011. An {R} Companion to Applied Regression, second edition. Thousand Oaks, CA.

Frynta D, Lisková S, Bültmann S, Burda H. 2010. Being attractive brings advantages: the case of parrot species in captivity. PloS ONE **5**:e12568.

Gaston KJ, Fuller RA. 2008. Commonness, population depletion and conservation biology. Trends in Ecology and Evolution **23**:14–19.

Generous N, Fairchild G, Deshpande A, Valle SY Del, Priedhorsky R. 2014. Global disease monitoring and forecasting with Wikipedia. PLoS Computational Biology **10**:e1003892.

Gunnthorsdottir A. 2001. Physical attractiveness of an animal species as a decision factor for its preservation. Anthrozoos **14**:204–214.

Hausmann A, Toivonen T, Slotow R, Tenkanen H, Moilanen A, Heikinheimo V, Di Minin E. 2018. Social Media Data Can Be Used to Understand Tourists' Preferences for Nature-Based Experiences in Protected Areas. Conservation Letters **11**:1–10.

Heiberger RM. 2018. HH: Statistical analysis and data display: Heiberger and Holland. Available from https://cran.r-project.org/package=HH.

Hothorn T, Zeileis A, Farebrother RW, Cummins C, Millo G, Mitchell D. 2018. lmtest: Testing Linear Regression Models. Available from https://cran.r-project.org/package=lmtest.

Inger R, Gregory R, Duffy JP, Stott I, Voříšek P, Gaston KJ. 2015. Common European birds are declining rapidly while less abundant species' numbers are rising. Ecology Letters **18**:28–36.

Kellert SR. 1982. Factors in Endangered Social and Perceptual Species Management. The Journal of Wildlife Management **49**:528–536.

Keyes O, Lewis J. 2016. pageviews: an API client for wikimedia traffic data. Available from https://cran.r-project.org/package=pageviews.

Kitchin R. 2014. Big Data, new epistemologies and paradigm shifts. Big Data & Society **1**:1–12.

Ladle RJ, Correia RA, Do Y, Joo GJ, Malhado ACM, Proulx R, Roberge JM, Jepson P. 2016. Conservation culturomics. Frontiers in Ecology and the Environment **14**:269–275.

Leader-Williams N, Dublin HT. 2000. Charismatic megafauna as "flagship species." Pages 53–81 in A. Entwistle and N. Dunstone, editors. Has the panda had its day? Priorities for the conservation of mammalian diversity. Cambridge University Press, Cambridge, UK.

Lindemann-Matthies P. 2005. "Loveable" mammals and "lifeless" plants: How children's interest in common local organisms can be enhanced through observation of nature. International Journal of Science Education **27**:655–677.

Lorimer J. 2007. Nonhuman charisma. Environment and Planning D: Society and Space **25**:911–932.

McIver DJ, Brownstein JS. 2014. Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. PLoS Computational Biology **10**:e1003581.

Mestyán M, Yasseri T, Kertész J. 2013. Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. PLoS ONE **8**:e71226.

Miller JR. 2005. Biodiversity conservation and the extinction of experience. Trends in Ecology and Evolution **20**:430–434.

Mittermeier JC, Roll U, Matthews TJ, Grenyer R. 2019. A season for all things: phenological imprints in Wikipedia usage and their relevance to conservation. PLoS Biology **17**:e3000146.

Moat HS, Curme C, Avakian A, Kenett DY, Stanley HE, Preis T. 2013. Quantifying Wikipedia Usage Patterns Before Stock Market Moves. Scientific Reports **3**:1–5.

Oksanen J et al. 2017. vegan: Community Ecology Package. Available from https://cran.r-project.org/package=vegan.

Oliver TH et al. 2015. Biodiversity and Resilience of Ecosystem Functions. Trends in Ecology and Evolution **30**:673–684.

Paul MJ, Dredze M. 2011. You are what you tweet: analyzing Twitter for public health. ICWSM **20**:265–272.

Pilgrim SE, Cullen LC, Smith DJ, Pretty J. 2008. Ecological knowledge is lost in wealthier communities and countries. Environmental Science and Technology **42**:1004–1009.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.r-project.org/.

Roll U, Mittermeier JC, Diaz GI, Novosolov M, Feldman A, Itescu Y, Meiri S, Grenyer R. 2016. Using Wikipedia page views to explore the cultural importance of global reptiles. Biological Conservation **204**:42–50.

Rosenberg K V et al. 2019. Decline of the North American avifauna **1313**:1–10.

Skiena S, Ward C. 2014. Who's Bigger? Where historical figures really rank. Cambridge University Press, New York, New York, USA.

Small E. 2012. The new Noah's Ark: Beautiful and useful species only. Part 2. The chosen species. Biodiversity **13**:37–53.

Soga M, Gaston KJ. 2016. Extinction of experience: The loss of human-nature interactions. Frontiers in Ecology and the Environment **14**:94–101.

Soriano-Redondo A, Bearhop S, Lock L, Votier SC, Hilton GM. 2017. Internet-based monitoring of public perception of conservation. Biological Conservation **206**:304–309.

Stokes DL. 2007. Things we like: human preferences among similar organisms and implications for conservation. Human Ecology **35**:361–369.

Sullivan BL, Wood CL, Iliff MJ, Bonney RE, Fink D, Kelling S. 2009. eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation **142**:2282–2292.

Veríssimo D, Vaughan G, Ridout M, Waterman C, MacMillan D, Smith RJ. 2017. Increased conservation marketing effort has major fundraising benefits for even the least popular species. Biological Conservation **211**.

Walker J, Taylor PD. 2017. Using eBird data to model population change of migratory bird species. Avian Conservation and Ecology **12**:4.

Wells N, Lekies K. 2006. Nature and the Life Course: Pathways from Childhood Nature

Experiences to Adult Environmentalism. Children Youth and Environments **16**:1–24.

Wikimedia. 2019a. List of Wikipedias. Available from https://meta.wikimedia.org/wiki/List_of_Wikipedias (accessed June 22, 2019).

Wikimedia. 2019b. Wikipedia article depth. Available from https://meta.wikimedia.org/wiki/Wikipedia_article_depth (accessed June 23, 2019).

Wood C, Sullivan B, Iliff M, Fink D, Kelling S. 2011. eBird: Engaging Birders in Science and Conservation. PLoS Biology **9**:e1001220.

Young HS, McCauley DJ, Galetti M, Dirzo R. 2016. Patterns, Causes, and Consequences of Anthropocene Defaunation. Annual Review of Ecology, Evolution, and Systematics **47**:333–358.

Yu AZ, Ronen S, Hu KZ, Hidalgo C a. 2016. Pantheon 1.0, a manually verified dataset of globally famous biographies. Scientific Data **3**:150075.

Zachte E. 2018. Wikimedia Traffic Analysis Report: page views per wikipedia language. Available from https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm (accessed July 23, 2018).

Zmihorski M, Dziarska-Palac J, Sparks TH, Tryjanowski P. 2013. Ecological correlates of the popularity of birds and butterflies in Internet information resources. Oikos **122**:183–190.

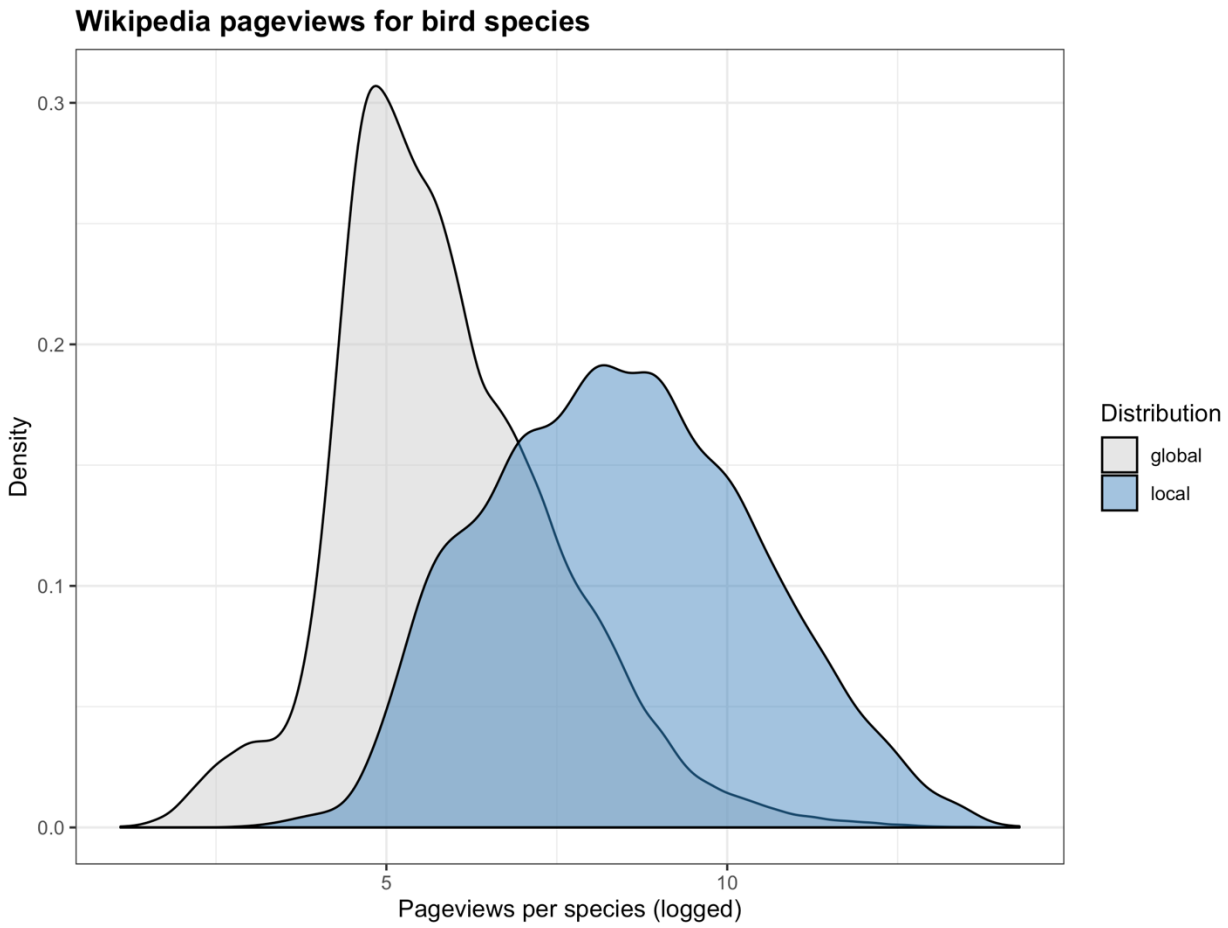


FIGURE 1. Based on 367 million Wikipedia pageviews for 78,415 pages for birds across 25 different languages, bird species whose wild distribution overlaps with the region responsible for the majority of a language edition's pageviews ("local species") tend to receive more views than species whose geographic distribution does not overlap with that region ("global species").

