Identifying the identifiers: How iNaturalist facilitates collaborative, research-relevant data generation and why it matters for biodiversity science

C. J. Campbell (ID), Vijay Barve (ID), Michael W. Belitz, Joshua R. Doby, Elizabeth White, Carrie Seltzer, Grace Di Cecco, Allen H. Hurlbert and Robert Guralnick (ID)

C. J. Campbell (caitjcampbell@gmail.com) is a PhD candidate in the Department of Biology and a fellow of the University of Florida Biodiversity Institute at the University of Florida, in Gainesville, Florida, in the United States. Michael W. Belitz is a PhD candidate, Joshua Doby and Elizabeth White are PhD students, and Robert Guralnick is a curator of biodiversity informatics at the Florida Museum of Natural History, in Gainesville, Florida, in the United States. Vijay Barve was a postdoctoral associate at the Florida Museum of Natural History, in Gainesville, Florida, and is presently a digitization project manager at the Natural History Museum of Los Angeles County, in Los Angeles, California, in the United States. Grace Di Cecco is a recent PhD graduate student who worked with Allen Hurlbert, a professor in the Department of Biology at the University of North Carolina, in Chapel Hill, North Carolina, in the United States. Carrie Seltzer is the stakeholder engagement strategist at iNaturalist, California Academy of Sciences in San Francisco, California, in the United States.

Abstract

The iNaturalist platform generates millions of research-grade biodiversity records via a system in which users collectively reach consensus on taxonomic identification. In the present article, we examine how identifiers and their efforts, an understudied component of the platform, support data generation. Identification is keeping pace with rapid growth of observations, assisted by a small subset of highly active users who tend to be taxonomically specialized. Identifier experience is the primary determinant of whether records reach research grade, and the time it takes to do so. Time to reach research grade has fallen rapidly with growing identification effort and use of computer vision, and research-grade identifications are generally stable. Most observations are vetted by experienced identifiers, although identifications are not free of biases. We close by providing suggestions for enhanced identification quality and continuing steps to enhance equitable credit and trust across the ecosystem of observers, identifiers, and data users.

Keywords:

Field-based community science, also called citizen or participatory science (Bonney et al. 2014, Brown and Williams 2019), already supplies the vast majority of digital knowledge about species occurrence for many regions and taxa (Amano et al. 2016). The rise of global digital platforms for sharing observations of organisms, such as iNaturalist (Di Cecco et al. 2021), is further accelerating the growth of knowledge about all diversity and its distribution across the tree of life, rather than select charismatic clades, such as birds. At the core of public involvement in ecological monitoring is the critical process of identifying observations to the finest possible taxonomic level, typically the species level. Taxonomic identification is a process that unites a specimen, specimen derivative (e.g., photograph), or observation with a collection of taxonomic concepts as input and produces an output with taxon designation associated with the specimen (Deck et al. 2015). Although this process is foundational for the use of species occurrences in biodiversity science, it is also one of the most complicated and challenging, especially in community science projects, in which varying levels of taxonomic expertise among the participants can lead to erroneous identifications (Hochmair et al. 2020, McMullin and Allen 2022). Identifying an observation to species relies on users having the skills and knowledge to process key characteristics, including diagnostic anatomical features, such as shape and size, along with other aspects of form and behavior, and linking those characteristics with a mental list of species that might be found in the area (Kelling et al. 2012).

Many community science projects have devoted significant effort to ensuring that their identifications of target taxa are of the highest quality possible in order to assure best downstream use. One of the most successful community science projects, eBird (Sullivan et al. 2009), uses multiple mechanisms for generating quality identifications. eBird asks volunteers to report a checklist of birds observed in a user-defined region over a user-defined survey length (Kelling et al. 2019), tied to eBird's preferred taxonomy. When checklists are submitted, further quality checks are performed to flag unusual sightings. eBird also captures information about observer skill using objective characteristics, such as the total number of checklists submitted, the number of species reported, and the total number of rejected flagged records (Kelling et al. 2015). This information can be used to filter or weight lists in downstream use.

