Digital data are accumulating at unprecedented rates. These contain a lot of information about the natural world, some of which can be used to answer key ecological questions. Here, we introduce iEcology (i.e., internet ecology), an emerging research approach that uses diverse online data sources and methods to generate insights about species distribution over space and time, interactions and dynamics of organisms and their environment, and anthropogenic impacts. We review iEcology data sources and methods, and provide examples of potential research applications. We also outline approaches to reduce potential biases and improve reliability and applicability. As technologies and expertise improve, and costs diminish, iEcology will become an increasingly important means to gain novel insights into the natural world.

Information Age, Big Data, and iEcology

The information age is characterized by rapid accumulation of myriad types of digital data [1]. Central to this revolution is the Internet, which is a source of unprecedented amounts of diverse and readily accessible data, via webpages, social media, and various other data platforms. These data are constantly created and stored in the digital realm and form an omnipresent part of the modern world. They also provide novel opportunities for research that the scientific community is only beginning to explore. Here, we describe an emerging research approach – iEcology (i.e., internet ecology), which we define as the study of ecological patterns and processes using online data generated for other purposes and stored digitally (Figure 1). These data can be used to address fundamental ecological questions and to analyze ecological processes at a range of spatiotemporal scales and across a diverse range of contexts. As such, iEcology has the potential to provide new understandings of ecological dynamics and mechanisms, complementing more traditional methods of obtaining ecological data.

While iEcology can be considered to fit within the wider scope of ecological informatics (see Glossary), it is distinct from other uses of Big Data sources in the biological sciences in that data are not specifically and intentionally generated to address ecological and environmental questions [2–4]. Moreover, iEcology expands on the traditional scope of ecological informatics with new data sources and dedicated methods to analyze them. iEcology is predominantly focused on collecting, collating, and exploring data generated online by human society, either passively or unintentionally (e.g., Internet search activity, social media interactions, and uploaded data and media), a process also referred to as passive crowdsourcing [5]. iEcology uses digital methods to access, handle, and analyze these data, in a manner akin to techniques from other research fields such as sociology, culture and media studies, biomedical sciences, computer sciences, and economics [6,7]. iEcology also shares part of its toolbox with conservation culturomics – an emerging research area in conservation science [8–10] – albeit with a different focus. Specifically, while conservation culturomics is
interested in understanding human engagement with nature, iEcology methods focus on the ecological knowledge that can be gained from these human–nature interactions in the digital realm. iEcology data predominantly give rise to insights that are correlative in nature, similar to other large-scale ecological explorations such as much of macroecology [11], and should be viewed as such.

Here, we present a broad overview and description of iEcology, including its scope, data types, sources and methods, as well as current major caveats and future prospects for the development of this emerging research approach.

Research Scope
Several recent studies have highlighted the potential of iEcology (Figure 2). The most common applications of such methods have been to explore species occurrences and their spatiotemporal trends (Figure 3). For example, a study comparing real-world encounter rates of bird species in the USA with Google Trends data found good agreement between the two sources (Figure 2A) [12]. This showcases the potential of using voluminous search engine data to explore species distributions in many regions. Others have explored species occurrences and distributions using various sources, such as Flickr, news articles, Twitter, YouTube, Facebook, and Google Trends [13–25], as well as population dynamics and phenology [14,20,23,26–31]. A particular illustration comes from assessing seasonal migration patterns of sockeye salmon (Oncorhynchus nerka) and Atlantic salmon (Salmo salar) from Wikipedia pageview frequencies (Figure 2B) [32]. In addition to mapping the distribution and occurrences of known species, images uploaded on social media have also been used to identify new species [33,34]. Trait dynamics, evolutionary trends, and biogeographic patterns can also be explored using iEcology methods. For instance, Google
Images were used to identify the presence and distribution of hybrid zones of hooded (Corvus cornix) and carrion (Corvus corone) crows in Europe (Figure 2C) [35]. Furthermore, spatiotemporal dynamics of biophysical environments, such as solar radiation and various other climatic parameters were characterized using Flickr tags [18].

iEcology sources, tools, and methods can also be used to explore biotic and abiotic interactions within and across species and their environments. For example, feeding patterns of yellow anaconda (Eunectes notaeus) and green anaconda (Eunectes murinus) were studied using online videos [36], while online images that simultaneously depicted African birds and herbivorous mammals were used to construct a web of associations between these two groups (Figure 2D) [37].