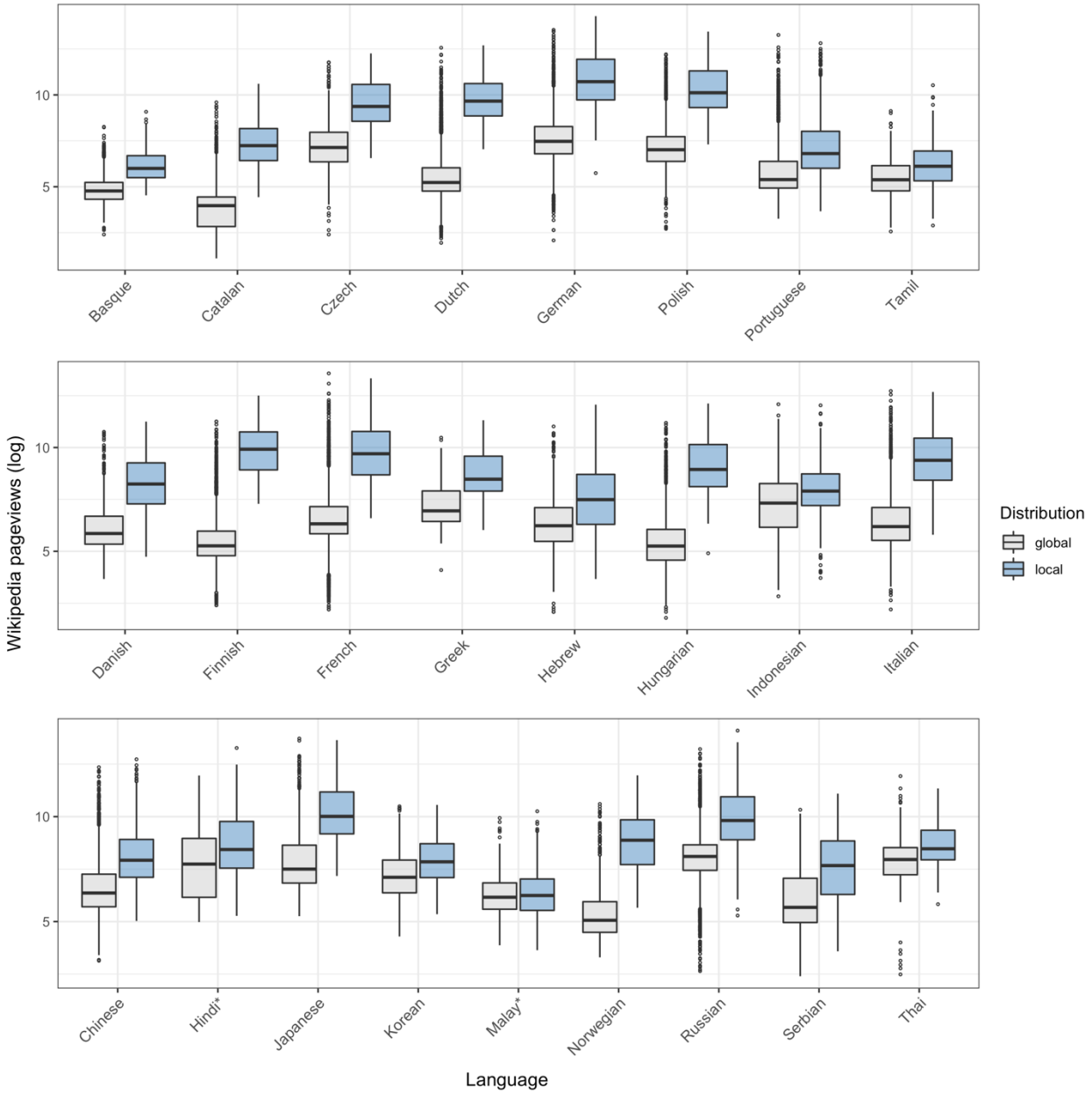


FIGURE 2. Species that occur in the wild in the region responsible for the majority of pageviews in a given Wikipedia language ("local species") tend to receive more views than those that do not occur within the region ("global species"). This pattern appears across multiple languages and regions of the world, and the difference between pageviews for local and global species is

statistically significant in 23 of 25 languages we tested. Two languages where the difference was not statistically significant (Malay and Hindi) are marked with an asterisk. While the mean pageviews for global species was always lower than for local species, global species also frequently contained extreme outliers (points in box plots indicate outliers occurring > 1.5 times interquartile range beyond the median).

Language	Region	Mean local	Mean global	Effect	p-value
Danish	Denmark	8770	1340	0.58	2.95E-81
Norwegian	Norway	15000	613	0.55	1.63E-142
Japanese	Japan	74700	12100	0.53	1.75E-115
Czech	Czech Republic	28700	4380	0.52	3.31E-73
Polish	Poland	66200	3950	0.46	2.65E-129
German	Germany	131000	6930	0.42	2.02E-164
Italian	Italian	34300	2380	0.42	2.77E-151
Russian	Russia	64400	7800	0.41	5.32E-124
Chinese	Taiwan	11800	3130	0.4	3.38E-97
Catalan	Catalonia	3670	106	0.39	1.03E-190
Serbian	Serbia	5360	1510	0.38	3.98E-19
Greek	Greece	12400	3270	0.38	1.45E-16
Portuguese	Brazil	7660	2000	0.37	6.60E-177
Finnish	Finland	36300	693	0.35	6.88E-165
French	France	50300	2380	0.34	4.13E-211
Hungarian	Hungary	20000	681	0.34	1.07E-149
Dutch	Netherlands	35100	856	0.29	3.68E-187
Hebrew	Israel	6260	1620	0.29	2.26E-27
Korean	South Korea	5350	3150	0.24	9.01E-10
Basque	Basque Country	723	156	0.23	1.10E-118
Tamil	Tamil Nadu	1140	511	0.23	1.71E-09
Indonesian	Indonesia	6540	5950	0.16	3.16E-12
Thai	Thailand	10800	7350	0.16	2.30E-07
Hindi*	India	23100	16600	0.05	0.0902
Malay*	Malaysia	1330	1170	0.02	0.526

TABLE 1. Mean Wikipedia pageview for birds that occur in the wild in a region (“local”) as opposed to those that do not (“global”) across 25 languages and regions. The mean views for local species was always higher but in two languages this difference was not statistically significant ($p > 0.05$; non-significant languages indicated with an asterisk). Effect size of the difference between the means is listed for each language.

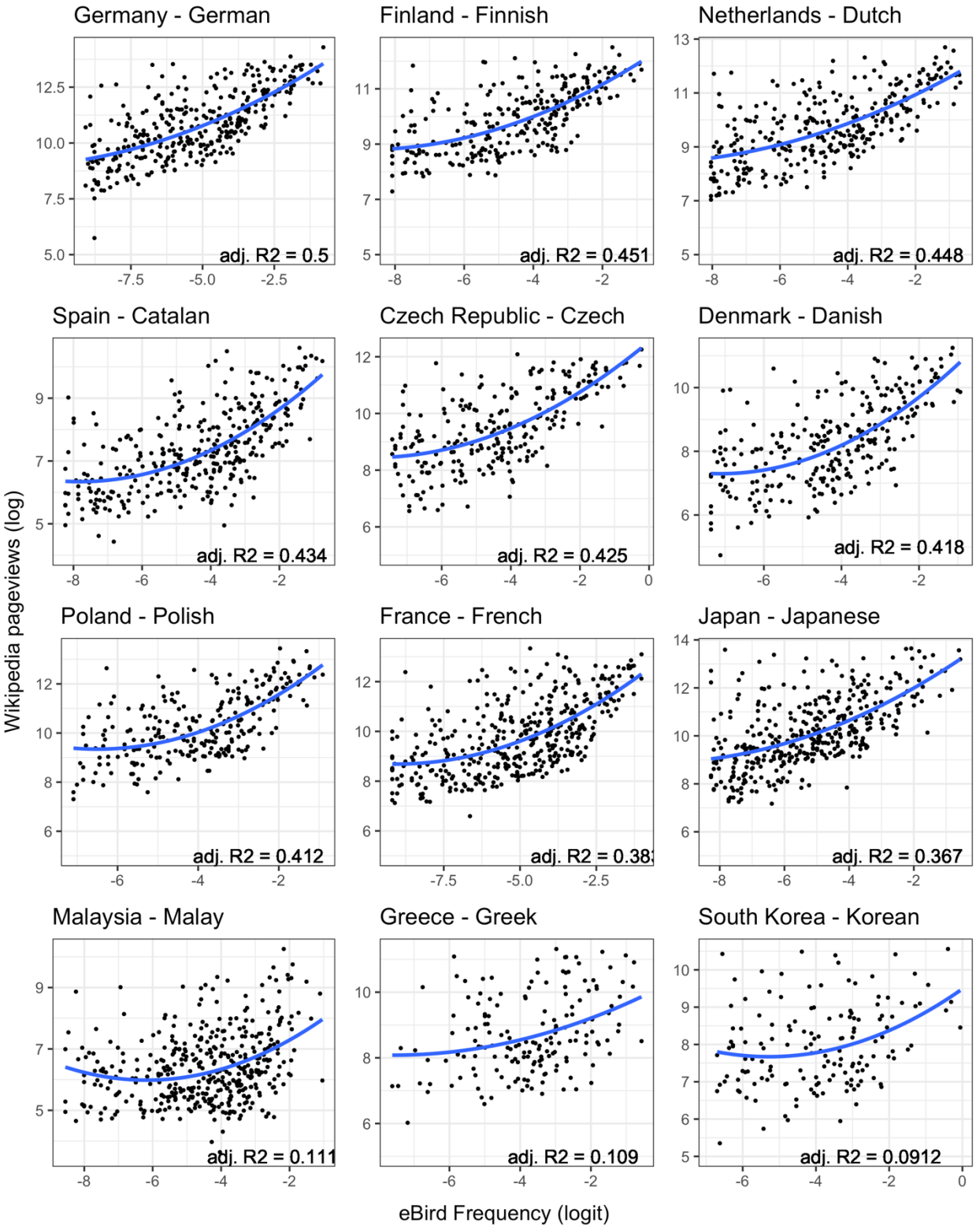

Germany - German
adj. R2 = 0.5
Finland - Finnish
adj. R2 = 0.451
Netherlands - Dutch
adj. R2 = 0.448
Spain - Catalan
adj. R2 = 0.434
Czech Republic - Czech
adj. R2 = 0.425
Denmark - Danish
adj. R2 = 0.418
Poland - Polish
adj. R2 = 0.412
France - French
adj. R2 = 0.38
Japan - Japanese
adj. R2 = 0.367
Malaysia - Malay
adj. R2 = 0.111
Greece - Greek
adj. R2 = 0.109
South Korea - Korean
adj. R2 = 0.0912
Wikipedia pageviews (log)
eBird Frequency (logit)

FIGURE 3. For multiple countries, the frequency with which a bird species is observed in a region shows a positive relationship with the number of pageviews that a species receives in the associated Wikipedia language. The top three rows show the nine regions in which eBird frequency accounted for the greatest proportion of variance in pageviews in our dataset (measured in terms of adj. R-squared). The bottom row shows the three languages in which frequency accounted for the smallest proportion of variance. Mean adj. R-squared across all 25 languages in our dataset was 0.31. Full results are listed in Table 2.

Language	Term	Estimate	Std. Error	Pr(>|t|)	Adj. R^2	p-value
Basque	x	0.706	0.114	1.80E-09	0.288	1.99E-22
	x^2	0.0549	0.114	3.65E-05		
Catalan	x	1.07	0.137	9.80E-14	0.434	2.78E-41
	x^2	0.0682	0.137	2.62E-06		
Chinese	x	0.754	0.113	7.70E-11	0.193	4.00E-23
	x^2	0.0444	0.113	3.03E-06		
Czech	x	1.02	0.17	8.48E-09	0.425	1.47E-29
	x^2	0.063	0.17	0.00142		
Danish	x	1.31	0.194	9.91E-11	0.418	6.37E-32
	x^2	0.0919	0.194	2.86E-05		
Dutch	x	0.752	0.128	1.00E-08	0.448	8.95E-41
	x^2	0.0361	0.128	0.00769		
Finnish	x	0.877	0.143	3.40E-09	0.451	5.70E-36
	x^2	0.0489	0.143	0.00102		
French	x	1.01	0.135	3.60E-13	0.383	7.71E-43
	x^2	0.0561	0.135	7.47E-06		
German	x	0.862	0.136	8.31E-10	0.500	3.92E-51
	x^2	0.0349	0.136	0.00643		
Greek	x	0.548	0.246	0.0274	0.109	2.63E-05
	x^2	0.0361	0.246	0.233		
Hebrew	x	1.23	0.215	2.58E-08	0.207	2.52E-15
	x^2	0.098	0.215	3.80E-05		
Hindi	x	0.194	0.266	0.468	0.117	0.00277
	x^2	-0.00311	0.266	0.897		
Hungarian	x	1.65	0.218	7.24E-13	0.355	1.02E-25
	x^2	0.152	0.218	6.12E-08		
Indonesian	x	1.07	0.22	1.46E-06	0.152	1.71E-18
	x^2	0.0764	0.22	6.67E-04		
Italian	x	1.15	0.183	1.20E-09	0.289	1.39E-23
	x^2	0.0814	0.183	3.99E-05		
Japanese	x	0.96	0.166	1.40E-08	0.367	5.45E-40
	x^2	0.0476	0.166	0.00364		
Korean	x	0.688	0.239	0.00452	0.0912	3.61E-04
	x^2	0.0651	0.239	0.0328		
Malay	x	0.926	0.172	1.22E-07	0.111	1.82E-11
	x^2	0.075	0.172	1.64E-05		
Norwegian	x	0.873	0.196	1.26E-05	0.353	6.29E-28
	x^2	0.0427	0.196	0.0226		
Polish	x	1.43	0.216	2.18E-10	0.412	1.79E-30
	x^2	0.11	0.216	3.11E-05		
Portuguese	x	1.6	0.114	2.15E-41	0.264	9.78E-77
	x^2	0.11	0.114	2.15E-26		
Russian	x	1.9	0.248	9.13E-14	0.258	5.01E-30
	x^2	0.159	0.248	1.43E-08		
Serbian	x	1.16	0.235	2.29E-06	0.205	1.11E-08
	x^2	0.0992	0.235	2.18E-04		
Tamil	x	0.721	0.135	1.83E-07	0.196	1.14E-13
	x^2	0.0503	0.135	2.88E-04		
Thai	x	0.715	0.197	3.77E-04	0.137	3.41E-06
	x^2	0.0561	0.197	0.00616		

TABLE 2. Model outputs for a polynomial linear regression with eBird frequency as predictor of Wikipedia pageviews by language. Outputs for both terms x and x^2 are shown for each language. The model was significant for all 25 languages editions ($p < 0.05$) and with adjusted R-squared values from 0.09 – 0.50 (mean = 0.31, stdev = 0.13).