Unlike community science efforts, such as eBird, that are restricted to a focal taxon, more taxon-agnostic community platforms have the challenge of managing species occurrences from across the tree of life (Di Cecco et al. 2021). This expands by many orders of magnitude the scope of taxonomic groups covered and greatly increases the challenge of properly identifying those occurrences. iNaturalist, one of the fastest growing and most popular of these taxon-agnostic platforms, asks users to upload a photograph or audio recording of the organism and share key metadata about location, date and time, and taxonomic identification. As of November 2022, nearly 2.5 million iNaturalist

observers have reported more than 135 million species occurrences, almost all of which are backed with a digital voucher, such as a photograph or audio recording.

The key to iNaturalist producing usable data for downstream scientific or management use is a unique community identification process. This process relies on the community of iNaturalist users to help identify observations. These identifiers are far less numerous than the observers (and, often, users are both) by nearly an order of magnitude (approximately 280,000 users who have made at least one identification as of November 2022) but play a critical role. For an iNaturalist record to move from a "needs identification" observation to a research-grade observation, the record needs at least two identifiers to agree on a species-level or finer taxonomic identification. Often, but not always, the original observer provides one of these identifications. Continued disagreement among identifiers can also mean that records move from research grade to "needs identification," with the bar set at greater than two-thirds agreement among all identifiers to keep research-grade status. Although this identification model is relatively simple, it has profound implications both for the overall quality of iNaturalist data and for spatial, taxonomic, and temporal biases in distribution of observations that reach research grade and those that do not.

Recent work by Di Cecco and colleagues (2021) provided a synthetic snapshot of how observers and their effort structure the spatial, temporal, and taxonomic coverage of species occurrences in iNaturalist while also categorizing observers as taxonomic specialists or generalists. But a full understanding of the utility of iNaturalist data for biodiversity science rests fundamentally on better understanding identification processes, because only research-grade observations are relevant for the vast majority of downstream science or management uses.

In the present article, we use the whole corpus of iNaturalist records from its inception through the end of 2021 to address key questions about identifiers and the identification process. We first explore questions related to patterns of identifications by users and how their rates and allocations play out across time, user activity, and in the context of the introduction of computer vision machine learning algorithms that assist in identifications. Given the growth of iNaturalist activity and its user base, we were particularly interested in whether identifications are scaling with the rapid growth of observations uploaded to the site. We then explore the patterns of identifier specialization across the axes of geography and taxonomy. Because research-grade observations on iNaturalist are published to the Global Biodiversity Information Facility (GBIF) for broadest use by the biodiversity research community, we were particularly interested in the proportion of identifications coming from taxonomic specialists. Next, we applied a hierarchical modeling framework to map the paths of observations from the default quality of "needs identification" to a vetted research-grade observation, testing key characteristics that determine whether an observation reaches research grade and how long it takes to get there. We were interested in the roles of identifier experience, geography, and the use of computer vision in affecting identification status. We close by discussing methods to decrease gaps and biases by incentivizing identification activity, best practices for enhanced identification quality, and continuing steps to further grow trust across the ecosystem of observers, identifiers, and data users.

Gathering key information about observations and identifications

All iNaturalist data from the first records posted in March 2008 up to the end of 2021 were downloaded by date of posting for each day using the get_inat_obs() function from the R package rinat (Barve and Hart 2022). The resulting data were stored in RDF format, from which relevant information on occurrence records and identifications provided by users were extracted and organized into monthly .csv files. Each downloaded record has identification histories through to the end of 21 January 2022. We then imported the monthly .csv files containing observation and identification records into a SQLite database. From iNaturalist's underlying taxonomy we derived rank level (a numerical description of taxonomic level, with a higher number corresponding to a coarser taxonomic resolution) and higher order taxonomic categorization for 99.7% of all observations and identifications made through 2021. The data assembled for our analyses are held at https://doi.org/10.17605/OSF.IO/J752N; the code used for the subsequent analyses is archived at https://doi.org/10.5281/zenodo.7681468 and is available at https://github.com/cjcampbell/iNaturalist-Identifiers. The analyses were conducted in R version 4.2.2 (R Core Team 2022), and we relied particularly on packages DBI (R Special Interest Group on Databases et al. 2021), tidyverse (Wickham et al. 2019), and data.table (Dowle and Srinivasan 2021).