iEcology also provides new opportunities to study animal behavior [15]. For instance, YouTube videos have been used to compare the behavior of red (Sciurus vulgaris) and grey squirrels (Sciurus carolinensis) in different habitats (Figure 2E) [38]. The sheer volume and coverage of such sources could also prove fertile ground for identifying and tracking the spread of new behaviors [39–41]. Disease ecology, including knowledge of the occurrence, distribution, prevalence, and severity of diseases, has also recently benefited from iEcology methods [42].

iEcology methods have also been used to investigate ecosystem and habitat dynamics in response to increasing anthropogenic impacts. For example, videos of the Tour of Flanders cycling race from over 35 years have been used to track phenological changes to vegetation in response to climate change (Figure 2F) [27]. Images of corals and tweets referring to corals have both been used to evaluate the state and trends of coral reefs in different areas, suffering from various human impacts [43,44]. Aspects of invasion dynamics [14,45] and overexploitation of fish [29,30] have also been studied using image analysis, tweets, and news articles. In the same way, behavioral changes in animals in response to anthropogenic impacts [46–48] can be tracked by such methods.

While inherently varied in scope, other fields within ecology and environmental science could conceivably benefit from iEcology tools and methods, such as functional ecology, macroecology, landscape ecology, and urban ecology.

### iEcology Research Toolbox – Data Types, Sources, and Methods

At their core, iEcology data sources fall into two categories: (i) new data uploaded by users for different purposes; and (ii) data on online activity, including data access and search engine usage. Types of data within the first category can comprise text, images, videos, and sounds (Figure 1). The second category is aggregated data and the exploration of frequencies (e.g., the number of times a term was searched or a webpage visited, but could also include interactions on social media such as shares and likes). Both categories have different types of associated metadata that are particularly important for iEcology, such as locality, timestamp, user identity, and links across data.

iEcology data sources differ in their scope, availability, ease of access, associated metadata, and therefore utility for different types of research. Potential data sources range from various social media platforms (e.g., Twitter and Flickr) [49], search engines (e.g., Google, Baidu, and Bing), online encyclopedias (e.g., Wikipedia and Encyclopedia Britannica online), and other online repositories (blogs, discussion forums, popular articles, books, etc.). Many of these sources can also be accessed through search engines. The scope of sources differs based on spatiotemporal coverage, linguistic or cultural breadth, data resolution, and the degree of multimedia composition (e.g., text, images, and video) per source. Data also differ in availability: while many sources are freely available, some platforms may restrict availability by limiting data collection (i.e., limits on...
Sources also differ in their ease of access, from simple online tools embedded at the source (e.g., Google Trends webpage), through open application programming interfaces (APIs) accessible via various dedicated computer scripts (e.g., Wikipedia and Flickr), to APIs with volume, time frame, or number of queries) or use (e.g., privileged access or paywall restrictions).

Figure 2. Examples of iEcology Studies. (A) High level of correlation observed for ruffed grouse (*Bonasa umbellus*) encounter rate and spatial distribution of societal interest, based on Google Trends [12]. (B) Sockeye salmon (*Oncorhynchus nerka*, blue line; upper photo) and Atlantic salmon (*Salmo salar*, red line; lower photo) popularity based on Wikipedia pageviews reflects their seasonal migration patterns [32]. (C) Distribution of two crow species in Europe, carrion crow (*Corvus corone*, upper photo) and hooded crow (*Corvus cornix*, lower photo), indicated by Google Images, corresponds well with their actual distribution and hybrid zones [35]. (D) Quantitative bird–mammal association webs for non-oxpecker and oxpecker species in African birds and herbivorous mammals revealed by the analysis of Google Images [37]; upper photo – yellow-billed oxpeckers (*Buphagus africanus*) on zebra (*Equus sp.*), lower photo – yellow-billed oxpecker on Cape buffalo (*Syncerus caffer*). (E) The network visualization of the behavior of red (*Sciurus vulgaris*, left) and eastern grey (*Sciurus carolinensis*, right) squirrels assessed by YouTube videos [38]. (F) Phenological changes in vegetation as a response to climate change identified through archive videos of the Tour of Flanders cycling race [27]. See the Supplemental Information online for image attributions.
restricted access (e.g., Facebook). However, data availability and ease of access to different sources can also change over time.

The analysis of iEcology data faces similar challenges and uses the same solutions as many other approaches for analysis of Big Data [2,50]. Many of the methods used in iEcology rely on high levels of automation, frequently adopting machine-learning techniques [51]. There are different tools that can aid each stage of the research: data access, downloading, handling, extraction, storage, pattern identification and recognition, data analysis, and visualization. These tools are in a constant state of evolution, as illustrated by developments in deep neural network analysis and other emerging technologies (Box 1).