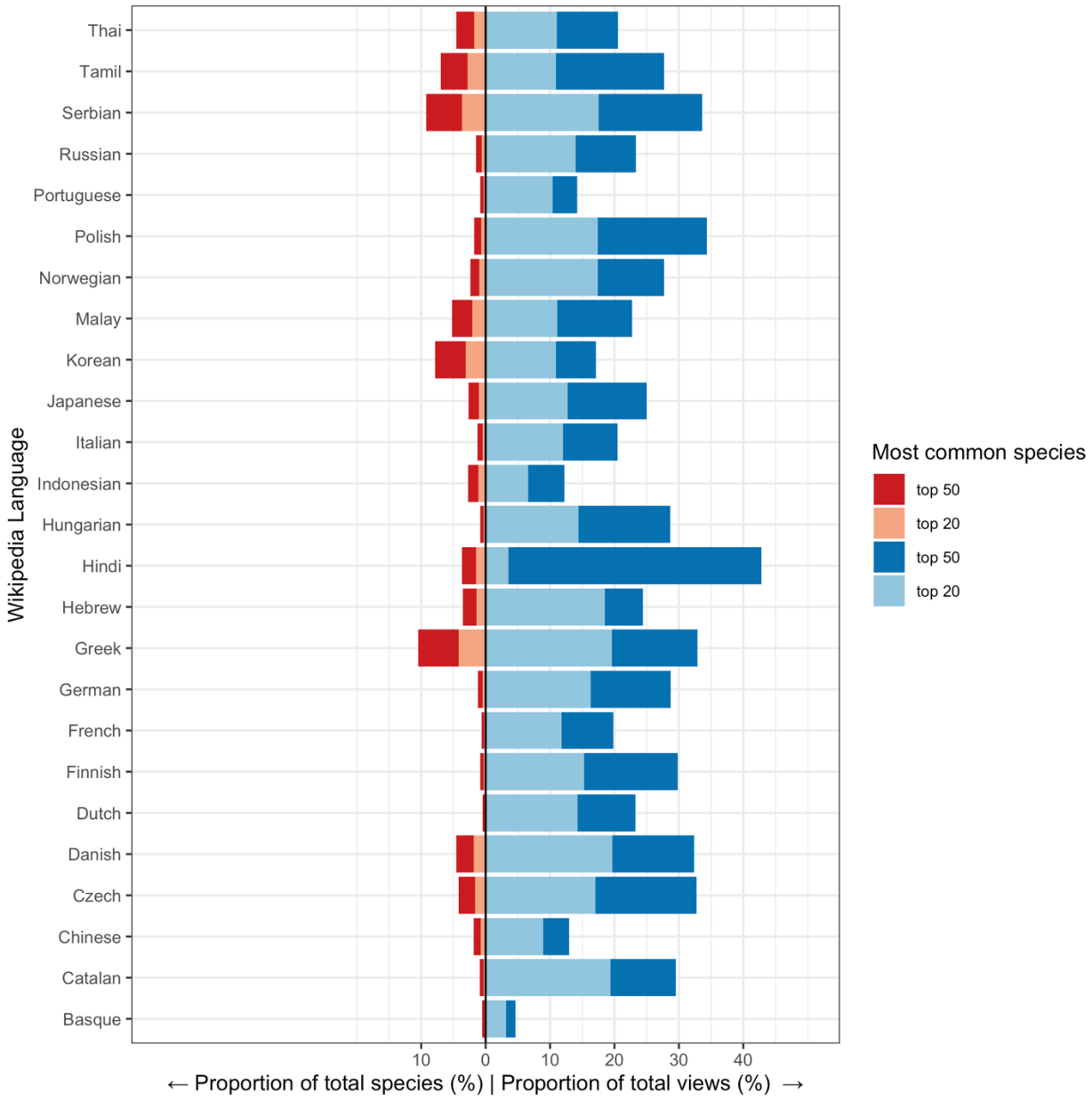


FIGURE 5. Common species account for a disproportionate amount of online interest in bird species. Across 25 languages, the 20 most commonly encountered birds account for 3.16 - 19.6% of the total pageviews in a language while only making of up 0.198 - 4.20% of the pages (lighter shades of color). The 50 most commonly encountered birds make up 4.61 – 42.8% of the pageviews with 0.494 – 10.5% of the pages (darker shades).

Language	Adj. R^2 both	Adj. R^2 freq.	Adj. R^2 mass
German	0.550	0.518	0.0611
Dutch	0.456	0.441	0.0186
Czech	0.453	0.447	0.0549
Catalan	0.428	0.424	0.0376
Polish	0.409	0.398	0.0505
Finnish	0.508	0.424	0.0808
Japanese	0.378	0.359	0.0267
Danish	0.471	0.385	0.0987
French	0.465	0.393	0.117
Norwegian	0.388	0.327	0.052
Hungarian	0.32	0.293	0.0507
Basque	0.294	0.265	0.0466
Italian	0.357	0.303	0.112
Hebrew	0.220	0.199	0.0648
Indonesian	0.127	0.129	0.00215
Hindi	0.118	0.118	-0.00926
Russian	0.319	0.227	0.119
Greek	0.184	0.154	0.0828
Serbian	0.244	0.174	0.114
Thai	0.246	0.203	0.144
Chinese	0.328	0.213	0.168
Tamil	0.325	0.216	0.173
Malay	0.128	0.0931	0.0575
Korean	0.105	0.0726	0.0398
Portuguese	0.441	0.177	0.256

TABLE 3. Variation partitioning based on partial linear regression for a model with both eBird frequency and mass as predictors of Wikipedia pageviews by language. Adjusted R-squared both shows the variation explained with both variables; adjusted R-squared frequency is variation explained by frequency absent mass; and adjusted R-squared mass is mass absent frequency.

Supplemental figures and data.

S1 DATA. Wikipedia language editions and eBird regions selected for our analysis. Total article counts and language edition depth (as of June 2019) for each Wikipedia. Proportion of views originating from the associated country for each Wikipedia, and total unique submission in eBird (as of April 2019) for the associated country/state.

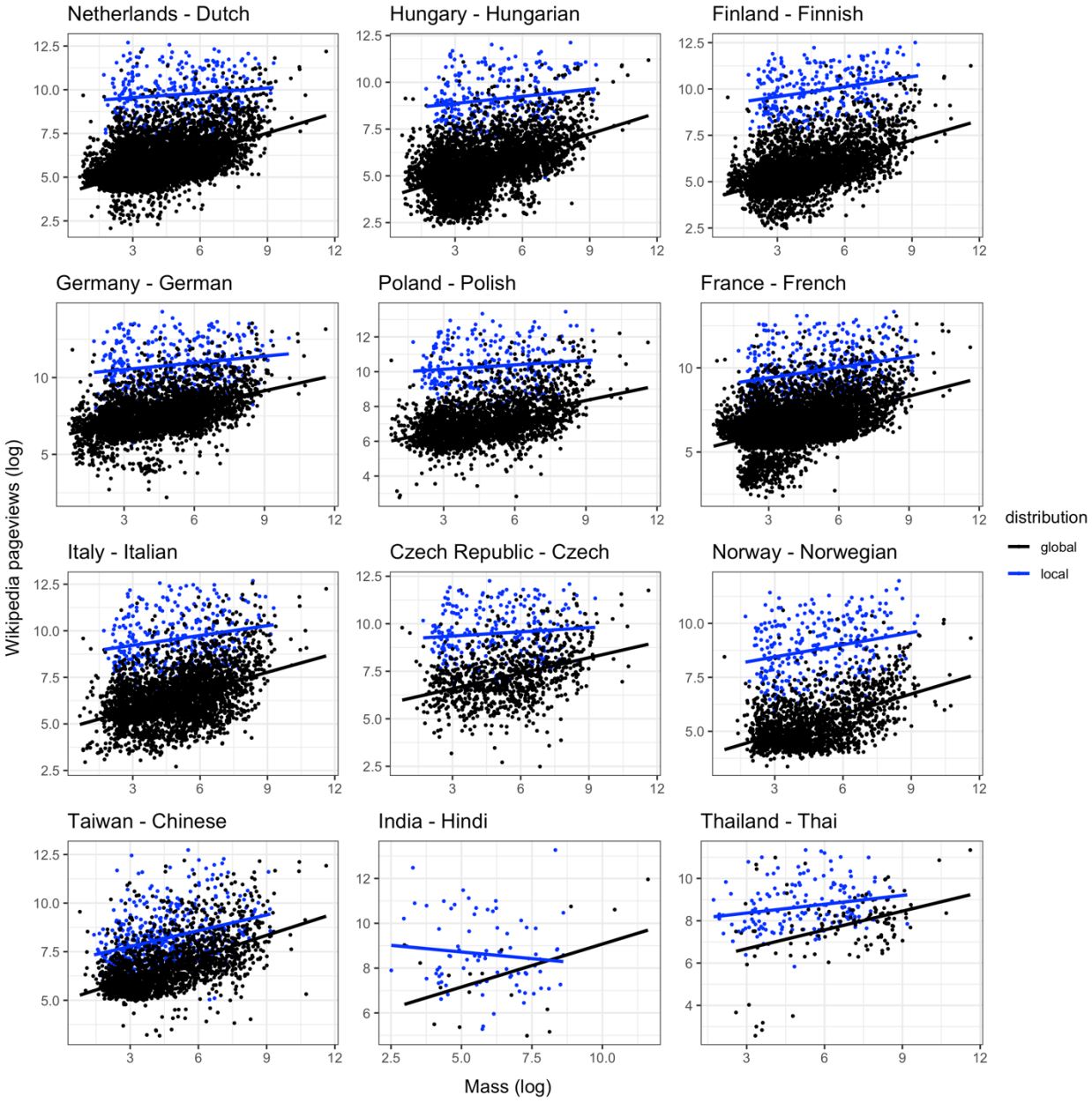
S2 DATA. Pages and pageviews for bird species for each Wikipedia edition and model inputs for comparing predictors of the effect size of the difference in means between local and global species. Predictors for each language are a) total Wikipedia articles; b) number of pages for species; c) number of pageviews for bird species; d) eBird species diversity in the associated region; e) number of eBird sampling events in the associated region; f) proportion of eBird species with a Wikipedia page.

S3 DATA. Linear model results for body mass as a predictor of pageviews across both local and global species. For each language, model estimate, standard error, p-value, and adjusted R-squared is listed.

S4 DATA. ANCOVA results for body mass as a predictor of pageviews for local vs. global species. For each language, interaction p-value is listed as well as coefficient for and p-values for both local and global species.

S5 DATA. Model outputs from linear regression models that include both body mass and eBird frequency as predictors of Wikipedia pageviews across languages. For each language, the

coefficient, standard error and p-value, for both the frequency and mass term is listed. Also provided is the adjusted R-Squared and p-value for the model containing both predictors.