After all the data were fully assembled, we extracted key metrics for each user, including the number of observations, the number of identifications, and the number of taxa identified. For users with more than 100 observations, we estimated their locality using the centroid of their observations and measured the distances from their locality to 5000 of their randomly sampled identifications. For each observation, we assessed changes to its quality grade with respect to each identification. We use iNaturalist's definition of research-grade status (i.e., more than two-thirds agreement among all the identifications at or below the species level), although we were unable to account for some additional data quality criteria including data and location accuracy flags that might shift observations to casual grade. To verify that the identification histories we reconstructed were correct, our team also hand checked 1000 user and 2500 observation histories, comparing the compiled data we collected with iNaturalist tallies. We found broad consistency, with slight differences due to further identification effort collated by iNaturalist past the date of our records. Finally, we also confirmed that taxon identifications and ranks and computer vision data were accurate for the subset of identifications. The summarized data were used in further analyses, as is described below.

Overall patterns of iNaturalist identifier activity

Our initial questions were focused on identification effort. We first asked whether the identification effort keeps pace with the observation effort, especially given the accelerating rates of observations (Di Cecco et al. 2021). Even though there is a nearly tenfold smaller number of identifiers than of observers (approximately 280,000 identifiers to approximately 2.5 million observers as of January 2023), identification activity is keeping pace with observations through time (figure 1a, b). Simultaneously, the median time to research-grade quality decreased precipitously from over 1 year in 2011 to just above 4 hours a decade later in 2021 (figure 1c). This increase in identifier effort is all the more extraordinary because almost all the work is done by highly active identifiers: 75% of identifications are performed by the top 1% of the pool of all participants (figure 1d), which includes only approximately 2000 individual users. These results are even more right-skewed (i.e., more effort concentrated in fewer hands) than for observers; Di Cecco and colleagues (2021) showed that approximately 62% of all observations are made by the top 1% of the
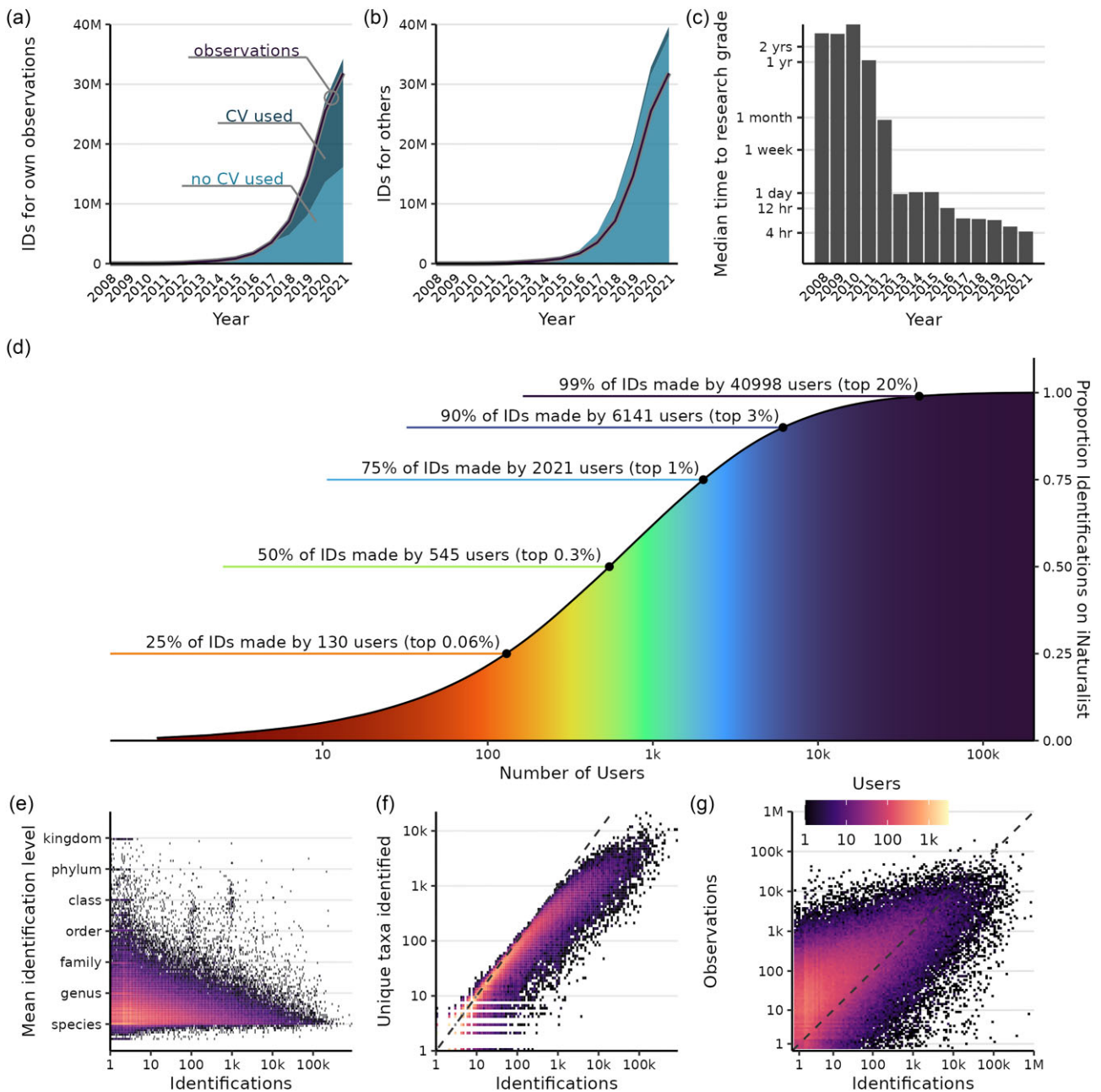
Figure 1. Summaries of identification effort on iNaturalist. Panels (a) and (b) summarize identifier activity (the shaded area) relative to the number of observations added to the site (the bold line), as a function of year. The identifications that were aided by computer vision are indicated by a darker shade. Panel (c) documents the median time per year since 2008 of observations first reaching research-grade status. Panel (d) shows accumulation plots of identification effort over different subsets of identifier activity, showing that the vast majority of identifications are made by a very small subset of users. Panel (e) shows that more active identifiers are more likely to identify to finer taxonomic levels (e.g., to species). Panel (f) highlights that the majority of users, including those with fewer than 1000 observations, identify close to 1 new taxon per identification, whereas the most active identifiers identify proportionally fewer taxa identified per identification (the dashed line indicates a one-to-one relationship). Panel (g) shows overall identification and observation activity for all users, with fewer users identifying more than observing (with a one-to-one relationship indicated with a dashed line).