**Caveats and Solutions**

While holding remarkable promise, iEcology is subject to several inherent challenges and gaps that require careful consideration when undertaking such research (see Outstanding Questions). Primarily, it is important always to keep in mind that while ever increasing, the digital realm only encompasses a

---

**Figure 3.** Overview of the Studied Taxa, Data Sources Used, and Knowledge Categories Addressed by the iEcology Studies Cited in this Article. Colors represent different taxa, width of lines represents relative number of publications connecting different categories.
subset of the world – one that is nonrandom in extent and depth. Indeed, as the data are not generated systematically, there is great variance in content generation among different users, regions, cultures, and time frames, with inherent risks of biases [52]. Such individual and cultural subjectivity can further complicate data interpretation. Moreover, multiple entries of the same data by single or different users could cause biases related to nonindependence. Therefore, underlying data for iEcology research should neither be treated as randomly distributed, nor used in raw form without addressing these issues. Indeed, rather than ignoring such considerations, specific investigations into aspects of the data such as the nonrandom distribution and the level of nonindependence can actually provide further insights into data structure and any discovered patterns.

Several approaches, many already recognized within other fields of research that rely on online data, can be used to tackle these challenges. Validation with common and reputable sources such as systematic surveys, remote sensing, and citizen science (i.e., ground truthing) can decrease the level of associated uncertainty and help reinforce confidence in the data and their interpretation [12,44,53,54]. This is particularly important when testing new tools or approaches.
The vast majority of iEcology studies that we have identified have used multiple data sources to validate results, including data from field research, citizen science, online databases, scientific literature, or their combination [12,13,16–25,31,32,35,43,44]. In most cases, authors report a satisfying to excellent level of consistency among data sources. When ground truthing is difficult, as is often the case, other metrics could be developed to assess data robustness. We also strongly advocate cross-referencing results across multiple iEcology data sources, to test consistency of patterns [54,55]. Furthermore, culturomics can provide critical support to understand societal perceptions, interests, and values that affect the process of data generation [8,56].

Correct taxonomic identification in iEcology may be a cause for concern when compared with traditional ecological research. This may be true at several levels – from species misidentification by data producers to challenges that experts face when identifying species based on a limited number of images or videos of an individual organism. Furthermore, automated classification of species also generates misidentifications. Such embedded errors could also arise in other types of ecological data, such as life history traits, behavior, and abiotic variables. However, we expect that as iEcology sources increase in size, and methods to validate them improve, so will the ability to identify the extent and type of such problems in the data. Furthermore, we also suggest assigning a validity attribute to data that can be nonbinary, and dependent on the contributor’s reputation and the likelihood of an observation – as is currently practiced on some citizen-science platforms.

iEcology research would greatly benefit from collaborative efforts and sharing of data, resources, and tools. These could be aided by developing specific metadata standards for sharing such data, which could include APIs and specific machine-learning algorithms used to extract or manipulate the data. Such developments could draw from similar efforts that are already being carried out by big ecological databases (e.g., www.gbif.org) to develop similar standards, which would make ecological data more interoperable [57].

iEcology repositories could be either centralized or remain decentralized, with benefits associated with both options. Centralized repositories would greatly aid the maintenance of high standards, as well as providing better reproducibility, open access, and versioning. Nevertheless, necessary effort on pre- and postprocessing of data and metadata for uploading, and generally rigid structure of a centralized repository may actually deter people involved in more local, small-scale explorations, and may ultimately hinder data sharing and collaboration. As iEcology is still in its infancy, there may be some advantages in its remaining decentralized and more flexible at this current stage. Yet, as the methods, tools, and associated data increase in breadth and scope, a move towards more collated, managed, and centralized repositories will become more natural and pertinent. Nevertheless, we advocate that good record keeping and maintaining high metadata standards is of particular importance to iEcology.

Other considerations of iEcology data sources involve interpretation and reproducibility. Some sources lack transparency in the way the considered data were produced and manipulated (e.g., search engines such as Google). Inability to publish raw data (as per provider guidelines) could also cause issues with scientific journal protocols that require making these available. Furthermore, some sources lack stability in data scope, underlying algorithms, and access options. These are inherent issues with many online sources. To alleviate these concerns, we advocate: (i) good record keeping of protocols for data access, handling, versioning, and
analysis; (ii) harmonization of methods [58] and standardization of metadata; (iii) publishing (when possible) raw data in freely accessible and stable repositories together with associated scripts; (iv) use of open-source data and software; and (v) keeping up to date with methodologies developed in other relevant fields (e.g., computational sociology) for assessing and addressing such issues.

iEcology research may give rise to several ethical issues, pertaining to both people and nature. Data shared online, especially on social media platforms, sometimes include explicit personal information, while implicit information could also be used to identify individuals or to extract sensitive information. Therefore, the privacy of individuals and their identifiers should be maintained in both data repositories and iEcology outputs, adhering to the highest ethical standards [59]. Moreover, data sources that include precise information on locations and other key attributes of rare or endangered species could increase their exposure to poachers and collectors [60]. This threat could be alleviated by either restricting access to data on species deemed at risk, or limiting precision of open-access information. In general, servers holding iEcology data should be securely maintained to avoid such abuse.