S4 FIGURE. Body mass as a predictor of Wikipedia pageviews for local and global species across languages. Birds with higher body mass tend to receive more pageviews across multiple Wikipedia languages, as indicated by the positive relationship between the two variables. The

strength of this relationship differs between local and global species in the language, however, with body mass frequently having a stronger positive correlation with pageviews for global species. Twelve of 14 languages that had a significant interaction ($p < 0.05$) are displayed. The top three rows are those with the most significant interactions, the bottom row is those least significant interactions that still met are threshold. For full results see table in S4 Data.

Chapter 3: Applications

My thesis highlights four ways in which online big data can be applied to species conservation. First, online big data can provide insight into why certain species attract more attention than others (this comes up as a theme in Patterns 1, 2, and 3). Second, online data offer the opportunity to systematically identify high interest species (Methods 1). Third, they can be used to monitor temporal and spatial patterns in people's interests, and in doing so identify which constituencies are most likely to support conservation initiatives and when (Patterns 2 and 3). Fourth, and finally, online big data can potentially supplement traditional biodiversity survey methods by tracking the geographic movements and abundance of species. This last point has yet to appear in any of the previous papers; I will touch on it here.

The purpose of this chapter is two-fold. First, Applications 1 investigates the last of these four applications of online big data to species conservation (whether these data can be used to monitor the abundance and distribution of species). As becomes clear from the paper, this application is more equivocal than the previous three. Second, Applications 2, provides an example of how the approach for identifying high interest species described in Methods 1 can be used in conservation policy analyses. I served as a collaborator on Applications 2, rather than lead author, and so I include only the abstract and the methods sections of the paper. The abstract provides a description of the paper's focus and conclusions, while the methods is part of the paper most relevant to my contribution in the context of this thesis.

Applications 1

In other research fields, online data have been used to monitor and track events in the real world. Perhaps the best-known example of this is "Google Flu" where people's Google search behaviour was found to be an effective approach for detecting influenza outbreaks (Ginsberg et al. 2009; see further discussions in Lazer et al. 2014; Kandula & Shaman 2019). As I demonstrated in Patterns 2 and Patterns 3, online data can respond to the physical presence of species in a landscape. The purpose of this paper, a brief correspondence, is to consider whether online data can be used to track the migratory movements of species.

This paper was initially submitted as a brief correspondence to the journal *Current Biology* and follows the strict formatting guidelines required for such a submission (maximum 1,000 words and ten references). After review, the paper was rejected with two criticisms. First, that it needed a more detailed explanation of the methods and was not appropriate for such a brief format. And second, the reviewers wanted to know how prevalent these patterns were amongst species. Does online data track migratory movements for only a few species or is it a widespread pattern? Both these comments are well taken. For purposes of the thesis, however, I include the original submission as an example of an application of online big data to species conservation. To provide some additional context, I briefly mention a few results that have emerged from beginning to address the reviewers' comments.

To increase the scale of the paper, I downloaded a sample of 1,803 Wikipedia pages for 614 migratory bird species across ten languages and compared the similarity between the pageview time-series and the daily occurrence of these species in each region. For 45% of the pages, pageviews performed better than a null model in their ability to predict species' daily occurrence. Pageviews for migratory birds were more likely to track seasonal movements in languages spoken

at higher latitudes than in more equatorial languages (79.7% of the sample pages in Finland, but only 19.0% in Brazil). Interestingly, pageviews did not respond to the movements of many highly migratory species, and how pronounced a bird's annual fluctuations in abundance were in a region did not predict whether its pageviews tracked its migratory movements.

These results will be further expanded and resubmitted to another journal. By including a brief description of them here, I hope to emphasize that—as per the reviewers' comments—the following paper fails to sufficiently clarify the extent of this pattern across species. For many species it will not work at all. For a selected subset of species, however, such as the Barn Swallows in Fig.1 of this paper, it might work, but even for these it requires careful calibration (see for example Fig. S2 in the paper).

Submission status: Submitted to Current Biology 3 April 2019. Mittermeier, J.C., Roll, U., Matthews, T., and R. Grenyer. *The digital imprints of bird migration.* Rejected following review.

Personal contribution*:* Lead author. Uri Roll and I devised with initial concept of the paper, I downloaded and curated the data, conducted the formal analysis, devised the methodology, and made the figures. Uri and I wrote first draft of the text together. All co-authors assisted with editing the final manuscript.

Correspondence: The digital imprints of bird migration

John C. Mittermeier[1†], Uri Roll[2], Richard Grenyer[1]

[1] School of Geography and the Environment, University of Oxford, South Parks Road, Oxford, OX1 3QY, UK

[2] Mitrani Department of Desert Ecology, The Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Midreshet Ben-Gurion 8499000, Israel

† corresponding author and lead contact: john.mittermeier@gmail.com

Type of article: correspondence

Keywords: Big Data, Conservation, Culturomics, eBird, Google Trends, Citizen Science, Wikipedia, Migration, Birds

Studying temporal fluctuations in online interest has an array of real-world applications. Wikipedia usage, for example, has been used to monitor influenza outbreaks [1], while Twitter data can reflect broad-scale patterns in people's happiness and well-being [2], and Google queries predict the behavior of financial markets [3]. Compared to traditional methods, these digital Big Data are often inexpensive, easily accessed and highly scalable. In the field of biodiversity conservation, effective policy development frequently requires large volumes of ecological and social data. Detailed knowledge of the populations and distributions of species as well as an understanding of public attitudes and responses to conservation actions are the foundation of good conservation practice. The potential to apply digital Big Data to the field of biodiversity conservation has recently started to gain traction with the emergence of "conservation culturomics" [4]. Here we present a novel example of how freely available and easily accessed digital Big Data can be applied to biodiversity conservation. Using a selection of migratory birds as a case study, we find instances where people's online behavior closely tracks the movements of a species through time and space. In some cases, this online interest and a bird's migratory behavior are so closely linked that it is possible to use Wikipedia pageviews and Google Trends to trace the migration birds across the landscape.

Bird migrations have fascinated people for millennia, and these annual movements often have strong cultural associations and significance. To explore how this interest has translated into the digital realm, we obtained time-series of search frequency in Google as well as daily pageviews to a species' Wikipedia page for a selection of migratory birds. We filtered Google search data to the country-level (via Google Trends) and used language as a proxy for geography in Wikipedia by assigning each language edition to the country that accounts for the greatest proportion of views in that language [5; see Supplemental Information for details]. We then

compared the temporal patterns in these online data to ecological data in the form of ringing records (from the United States Geological Survey Bird Banding Laboratory and the Yamashina Institute for Ornithology) and citizen science records submitted to the Cornell University Lab of Ornithology's eBird project (www.ebird.org [6]).

In multiple instances, we found strong spatial and temporal correspondence between the four data sources we compared demonstrating that the phenology of avian migration is reflected across a range of social communities (e.g. trained bird ringers, bird-watchers logging observational data, and the general public searching for information; Fig 1a) and migratory behaviors (e.g. summering as opposed to wintering birds, early versus late spring migrants, and austral versus boreal migrants; Fig 1b-c). These patterns appear in different regions and cultural contexts (Japan, Germany and the United States, for example). In languages/countries outside of species' migratory route there is often less temporal variation in the online interest further reinforcing that the role of a species' physical presence in the landscape is critical in influencing online interest. In a particularly striking example, views to the Wikipedia pages for the Barn Swallow *Hirundo rustica* respond to the annual migration of this species with such high temporal resolution that it is possible to track the swallow's northward journey using pageviews to regional Wikipedia language editions. Views to Barn Swallow pages peak first in southern European languages (Italian and Catalan), followed by central European ones (German and Czech), and finally Scandinavian languages (Swedish and Norwegian; Fig 1d).

The data sources we compared differ in the timing of their peaks; for example, ringing records are often highest in the autumn whereas eBird frequency is highest in the spring, and peaks online interest in both Google and Wikipedia lag behind the peaks in eBird frequency by several weeks. Since the data result from distinct forms of human interactions with the natural

world, these differences are expected. Some of them are easily explained; e.g. vocal spring migrants are more likely to be detected by the bird-watcher observations represented in eBird whereas high numbers of juvenile birds migrating in the fall lead to an increase in ringing records. In contrast, determining what drives people's online search behavior is challenging and at the largest scales understanding the mechanisms behind these behaviors may not always be possible. As has been demonstrated in other areas of emerging Big Data research, however, often "correlation is enough" and meaningful interpretations and conclusions can be drawn without explicit understandings of the underlying mechanisms [7].

Patterns in online behavior can have biases and errors that need to be carefully considered [8], but these freely-available data provide unique opportunities to obtain broad-scale insights relevant to biodiversity conservation, particularly given that the use of and access to online resources are projected to expand significantly in the future [9]. We foresee several promising applications from our findings. First, the high spatio-temporal correspondence between the physical presence of a species and views of its digital representations offers the potential to track migration using online data. Though clearly not a substitute for traditional survey methods, the scale and ease of access of these data could be of particular utility in regions where direct interactions with birds are scarce, difficult, or seldom recorded. Implementing this will require developing methods to calibrate the lag times between online data and migratory timings. Over time, these digital records should also provide an archive of migratory responses to ongoing environmental and climatic change and may have potential to detect range expansion and the surveillance of invasive species. In addition to their ecological applications, these patterns offer a valuable opportunity to gain perspective on the cultural and social aspects of human interest in avian migration [4]. Why does online interest track migration for some species and not others,

for example? Understanding these and other questions will aide conservation policy makers in considering human preferences and attitudes towards species and in doing so benefit conservation policy and the implementation of conservation actions. Lastly, these results show that even in this age of increased detachment from nature [10], migration still holds an important place in our hearts and minds.

References

1. McIver, D.J., and Brownstein, J.S. (2014). Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. PLoS Comput Biol. 10(4), e1003581.
2. Mitchell, L., Frank, M.R., Harris, K.D., Dodds, P.S., and Danforth, C.M. (2013). The Geography of Happiness: connecting Twitter sentiment and expression, demographics, and objective characteristics of place. PLoS One. 8(5): e64417.
3. Preis, T., Moat, H.S., and Stanley, H.E. (2013) Quantifying Trading Behavior in Financial Markets Using Google Trends. Sci. Rep. 3(1684): 1–6.
4. Ladle, R.J., Correia, R.A., Do, Y., Joo, G.J., Malhado, A.C.M., Proulx, R., et al. (2016) Conservation culturomics. Front. Ecol. Environ. 14(5): 269–75.
5. Zachte, E. Wikimedia Traffic Analysis Report: page views per wikipedia language [Internet]. 2018 [cited 2018 Jul 23]. Available from: https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm
6. Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., and Kelling, S. (2009) eBird: A citizen-based bird observation network in the biological sciences. Biol. Conserv.

142(10): 2282–92.

7. Anderson, C. (2008) The end of theory: the data deluge makes the scientific method obsolete. Wired June 30. Available from: http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

8. Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The Parable of Google Flu: traps in Big Data analysis. Science 343(6176): 1203-1205.

9. World Bank Group. Individuals using the Internet (% of population) [Internet]. The World Bank: Data. The World Bank: Data; 2018 [cited 2018 Nov 2]. Available from: https://data.worldbank.org/indicator/IT.NET.USER.ZS

10. Louv, R. (2005). Last child in the woods: saving our children from nature-deficit disorder (Chapel Hill, NC: Algonquin Books of Chapel Hill).