pool. Highly active identifiers have different profiles of identification activity than do other identifiers, including a tendency toward specializing on specific groups and identifying to the species or lower level relative to less active identifiers (figure 1e, f). Finally, users who are active identifiers also tend to be active observers, but this is highly variable, with a trend toward highly active identifiers contributing more identifications than observations (figure 1g).

One of the key innovations for iNaturalist was the incorporation of computer vision in 2017 into its identification process. Both observers and identifiers can use computer vision to see suggestions for the identification of an observation. Users can simply click a button to get a list of candidate species with visual similarity, ranked by those observed in the area, and can select among them. The use of computer vision can provide direct assistance in identification by suggesting candidates but can

also increase the ease of use by saving time on recall and typing. Since computer vision was added in 2017, it has quickly become a critical means for early identification of an observation, accounting for nearly half of the observing users' identifications by the end of 2021 (figure 1a). However, beyond an initial identification by an observer, other subsequent identifiers are much less reliant on computer vision, as is shown in figure 1b. Computer vision may also have some role in reducing the time it takes to hit research grade (see more below), which has decreased markedly over time (figure 1c). However, the most noticeable shifts toward faster time to research grade occurred prior to computer vision's introduction, potentially when a critical mass of active identifiers became active on the platform, also likely hastened by an efficient "identify" user interface introduced in 2016 (Ueda 2016).