Concluding Remarks and Future Perspectives
The field of ecology is undergoing a rapid shift towards indirect, technology-based, and automated observations of nature and biodiversity [53,61,62], where iEcology is likely to play a critical role. Utilization of iEcology methods and data has greatly expanded recently, with most publications appearing during the past few years. Ecology is likely to experience rapid development over the coming years and become one of the major research techniques in ecology. While classical biodiversity research is irreplaceable for understanding the natural world through targeted observations and experiments, iEcology could provide novel and low-cost support for ongoing research efforts. The value of iEcology is likely to greatly increase as the global coverage of the Internet, mobile computing, sensor networks, and their users, expand. This will be augmented by leaps in computational capabilities, emergence of new data sources and types (such as odors, obtained by electronic noses) [63], and other emerging tools and technologies for using these data (Box 1). Combining these with other Big Data sources and efforts to understand nature, such as ecoinformatics, could also prove valuable.

In the near future, we foresee complete automation of all stages of data handling within iEcology, from access to visualization, to creation of ever-expanding datasets of biological entities, traits, behaviors, etc. This could give rise to a global digital monitoring initiative for the natural world. For example, an ecologist interested in animal behavior could produce tools to automatically scrape all uploaded YouTube videos for animal representations, automatically analyze them for different types of behaviors, include these into a constantly updated dataset, and ultimately analyze them in real time, to produce continuously updated research outputs. However, good expertise regarding the organisms studied and underlying ecological mechanisms at play will always be invaluable to make sense of these rapidly accumulating data and their inherent biases. Furthermore, many of the examples presented above demonstrate imagination and creativity in using data sources for ecology that were collected for other purposes. Above all, iEcology will benefit from such creativity to find new ways to harness data beyond their original purposes.

iEcology provides fertile ground for interdisciplinary collaborations, enhanced by a wide range of expertise and specializations. Furthermore, iEcology will create new opportunities for partnerships between academia, industry, governmental, and non-governmental organizations, working synergistically to produce original insights into the natural world.

Outstanding Questions

How do iEcology insights differ from those uncovered by other ecological methods (in scope, reliability, applicability etc.)?

Can insights and observed patterns from regions with high digital coverage be generalized to those without?

How should we attach uncertainty to data from iEcology sources?

Should we aim for centralized or decentralized repositories and datasets?

Are there particular skills that should be taught to students to help them develop iEcology expertise and how can these be integrated in curricula?

Are Linnean or Wallacean shortfalls manifested in iEcology sources similar to classical ecological sources?

Will the future development of iEcology cause greater detachment from nature even among naturalists or ecologists and how could this be averted?

Will the use of iEcology alleviate some of the ethical concerns of handling animals, and will it give rise to new ethical concerns?
Acknowledgments

I.J.’s work was supported by the J. E. Purkyňě Fellowship of the Czech Academy of Sciences. F.C. was supported through the Invacost grants by the ANR, the AXA Research Fund Chair for Invasion Biology and Biodiversa AlienScenarios. R.A.C. acknowledges funding from the Netherlands Organization for Scientific Research (NWO) and from Social media data sources. F.C. was supported through Invacost grants by the ANR, the AXA Research Fund Chair for Invasion Biology and Biodiversa AlienScenarios. R.A.C. was supported by a research fellowship from Merton College and the Israel Science Foundation (grant No. 406/19). J.A.F. was supported by a research fellowship from Merton College and (grant agreement #802933). A.T.S. acknowledges the EU Horizon 2020 research and innovation programme (grant No. 406/19). J.A.F. was supported by a research fellowship from Merton College and (grant agreement #802933). A.T.S. acknowledges the EU Horizon 2020 research and innovation programme (grant agreement #802933). A.T.S. acknowledges the EU Horizon 2020 research and innovation programme (grant agreement #802933). A.T.S. acknowledges the EU Horizon 2020 research and innovation programme (grant agreement #802933).