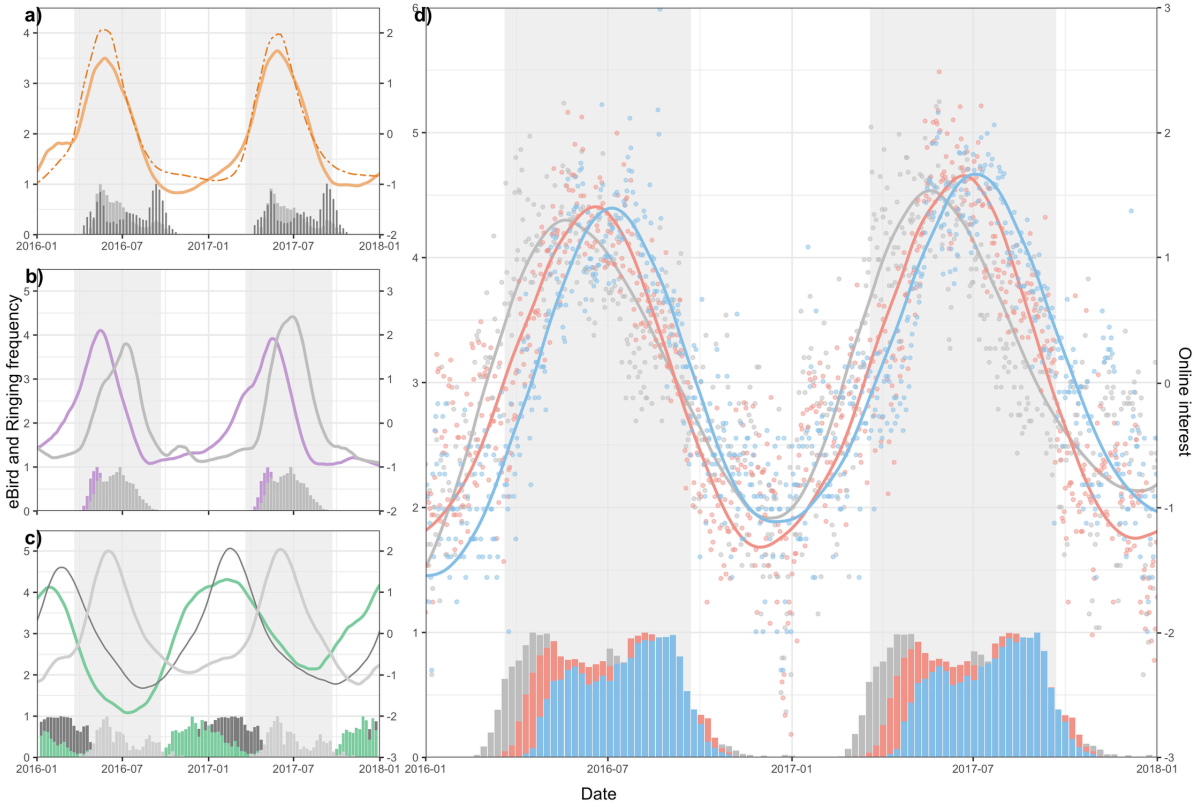


FIGURE 1. Patterns in Google Trends and Wikipedia pageviews track bird migrations around the world. a) English-language Wikipedia pageviews (solid orange line) and Google search trends in the United States (dashed orange) for Scarlet Tanager *Piranga olivacea* increase in the northern hemisphere spring and summer (dates shaded gray) when the tanagers are present on their breeding grounds in the eastern United States. This corresponds to the increased frequency of eBird reports (gray histogram) and ringing records (black bars) for Scarlet Tanagers the US. b) In German-language Wikipedia, page-views for the Common Nightingale *Luscinia megarhynchos* (purple) peak earlier in the year than those for Common Swift *Apus apus* (gray). This corresponds with the later arrival of swifts to their German-breeding grounds, as evidenced by the eBird frequency of both species in Germany (histograms). c) Similarly, Japanese Wikipedia pageviews mirror the presence of migratory species in Japan for two winter migrants, the Tundra Swan *Cygnus columbianus* (green) and the Dusky Thrush (black), and one summer

visitor to Japan, the Japanese Paradise-Flycatcher *Terpsiphone atrocaudata* (gray). Finally, d) the timing of a peak in Wikipedia page-views in regional European languages tracks the spring migration of Barn Swallows *Hirundo rustica* northward across the European continent. Views to the Wikipedia page for Barn Swallow in a southern European language (Italian, gray) peak first, followed by central European language (German, red) and finally a Scandinavian language (Swedish, blue); this sequence mirrors the spring arrival of swallows in each country as shown by the frequency with which swallows are reported in eBird in Italy (gray histogram), Germany (red histogram), and Sweden (blue histogram). eBird frequency and ringing records are scaled 0-1 (left axis). Wikipedia page-views and Google search trends are scaled by subtracting the mean and dividing by the standard deviation (right axis) and smoothed with a local polynomial regression of with a set span of 0.2.

Supplementary Materials

Supplementary Methods

Wikipedia page views for selected bird species and language editions for the period between 01 January 2016-01 January 2018 were extracted from the public Wikipedia API and summarized using R [1,2]. Fitted values for page views over this period were calculated with a local polynomial regression with a set span of 0.2 [3] and views were scaled in plots to aide in visual comparisons. Google Trend data for the period between 01 January 2016-01 January 2018 was downloaded from Google Trends (https://trends.google.com); with species being specified by their common name and the identifying topic 'bird' and the geographic region restricted to a specific country (e.g. the United States, Japan). eBird frequency data for target species was downloaded manually from the public eBird API (https://eBird.org) for the years 2016 and 2017 as separate years, with frequency records restricted to the target country (e.g. Germany, Japan). In the United States, we aggregated eBird frequency from four states in eastern North America (New York, Pennsylvania, Ohio and the Massachusetts) as a proxy for the eastern United States region. Ringing, also referred to as banding, records for target species were obtained directly from the United States Geological Survey Bird Banding Laboratory and the Yamashina Institute for Ornithology. Ring records for 2016-2018 were summarized by week.

References

1. Keyes, O. & Lewis, J. (2016). pageviews: an API client for wikimedia traffic data version 0.3.0. Available at: https://cran.rstudio.com/web/packages/pageviews/index.html. (Accessed: 1st January 2017)
2. R Core Team. (2018). R: A language and environment for statistical computing, v. 3.5.0.

3. Wang, X.-F. (2010). fANCOVA: nonparametric analysis of covariance. Available at: https://cran.r-project.org/package=fANCOVA. (Accessed: 1st May 2018)

Supplementary figures

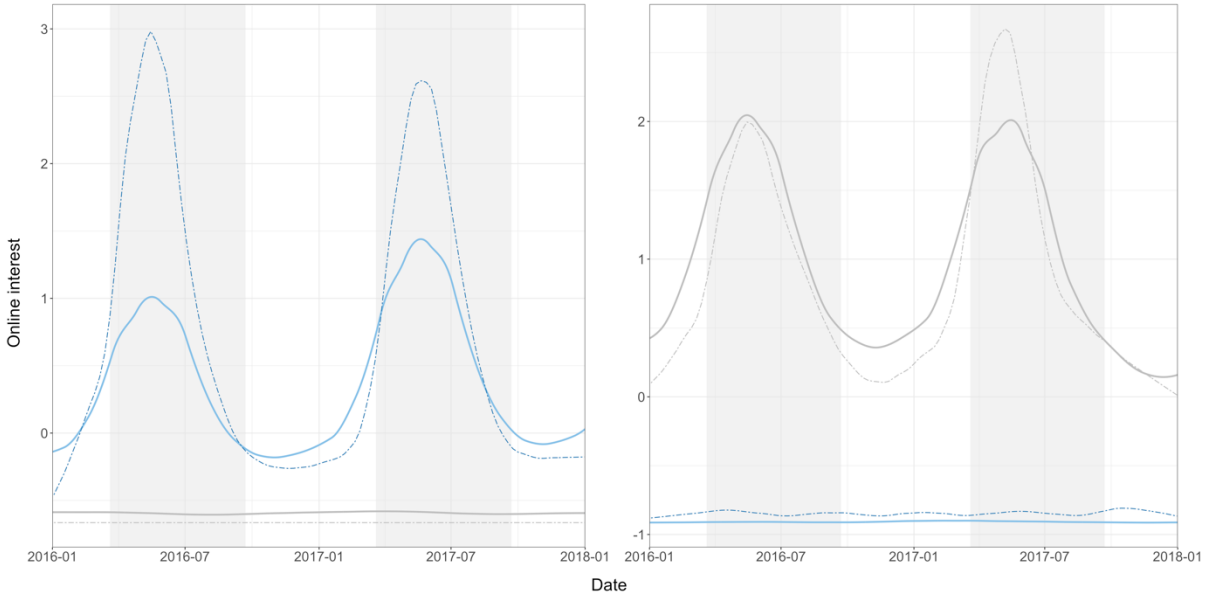


FIGURE S1. Online interest in two migratory birds responds to presence of a species in a region. a) Google Search Trends (dashed lines) in the United States and English-language Wikipedia page views (solid lines) for the Indigo Bunting *Passerina cyanea* (blue), a migratory bird that occurs in the eastern United States, show a clear seasonal pattern with increased views in the spring and summer (light gray background shading) when the buntings are present in the eastern United States. Those for the Blue-and-white Flycatcher *Cyanoptila cyanomelana* (Google Trends: gray dashed; English-language Wikipedia: solid gray), a species that does not occur in the United States show no seasonal pattern. b) Google Search Trends in Japan and Japanese-language Wikipedia page views for Blue-and-white Flycatcher (gray dashed and gray solid, respectively) increase in the spring and summer when the migratory flycatchers are present in

Japan but there is no clear seasonal pattern for Google Searches or Japanese-language Wikipedia views for Indigo Bunting (blue dashed and blue solid) which does not occur in Japan.

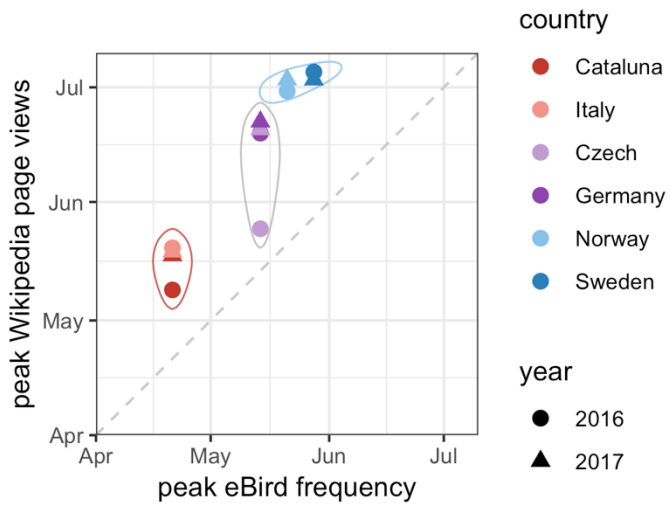


FIGURE S2. The timing of the spring peak in eBird frequency in six European regions tracks the northward migration of the Barn Swallows *Hirundo rustica* across Europe over the course of two years (2016-2017) with southern European regions (Cataluna, Italy; reds) peaking first followed by central European regions (Czech Republic and Germany; purples) and finally Scandinavian regions (Norway, Sweden; blues). The peak in Wikipedia page views in the language primarily associated with each of these regions follows a similar sequence (Catalan and Italian peak first, then Czech and German, and finally Norwegian and Swedish) but consistently lags behind the peak eBird frequency by several weeks.

Applications 2

An exciting application of online big data to species conservation is as a resource to systematically identify high interest species. In many cases, these high interest species can function as flagship species. I explored this in Methods 1, and it also emerges in the context of global reptile species in Patterns 1. This paper demonstrates an example of how high interest species identified using online data can be used into conservation policy assessments. As a co-author to the paper, I identified bird species from around the world that received the most Wikipedia pageviews across many languages. These species were manually reviewed to select those that could qualify as flagship species, and information on their occurrence and distribution was incorporated into the study. While I am not lead author on this paper, I include it here as a demonstration of one of the potential applications of online big data to conservation. Rather than providing the full paper, I include the abstract and methods sections which speak to my contributions and are the relevant sections in the context of this thesis.

Submission status: Accepted. McGowan, J., Beaumont, L.J., Smith, R., Chauvenet, A.L.M., Harcourt, R., Atkinson, S., Mittermeier, J.C., Esperon-Rodriguez, M., Baumgartner, J.B., Beattie, A., Dudaniec, R.Y., Grenyer, R., Nipperess, D., Stow, A., and H.P. Possingham. 2019. *Flagship species can deliver efficient conservation*. Nature Communications.

Personal contribution: I provided quantitative data to select flagship bird species at a global scale and helped to review and edit the manuscript.

Flagship species can deliver efficient conservation

J. McGowan[1,2,3,4*], L. J. Beaumont[1], R. J. Smith[5], A. L. M. Chauvenet[2,6], R. Harcourt[1], S. Atkinson[2], J. C. Mittermeier[7], M. Esperon-Rodriguez[1,8], J. B. Baumgartner[1], A. Beattie[1], R. Y. Dudaniec[1], R. Grenyer[7], D. Nipperess[1], A. Stow[1], and H. P Possingham[3]

1. Department of Biological Sciences, Macquarie University, NSW, 2109, Australia.
2. Centre for Biodiversity and Conservation Science, University of Queensland, QLD,4072, Australia
3. The Nature Conservancy, Arlington, VA, USA
4. WildArk.org, Atlanta, GA USA
5. Durrell Institute of Conservation and Ecology, University of Kent, Canterbury, Kent, UK
6. Environmental Futures Research Institute & School of Environment and Science, Griffith University, QLD, 4222, Australia.
7. School of Geography and Environment, Oxford University, South Parks Road, Oxford, OX1 3QY, UK.
8. Hawkesbury Institute for the Environment, Western Sydney University, 2753 NSW Australia

Abstract

Conservation strategies based on charismatic flagship species, such as tigers, pandas and gorillas, successfully attract funding from individuals and corporate donors. However, advocates of place-based conservation criticise this approach, arguing it is more effective to focus on areas with high biodiversity value (e.g. places with concentrations of threatened or endemic species)

than on individual species whose conservation may not deliver benefits to broader biodiversity. Here, we ask – to what extent is fundraising through a flagship species approach a major constraint on delivering efficient place-based conservation? To answer this, we constructed a novel place-based approach that incorporates flagship species while simultaneously maximizing a biodiversity objective (defined as the conservation of 19,616 vertebrate species). We tested our approach using eight global planning scenarios and compared the results to a purely place-based approach (e.g. one that efficiently maximizes the biodiversity objective, irrespective of the presence of flagship species) and two null models. Our integrated approach achieved 80-89% of the objective across the range of prioritization scenarios. For the first time, we provide strong evidence that prudently selected flagship species can help deliver efficient conservation. This allows organizations and private ventures, whose role in conservation continues to grow, to maximise public awareness and attract funding while accommodating important attributes of both species-based and place-based conservation that are relevant to their conservation goals.

Materials and Methods

Selecting candidate flagship species

We used two approaches to identify plausibly charismatic candidate species. For mammals, we used existing conservation flagships (N = 80) and the "Cinderella" species (N = 183) identified by Smith et al. (*3*). Cinderella species have similar physical characteristics to flagships, namely large body size and forward-facing eyes, but are not known to be conservation flagships (*3*). For reptiles and birds, we identified candidate species using an approach developed in Roll et al. (*35*) that quantifies interest in species based on their online popularity, measured via the number of Wikipedia page views for a given year. Popular reptile species were taken from the previous

work of Roll et al. (*35*). Bird species were similarly identified matching the global species taxonomy of the International Ornithological Committee (IOC World Bird List version 7.1) to English language Wikipedia pages and extracting views to each species page for the period between 1 January 2016—1 February 2017 (Mittermeier et al., unpublished). The top 100 reptile and 500 bird species, measured by total page views, were identified as the potential flagship representatives of these groups. However, as range maps were not available for all candidates, our final list was reduced to 534 species (Table S1). We assumed that all species in the list of candidates have equal capacity to serve as a conservation flagship given dedicated marketing efforts (*1)*.

References (selected)

1. D. Veríssimo *et al.*, Increased conservation marketing effort has major fundraising benefits for even the least popular species. *Biological Conservation* **211**, 95-101 (2017).

3. R. J. Smith, D. Veríssimo, J. B. Isaac Nicholas, E. Jones Kate, Identifying Cinderella species: uncovering mammals with conservation flagship appeal. *Conservation Letters* **5**, 205-212 (2012).

35. U. Roll *et al.*, Using Wikipedia page views to explore the cultural importance of global reptiles. *Biological Conservation* **204**, 42-50 (2016).

Conclusion

At the start of the thesis, I introduced a global challenge, the conservation of earth's biological diversity, and described the rapid emergence of a new technological resource, online big data. The purpose of the thesis has been to explore whether insights from this new resource can help to address the global conservation challenge, and specifically whether online big data can be used to better understand some of the human variables associated with species conservation. When I began my research five years ago, this was an untested idea in conservation. Together with my supervisors and co-authors, I spent considerable effort thinking about whether online data such as Wikipedia pageviews could be a valid resource for addressing conservation questions at all. Times change. My thesis has coincided with a broader recognition of the potential applications of online big data to conservation. This is reflected in the growing interest in conservation culturomics as a research area. At the start of my thesis, the term "conservation culturomics" did not yet exist (it was first coined by Ladle et al. in 2016); now, there is a dedicated conservation cultromics working group in the Society for Conservation Biology and a special journal issue on "Advancing Conservation Culturomics" scheduled to appear in *Conservation Biology* later this year. The papers presented here have helped to contribute to the development of conservation culturomics as a research area. In particular, they have added to this new field by developing Wikipedia as a conservation culturomics resource and by evaluating patterns of interest in species at much larger scales than any previous studies.

As I outlined in the Introduction, the goals of the thesis are threefold. First, to introduce *methods* to access and analyse Wikipedia data as a conservation culturomics resource. Second, to explore some of the *patterns* demonstrated by these data. Third, to touch upon a few of the

conservation *applications* of these new methods. By way of concluding, I will briefly review the main findings of the thesis with regard to each of these goals and reflect on some of the things I have learned while developing these ideas.

Methods

My early research proposals included plans to combine data from digital books, newspapers, social media platforms, Google searches, and Wikipedia. In hindsight, these plans had almost comically little awareness of some of the challenges associated with big data. Developing methods to access and interpret online big data in the context of conservation can be complex. I imagine that many researchers coming from fields outside of computer science or big data analytics, as I did, may face a similar shock when they first begin conservation culturomics research. I would summarize my methodological conclusions from the thesis into three points. The first two are words of caution: do not underestimate challenges associated with data access, and do not underestimate challenges associated with data interpretation. The third is a recommendation. The more I worked with Wikipedia, the more I came to appreciate that it is a unique and valuable tool for researchers. Wikipedia data should always be on the short-list of potential information sources for conservation culturomics research.

1. Do not underestimate the technical challenges associated with big data

Accessing online data can be challenging. In order to get a year's worth of Wikipedia pageviews for reptiles in 2015, Gonzalo Diaz and I had to download nearly 9,000 files with hourly views to all of the pages in Wikipedia and write custom scripts to sift through each file to find pages related to reptiles. Decompressed, the hourly view files were large enough that we could

only open a few at a time, and the entire operation of downloading, uncompressing, and checking for reptile pages was so computationally intensive that we ran it on server in the computer science department. The code took nearly a week to run, and any error meant that we had to restart the process from the beginning. Today Wikipedia has restructured its pageview files and this process is vastly easier (see the methods discussions in Patterns 2 and 3), but the example of our initial experience still holds true in many big data analyses.

In the context of big data, even small, seemingly simple questions can be technically and computationally difficult. While Wikipedia has improved access for researchers, other platforms have not. Some platforms have even reduced access for researchers; Instagram, for example, shut down its public application programming interface in 2018. The first step for any culturomics research project is to carefully assess the technical hurdles associated with accessing the relevant data.

In the context of this thesis, my solution to the challenge of access was to focus primarily on a single data source, Wikipedia. This allowed me to delve deeply into the unique technicalities associated with access. Focussing on a single data source obviously has limitations, but as an initial stage in the development of conservation culturomics research, it provided a valuable opportunity to create methodological foundations.

2. Do not underestimate the interpretation challenges associated with big data

The sheer volume of online big data generates challenges in interpreting how patterns in these data translate into actions and behaviours in the real world. The reason that people in the United States view the Wikipedia page for Wild Turkey in English may be quite different from the reason that people in Japan view the Japanese-language page for the species, for example. While

it would be possible to delve into these differences for small subset of pages, understanding these fine-scale patterns at the scale of hundreds of languages and millions of articles becomes much harder, and often impossible. In a dataset of nearly 3 billion pageviews, such as the one I used in Patterns 2, we can guess at what is motivating most people when they click on a page but having specificity beyond that is difficult.

My solution to this challenge in the context of the thesis was to focus primarily on Wikipedia pageviews. As I described in Methods 1, pageviews have a clearer intention associated with them (searching for additional knowledge about a subject) than many other data types, and have the advantage of an existing body of research that explores their utility as a measure of public awareness and interest (e.g. Skiena & Ward 2014; Yu et al. 2016). Similar to the issues associated with focussing on a single dataset that I mentioned above, however, pageviews only offer a partial picture of people's online behaviour. Now that we understand some of the patterns that appear in pageview data, however, we can use these as a build block against which to compare patterns in other types of Wikipedia data and from other data sources.

3. Wikipedia should be on the short-list list of conservation culturomics datasets

I was initially slow to embrace Wikipedia as a data source (perhaps due to the many reminders I had received from undergraduate professors that Wikipedia is not a source for academic research). As I have progressed, however, I have become increasingly appreciative of Wikipedia's value as a research tool. Wikipedia data have limitations, obviously. As I have described, they are not effective at measuring people's sentiment towards an entity, something with which is often of interest to conservationists, and they lack fine-scale geographic information. Like online data in general, they present only one view of the world and have particular biases.

With those in mind, however, Wikipedia offers a rich resource for conservation research. To have such a widely used website be well structured and completely open access is truly exceptional. Wikipedia pageviews, in particular, are one of the best, if not *the best*, resource currently available for researchers interested in comparing public interest across large numbers of languages and biodiversity features. Even if Wikipedia is not used as a primary data source, as I have done for many of the papers in this thesis, the increasing ease of access to Wikipedia data together with our newfound understanding of how to interpret these data in a conservation context mean that insights from Wikipedia can be added to wide array of studies. In conservation culturomics specifically, future research can integrate the approaches I describe in this thesis with data from other sources to form more holistic understandings of human interactions. In the not so distant future, I imagine that my original aspirations of combining multiple data sources will be possible.

Patterns

After considering methods to access and interpret data from Wikipedia and specifically Wikipedia pageviews, I explored patterns in these data in Chapter 2. The strength of Wikipedia for researchers is its ability to make comparisons across very broad scales. The thousands of species and dozens of languages that I compared in Patterns 1, 2, and 3, for example, are larger sample sizes than any previous studies of human interest in species. For me, four main conclusions from these large-scale analyses stand out.

1. Biological traits influence public interest (but less than might be expected)

It is well-known in conservation that the biological traits of species impact their public attention, and previous studies have identified features such as large body size, forward-facing

eyes and bright colours as important to generating higher public interest (e.g. Leader-Williams & Dublin 2000; Frynta et al. 2010; Albert et al. 2018). These existing studies, however, have all used relatively small samples sizes (at most a few dozen species). As a result, revisiting these questions through the lens of big data with sample sizes of thousands or tens of thousands of species, as I did in Chapter 2, provides an opportunity to gain new perspective.

In some cases, I found that this big data perspective confirmed the results of previous studies. Larger species, for example, also attract more attention in Wikipedia, and body size is a strong predictor of public interest in both reptiles (Patterns 1) and birds (Patterns 2). In other cases, however, the big data perspective offers some new insights. It indicates that certain patterns may not be widespread as has been previously assumed. Being colourful, for example, does not seem to be an important factor in determining the most popular reptiles (Patterns 1) or birds (Methods 1) in Wikipedia; none of the most-viewed species in either group are especially brightly coloured. Most importantly, however, the perspective of big data demonstrates that while the biological traits of species can be important to determining public interest, at a broad scale they are often secondary to factors such as timing, geography, and the local presence and abundance of species (Patterns 2 and 3).

2. Temporal patterns influence interest in biodiversity

Evaluating public interest in biodiversity with online big data reveals that temporal patterns are an important, and often overlooked, variable is in determining people's interactions with the natural world. Wikipedia pages for species are more than three times as likely to have a seasonal pattern than non-species pages and, across languages many languages, around 20% of all species-pages show a seasonal pattern in their pageviews (Patterns 2). For some species, temporal factors

are so significant that the Wikipedia pageviews for species track their migratory movements (Applications 1). The prevalence of these temporal patterns demonstrates that timing and seasonality are important drivers in people's interactions with species and, as I will discuss more below, that the physical presence and visibility of some species can be a significant to determining their public interest.

3. Geographic patterns influence interest in biodiversity

In addition to revealing temporal patterns, the big data perspective demonstrates the importance of geography in influencing people's interest in biodiversity. Geographic patterns are visible in the results throughout the papers in the thesis, with particularly conspicuous examples being the way in which languages cluster geographically in Methods 1 and the fact that people are consistently more interested in local species than non-local ones (Patterns 3). One outcome of these patterns is the existence of 'geographic scale dependence' in people's interest. As I describe in Patterns 3, for species that occur far away from where people live, being large has increased importance in determining interest, but for those species that occur locally, being common in the local region is the primary driver of public attention.

4. The local presence and abundance of species is an important driver of interest

This final point is closely intertwined with the temporal and geographic patterns described above. I found it to be so striking, however, that I feel it deserves a separate heading: the physical presence of a species in the landscape is often highly significant in correlating with increased public interest. Over the course of the thesis, this relationship emerged in a temporal context, with pageviews tracking the migratory movements of some species in Applications 1, and in a

geographic context, with local, commonly observed bird species generating higher interest in Patterns 3.

Given the widespread narrative in conservation that people are increasingly disconnected with the natural world and unaware of the species outside their windows, I did not anticipate this relationship being so significant. I find its prevalence reassuring. Despite the concerns of conservationists, many people are still paying attention to and are interested in the migratory birds and emerging flowers around their homes. This relationship is potentially concerning, however. As I point out in Patterns 3, if interest in a species is contingent on how frequently people encounter it in the wild, then the ongoing declines of wild animals around the world could lead to reduced public interest in many species.

Applications

The principal aim of the thesis has been to introduce and develop methodological tools that have conservation applications. In Chapter 3 and to varying extents in the preceding papers, I identified four ways in which these methods can be applied to conservation. These four areas are a useful starting point, but there are undoubtedly more possibilities beyond what I discuss here. My hope is that, in addition to being useful themselves, these applications serve as inspiration for conservationists considering additional ways to apply the methods in the thesis to their own research.

1. Understanding why some species attract more public interest

As I have described, online big data are an excellent resource for investigating what influences public interest in species. Many conservation measures require public support to be

effective, and thus understanding these drivers of public interest have applied as well as theoretical relevance. The relationship between public interest and the local abundance of species that I described in the previous section, for example, could have practical repercussions: if interest in species is tied to encounter rates with wild individuals then one way to increase interest could be through creating opportunities for people to have greater engagement with common, local species (Patterns 3).

2. Systematically identifying high interest species

The ability of online big data to quantitively compare species also make these data useful for identifying high interest species. While the species that attract the greatest interest online are not necessarily good flagship species, many of them could function in this role. Incorporating online data such as Wikipedia pageviews into flagship species selection processes could help to make these selections more systematic, data driven, and relevant to constituencies outside of conservation. One of the benefits of online big data is their ability to identify interactions that may not be obvious to people outside of the groups involved in those interactions. For researchers based in the United States, for example, it be a surprise that the Shoebill is amongst the most viewed bird species in Japanese Wikipedia. Likewise, for people who do not watch the Discovery Channel or go hunting, the peaks in online interest for Great White Sharks during the broadcasting of "Shark Week" and for Wild Turkeys during the spring hunting season may not be readily apparent (Patterns 2). Using online data to identify high interest species requires carefully considering the appropriate data source as well as variables associated with geography, language, and metadata (Methods 1). Wikipedia pageviews are an especially useful resource for making these assessments,

and the rankings of species generated by Wikipedia pageviews are applicable to a variety of conservation applications (e.g. Methods 1, Applications 2).

3. Monitoring temporal and spatial trends in people's interest

Online big data provide opportunities to monitor temporal and spatial trends in people's interest and assess how those trends change over time. Increasing awareness of conservation needs and actions is critically important to modern conservation, particularly in the context of the failures to meet conservation targets that I mentioned at the beginning of the thesis. Conservation marketing and environmental education are important tools for accomplishing increased awareness, and online data can help improve the scope and effectiveness of these by providing tools to monitor where the people are who are most interested in a particular species (e.g. Methods 1, Patterns 3), when those people are most interested in those species (Patterns 2), and whether patterns in their interest change as a result of conservation interventions and marketing campaigns.

4. Monitoring temporal and spatial trends in biodiversity

Collecting biodiversity data using standard methods is often costly and time-consuming and as a result many types of biodiversity analyses suffer from a chronic shortage of data on the distribution and abundance of species. The possibility that online big data could help resolve these shortfalls is appealing, especially given that public interest in some species can respond to the migratory patterns and abundance of species (Patterns 2 and 3). Could Wikipedia pageviews be used to create a Google Flu type monitoring system for migratory birds, for example? In contrast to the previous three points, where I argue that there are strong and obvious benefits to using online data, the applicability of online data to biodiversity monitoring is more equivocal. With proper

calibrating it could be an effective tool in some situations (the swallow migrations in Fig. 1 of Applications 1, for example) but it is unlikely to work for the majority of species. Even for those species where online patterns do closely mirror real-world distributions, using these data for planning and monitoring would require careful calibration and validation to be effective.

Parting thoughts

This thesis has covered a lot of ground. Here are the conclusions I hope you will take with you.

Online big data can contribute a lot to biodiversity conservation. These data are not a replacement to existing data sources or methods, but a complementary approach that can add new and potentially powerful dimensions to the modern conservationist's tool kit. The challenges associated with accessing and interpreting these data can be substantial and need to be carefully considered and critically assessed. When used appropriately, however, online big data offer exciting opportunities to address questions that were difficult, or even impossible, to study in the past.

Within the growing ecosystem of online data platforms, Wikipedia is one of the best available resources and should be one of the first sources considered for many conservation culturomics applications. Wikipedia pageviews, in particular, are effective for investigating patterns of public interest and attention across large numbers of features such as species and amongst many different languages. Patterns in Wikipedia pageviews help to reveal the biological traits that correspond with higher public interest in species and emphasise the importance of temporal and geographic patterns in determining people's awareness of biodiversity.

Wikipedia pageviews also have a variety of conservation applications. They can be used to understand why some species attract greater public interest than others, to systematically identify high interest species at regional and global scales, and to monitor temporal and spatial patterns in people's interest. In some cases, it may even be possible to use Wikipedia pageviews to track the geographic movements and distributions of species.

The work shared in this thesis is a first step in developing the potential of online big data and in particular Wikipedia to conservation. As the quantity and resolution of data in the digital universe continues its exponential expansion, the opportunities these data offer for twenty-first century conservationists will only continue to grow.

References

Aiden E, Michel J-B. 2013. Uncharted: Big Data as a lens on human culture. Riverhead Books, New York, New York, USA.

Albert C, Luque GM, Courchamp F. 2018. The twenty most charismatic species. PLoS ONE **13**:e0199149.

Anderson C. 2008, June 30. The end of theory: the data deluge makes the scientific method obsolete. Wired:Online. Available from http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory.

Arts K, van der Wal R, Adams WM. 2015. Digital technology and the conservation of nature. Ambio **44**:661–673.

Barnosky AD et al. 2011. Has the Earth's sixth mass extinction already arrived? Nature **471**:51–57.

Bohannon J. 2011. Google Books, Wikipedia, and the future of Culturomics. Science **331**:135.

Bollen J, Mao H, Zeng X-J. 2011. Twitter mood predicts the stock market. Journal of Computational Science **2**:1–8.

boyd d, Crawford K. 2012. Critical Questions for Big Data. Information, Communication & Society **15**:662–679.

Bright AD, Manfredo MJ. 1996. A conceptual model of attitudes toward natural resource issues: a case study of wolf reintroduction. Human Dimensions of Wildlife **1**:1–21.

Burivalova Z, Butler RA, Wilcove DS. 2018. Analyzing Google search data to debunk myths about the public's interest in conservation. Frontiers in Ecology and the Environment:1–6.

Buscher B, Sullivan S, Neves K, Igoe J, Brockington D. 2012. Towards a synthesized critique of neoliberal biodiversity conservation. Capitalism, Nature, Socialism **23**:4–30.

Butchart SHM et al. 2010. Global biodiversity: indicators of recent declines. Science **328**:1164–8.

Butchart SHM, Stattersfield AJ, Collar NJ. 2006. How many bird extinctions have we prevented? Oryx **40**:266–278.

Cantú-Salazar L, Orme CDL, Rasmussen PC, Blackburn TM, Gaston KJ. 2013. The performance of the global protected area system in capturing vertebrate geographic ranges. Biodiversity and Conservation **22**:1033–1047.

Caro T. 2010. Conservation by proxy: indicator, umbrella, keystone, flagship, and other surrogate species. Island Press, New York, New York, USA.

CBD. 2002. 2010 Biodiversity Target. Available from https://www.cbd.int/2010-target (accessed October 5, 2019).

Child MF. 2009. The Thoreau Ideal as a Unifying Thread in the Conservation Movement. Conservation Biology **23**:241–243.

Cioffi-Revilla C. 2010. Computational social science. Wiley Interdisciplinary Reviews: Computational Statistics **2**:259–271.

Correia RA et al. 2020. Digital data sources and methods for conservation culturomics. Conservation Biology.

Correia RA, Di Minin E, Jarić I, Jepson P, Ladle R, Mittermeier J, Roll U, Soriano-Redondo A, Veríssimo D. 2019. Inferring public interest from search engine data requires caution. Frontiers in Ecology and the Environment **17**:254–255.

Correia RA, Jepson PR, Malhado ACM, Ladle RJ. 2016. Familiarity breeds content: assessing bird species popularity with culturomics. PeerJ **4**:1–15.

Costanza R et al. 1997. The value of the world's ecosystem services and natural capital. Science

387:253–260.

Cristancho S, Vining J. 2004. Culturally defined keystone species. Human Ecology Review **11**:153–164.

De Goes J. 2013. “Big Data” is dead. What’s next? Venture Beat News:Online. Available from http://venturebeat.com/2013/02/22/big-data-is-dead-whats-next/.

de Groot RS, Wilson MA, Boumans RMJ. 2002. A typology for the classification, description and valuation of ecosystem functions, goods and services. Ecological Economics **41**:393–408.

De Vos JM, Joppa LN, Gittleman JL, Stephens PR, Pimm SL. 2014. Estimating the Normal Background Rate of Species Extinction. Conservation Biology **00**:1–10.

Di Minin E, Tenkanen H, Toivonen T. 2015. Prospects and challenges for social media data in conservation science. Frontiers in Environmental Science **3**:1–6.

Diebold FX. 2012. A Personal Perspective on the Origin(s) and Development of “Big Data”: The phenomenon, the term, and the discipline, second version. Page Pier Working Paper Archive.

Dirzo R, Young HS, Galetti M, Ceballos G, Isaac NJB, Collen B. 2014. Defaunation in the Anthropocene. Science **345**:401–406.

Doak DF, Bakker VJ, Goldstein BE, Hale B. 2013. What is the future of conservation? Trends in Ecology & Evolution **29**:77–81.

Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM. 2011. Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. PLoS ONE **6**:e26752.

Domo Inc. 2018. Data Never Sleeps 6.0. Available from https://www.domo.com/learn/data-

never-sleeps-6 (accessed April 5, 2019).

Ehrlich PR, Ehrlich AH. 1981. Extinction: the causes and consequences of the disappearance of species. Random House, New York, New York, USA.

Einav L, Levin J. 2013. The Data Revolution and Economic Analysis. Stanford Univerity and NBER:1–29.

EMC Digital Universe. 2014. The Digital Universe of Opportunities: rich data and the increasing value of the internet of things. Executive summary. Available from http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm.

Faith DP. 1992. Conservation evaluation and phylogenetic diversity. Biological Conservation **61**:1–10.

Frynta D, Lisková S, Bültmann S, Burda H. 2010. Being attractive brings advantages: the case of parrot species in captivity. PloS ONE **5**:e12568.

Garibaldi A, Turner N. 2004. Cultural keystone species: Implications for ecological conservation and restoration. Ecology and Society **9**:1.

Gilbert N. 2009. Efforts to sustain biodiversity fall short. Nature **462**:263.

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. 2009. Detecting influenza epidemics using search engine query data. Nature **457**:1012–1014. Nature Publishing Group. Available from http://dx.doi.org/10.1038/nature07634.

Goble DD, Wiens JA, Scott JM, Male TD, Hall JA. 2012. Conservation-Reliant Species. BioScience **62**:869–873.

Graham S, Milligan I, Weingart S. 2016. Exploring Big Historical Data: the historian's macroscope. Imperial College Press, London, UK.

Gunnthorsdottir A. 2001a. Physical attractiveness of an animal species as a decision factor for its

preservation. Anthrozoos **14**:204–215.

Gunnthorsdottir A. 2001b. Physical attractiveness of an animal species as a decision factor for its preservation. Anthrozoos **14**:204–214.

Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, Duke CS, Porter JH. 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment **11**:156–162.

Hausmann A, Toivonen T, Heikinheimo V, Tenkanen H, Slotow R, Minin E DI. 2017. Social media reveal that charismatic species are not the main attractor of ecotourists to sub-Saharan protected areas. Scientific Reports **7**:1–9.

Hausmann A, Toivonen T, Slotow R, Tenkanen H, Moilanen A, Heikinheimo V, Di Minin E. 2018. Social Media Data Can Be Used to Understand Tourists' Preferences for Nature-Based Experiences in Protected Areas. Conservation Letters **11**:1–10.

Hey J. 2001. The mind of the species problem. Trends in Ecology and Evolution **16**:326–329.

Hey T, Tansley S, Tolle K. 2009. Jim Gray on eScience: a transformed scientific method. Pages xvii–xxxi in T. Hey, S. Tansley, and K. Tolle, editors. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond.

Hoffmann M et al. 2010. The Impact of Conservation on the Status of the World's Vertebrates. Science **330**:1503–1509.

Howe D, Yon S. 2008. The future of biocuration. Nature **455**:47–50.

IUCN. 2014. The IUCN Red List of Threatened Species. Version 2014.3. Available from www.iucnredlist.org (accessed November 17, 2014).

Justus J, Colyvan M, Regan H, Maguire L. 2009. Buying into conservation: intrinsic versus instrumental value. Trends in Ecology and Evolution **24**:187–191.

Kandula S, Shaman J. 2019. Reappraising the utility of Google Flu Trends. PLOS Computational Biology **15**:e1007258.

Kareiva P, Marvier M. 2012. What Is Conservation Science? BioScience **62**:962–969.

Kayyali B, Knott D, Kuiken S Van. 2013. The big-data revolution in US healthcare: accelerating value and innovation. McKinsey & Company:1–6.

Kellert SR. 1982. Factors in Endangered Social and Perceptual Species Management. The Journal of Wildlife Management **49**:528–536.

Kelling S, Fink D, La Sorte FA, Johnston A, Bruns NE, Hochachka WM. 2015. Taking a 'Big Data' approach to data quality in a citizen science project. Ambio **44**:601–611.

Kidd LR, Gregg EA, Bekessy SA, Robinson JA, Garrard GE. 2018. Tweeting for their lives: Visibility of threatened species on twitter. Journal for Nature Conservation **46**:106–109.

Kitchin R. 2014a. The Data Revolution: big data, open data, data infrastructures and their consequences. SAGE Publications, London, UK.

Kitchin R. 2014b. Big Data, new epistemologies and paradigm shifts. Big Data & Society **1**:1–12.

Ladle RJ, Correia RA, Do Y, Joo GJ, Malhado ACM, Proulx R, Roberge JM, Jepson P. 2016. Conservation culturomics. Frontiers in Ecology and the Environment **14**:269–275.

Ladle RJ, Jepson P. 2008. Toward a biocultural theory of avoided extinction. Conservation Letters **1**:111–118.

Ladle RJ, Jepson P. 2010. Origins, Uses, and Transformation of Extinction Rhetoric. Environment and Society: Advances in Research **1**:96–115.

Lazer D et al. 2009. Life in the network: the coming age of computational social science. Science **323**:721–723.

Lazer D, Kennedy R, King G, Vespignani A. 2014. The Parable of Google Flu: traps in Big Data analysis. Science **343**:1203–1205. Available from http://www.sciencemag.org/content/343/6176/1203.

Leader-Williams N, Dublin HT. 2000. Charismatic megafauna as "flagship species." Pages 53–81 in A. Entwistle and N. Dunstone, editors. Has the panda had its day? Priorities for the conservation of mammalian diversity. Cambridge University Press, Cambridge, UK.

Leinweber DJ. 2007. Stupid Data Miner Tricks. The Journal of Investing **16**:15–22.

Lindemann-Matthies P. 2002. The Influence of an Educational Program on Children's Perception of Biodiversity. The Journal of Environmental Education **33**:22–31.

Lindemann-Matthies P. 2005. "Loveable" mammals and "lifeless" plants: How children's interest in common local organisms can be enhanced through observation of nature. International Journal of Science Education **27**:655–677.

Mace GM. 2014. Whose conservation? Science **345**:1558–1560.

Maguire LA, Justus J. 2008. Why Intrinsic Value Is a Poor Basis for Conservation Decisions. BioScience **58**:910.

Marr B. 2017. Data Strategy: how to profit from a world of big data, analytics and the internet of things. Kogan Page Limited, London, UK.

Martín-Forés I, Martín-López B, Montes C. 2013. Anthropomorphic Factors Influencing Spanish Conservation Policies of Vertebrates. International Journal of Biodiversity **2013**:1–9.

Marvier M, Kareiva P. 2014. The evidence and values underlying "new conservation." Trends in Ecology and Evolution **29**:131–132.

Mayer-Schönberger V, Cukier K. 2013. Big Data: a revolution that will transform how we live, work and think. Houghton Mifflin Harcourt, Boston, MA.

McAfee A, Brynjolfsson E. 2012. Big Data. The management revolution. Harvard Buiness Review **90**:61–68.

McCauley DJ. 2006. Selling out on nature. Nature **443**:27–28.

Michel J, Shen Y, Aiden A, Veres A. 2011. Quantitative analysis of culture using millions of digitized books. Science **331**:176–183.

Mills LS, Soule ME, Doak DF. 1993. The keystone-species concept in ecology and conservation. BioScience **43**:219–224.

Myers N. 1983. A priority-ranking strategy for threatened species? The Environmentalist **3**:97–120.

Nash RF. 1990. The Rights of Nature: history of environmental ethics. University of Wisconsin Press, Madison, Wisconsin.

Nghiem LTP, Papworth SK, Lim FKS, Carrasco LR. 2016. Analysis of the capacity of google trends to measure interest in conservation topics and the role of online news. PLoS ONE **11**:1–12.

Pimm SL, Russell GJ, Gittleman JL, Brooks TM. 1995. The future of biodiversity. Science **269**:347–350.

Prensky M. 2009. H. sapiens digital: from digital immigrants and digital natives to digital wisdom. Available from http://www.wisdompage.com/Prensky01.html.

Primack RB. 2014. Essentials of conservation biology, sixth edition. Sinauer Associates, Inc., Sunderland, MA.

Proulx R, Massicotte P, Pepino M. 2013. Googling Trends in Conservation Biology. Conservation Biology **28**:44–51.

Raghupathi W, Raghupathi V. 2014. Big data analytics in healthcare: promise and potential.

Health Information Science and Systems **2**:1–10.

Rands MRW et al. 2010. Biodiversity conservation: challenges beyond 2010. Science **329**:1298–1303.

Rieder B, Rohle T. 2017. Digital Methods. Page in M. T. Schafer and K. van Es, editors. The Datafied Society: Studying Culture through Data. Amsterdam University Press, Amsterdam.

Roberge JM. 2014. Using data from online social networks in conservation science: Which species engage people the most on Twitter? Biodiversity and Conservation **23**:715–726.

Rodrigues ASL, Pilgrim JD, Lamoreux JF, Hoffmann M, Brooks TM. 2006. The value of the IUCN Red List for conservation. Trends in Ecology & Evolution **21**:71–76.

Rosenberg K V et al. 2019. Decline of the North American avifauna. Science **1313**:1–10.

Santana C. 2014. Save the planet: eliminate biodiversity. Biology and Philosophy:1–20.

Schich M, Song C, Ahn Y-Y, Mirsky A, Martino M, Barabási A-L, Helbing D. 2014. A network framework of cultural history. Science **345**:558–562.

Schwartz MW. 2008. The Performance of the Endangered Species Act. Annual Review of Ecology, Evolution, and Systematics **39**:279–299.

Scott JM, Goble DD, Haines AM, Wiens JA, Neel MC. 2010. Conservation-reliant species and the future of conservation. Conservation Letters **3**:91–97.

Scott JM, Goble DD, Wiens JA, Wilcove DS, Bean M, Male T. 2005. Recovery of imperiled species under the Endangered Species Act: the need for a new approach. Frontiers in Ecology and the Environment **3**:383–389.

Sheil D, Meijaard E. 2010. Purity and Prejudice: Deluding Ourselves About Biodiversity Conservation. Biotropica **42**:566–568.

Sivarajah U, Kamal MM, Irani Z, Weerakkody V. 2017. Critical analysis of Big Data challenges

and analytical methods. Journal of Business Research **70**:263–286. The Authors. Available from http://dx.doi.org/10.1016/j.jbusres.2016.08.001.

Skiena S, Ward C. 2014. Who's Bigger? Where historical figures really rank. Cambridge University Press, New York, New York, USA.

Soriano-Redondo A, Bearhop S, Lock L, Votier SC, Hilton GM. 2017. Internet-based monitoring of public perception of conservation. Biological Conservation **206**:304–309.

Soule ME. 1985. What is Conservation Biology? BioScience **35**:727–734.

Stokes DL. 2007. Things we like: human preferences among similar organisms and implications for conservation. Human Ecology **35**:361–369.

Sutherland WJ et al. 2018. A 2018 Horizon Scan of Emerging Issues for Global Conservation and Biological Diversity. Trends in Ecology and Evolution **33**:47–58. Elsevier Ltd. Available from http://dx.doi.org/10.1016/j.tree.2017.11.006.

Tallis H, Lubchenco J. 2014. A call for inclusive conservation. Nature **515**:27–28.

TEEB. 2010. The Economics of Ecosystems and Biodiversity: mainstreaming the economics of nature: a synthesis of the approach, conclusions and recommendations of TEEB. Available from www.teebweb.org (accessed August 1, 2017).

Tisdell CA. 2014. Human values and biodiversity conservation: the survival of wild species. Edward Elgar, Cheltenham, UK.

Toivonen T, Heikinheimo V, Fink C, Hausmann A, Hiippala T, Järv O, Tenkanen H, Di Minin E. 2019. Social media data for conservation science: A methodological overview. Biological Conservation **233**:298–315. Elsevier. Available from https://doi.org/10.1016/j.biocon.2019.01.023.

van der Wal R, Arts K. 2015. Digital conservation: An introduction. Ambio **44**:517–521.

van Es K, Lopez Coombs N, Boeschoten T. 2017. Towards a reflexive digital data analysis. Page in M. T. Schafer and K. van Es, editors. The Datafied Society: Studying Culture through Data. Amsterdam University Press, Amsterdam.

Vane-Wright RI, Humphries CJ, Williams PH. 1991. What to protect?—Systematics and the agony of choice. Biological Conservation **55**:235–254.

Walpole M et al. 2009. Tracking progress toward the 2010 biodiversity target and beyond. Science **325**:1503–1504.

Walpole MJ, Leader-Williams N. 2002. Tourism and flagship species in conservation. Biodiversity and Conservation **11**:543–547.

Williams CK, Ericsson G, Heberlein T. 2002. A quantitative summary of attitudes toward wolves and their reintroduction (1972-2000). Wildlife Society Bulletin **30**:575–584.

Wilson EO. 1984. Biophilia. Harvard University Press, Cambridge, MA.

Wood SA, Guerry AD, Silver JM, Lacayo M. 2013. Using social media to quantify nature-based tourism and recreation. Scientific Reports **3**:2976.

Young HS, McCauley DJ, Galetti M, Dirzo R. 2016. Patterns, Causes, and Consequences of Anthropocene Defaunation. Annual Review of Ecology, Evolution, and Systematics **47**:333–358.

Yu AZ, Ronen S, Hu KZ, Hidalgo C a. 2016. Pantheon 1.0, a manually verified dataset of globally famous biographies. Scientific Data **3**:150075.

Zikopoulos PC, Eaton C, DeRoos D, Deutsch T, Lapis G. 2012. Understanding Big Data. McGraw Hill, New York, New York, USA.

Zmihorski M, Dziarska-Palac J, Sparks TH, Tryjanowski P. 2013. Ecological correlates of the popularity of birds and butterflies in Internet information resources. Oikos **122**:183–190.