Geographic and taxonomic patterns of identifications or identifiers

Biases in usable data from community science projects remain a key concern because rapidly increasing data resources may reduce some types of data gaps but ultimately increase spatial, temporal, and taxonomic biases (Shirey et al. 2021). Identification processes can contribute to these biases, because even if the observations are made in regions that are otherwise poorly sampled, there must still be effort to assess identification quality before these data are usable in research. Our working hypothesis is that identifiers are typically less constrained geographically than observers are, because they work remotely and are therefore freed from the costs associated with collecting observations. We measure geographic specialization as the percentile rank of users according to the standard deviation of the distance from inferred residence to identification (measured as the distance from a centroid of observations to 5000 random identifications along an ellipsoid, ranked among users with at least 100 identifications). We measured taxonomic specialization as the percentile rank of a user's number of identified taxa over their number of identifications (also among users with at least 100 identifications). From both percentile ranks of specialization or generalization, we considered users "highly specialized" if they were in the bottom quartile, "specialized" if they were in the second, "generalized" if they were in the third, and "highly generalized" if they were in the top quartile. We also expected that more active identifiers are likely to be taxonomic specialists, defined as focusing their identification efforts more narrowly across the tree of life, given that identification is facilitated by taxonomic expertise.

Identifiers are commonly geographically and taxonomically specialized; however, there is significant variation along both axes (figure 2). The most active identifiers are taxonomic specialists, with similar numbers identifying geographically broadly and narrowly (figure 2a, b). Most striking is the trend toward most identifications coming from taxonomic specialists (84% of the identifications from the top quantile of taxonomic specialists and 96% from the top half; figure 2b). We also note that there is still significant variation in identifier approach, exemplified by a selection of different identifiers in figure 2c–f, showing values along key axes (percentile rank of the number of identifications, the mean distance to identifications, the geographic generalist or specialist, and the taxonomic generalist or specialist) and the actual geographic patterns of identification.

Identifier experience is essential for reaching research-grade status

Our work thus far summarized identifier patterns over space and time, but we had yet to identify which key factors determine whether and how long it takes for observations to reach research grade. We built two linear mixed models to statistically test these questions, controlling for the length of time observations were posted on iNaturalist and were therefore subject to potential identification. The first model had a binary response of whether an observation had reached research grade or not, and our predictor variables included the time the observation was active on the platform, whether computer vision was used at the initial upload, the cumulative experience (the number of identifications) of all identifiers on the observation, the taxonomic rank of the observation when identified at upload, and the number of observations that have been recorded in the biome of the observation (Olson et al. 2001). This model was fitted to a binomial error distribution using a logit link function and included a random intercept for the region (defined by the United Nations) to control for unmeasured geopolitical factors that may influence whether an observation hits research grade. We used a variable selection approach for our model using a L1-penalized (LASSO) estimation, which uses a gradient ascent algorithm to reestimate a model that includes only the variables corresponding to the nonzero fixed effects (Groll and Tutz 2014). We did not have the memory to run our gradient ascent approach on all data points and therefore ran the procedure on 10 subsets of 10,000 random data points sampled with replacement. If covariates were retained in eight or more of the iterations, those were included in a single top model that was fitted to a larger random subset of data (934,246 observations). The results from that final top model are shown in figure 1.

In our second model, we created subsets of records including only those that reached research grade, and we tested which predictors influenced how long it took to get to research grade. The predictor variables included the time the observation was active on the platform, whether computer vision was used at the initial upload, the cumulative identification experience (the number of identifications) of all users who identified the observation prior to it hitting research grade, the taxonomic rank of the first identification, the number of observations that have been recorded in the biome in which the observation was observed, the number of observations on iNaturalist in the genus of the observation, and the number of observations in the family of the observation. We included the same random intercept for UN regions and used the same gradient ascent approach described above. The final model was fit to the 604,417 random observations that were classified as research grade.

The model's results suggest four key take-home points (figure 3). First, identifier experience was the most important predictor of whether an observation reached research grade and was far more important than whether computer vision was used, which also positively affected the likelihood of reaching research grade. Second, more identifier experience was also crucial for decreasing the time it took for an observation to hit research grade. The taxon rank of the first identification, a parameter lowered substantially by the use of computer vision, was another key factor in reducing the time to research grade. Third, the records reported at higher taxon ranks were less likely to reach research grade and took more time to get there if they did reach it. Fourth, we found a lot of variation across regions, with generally fewer records reaching research grade and with those that did taking longer in Asia, Africa, and South America than in Europe,
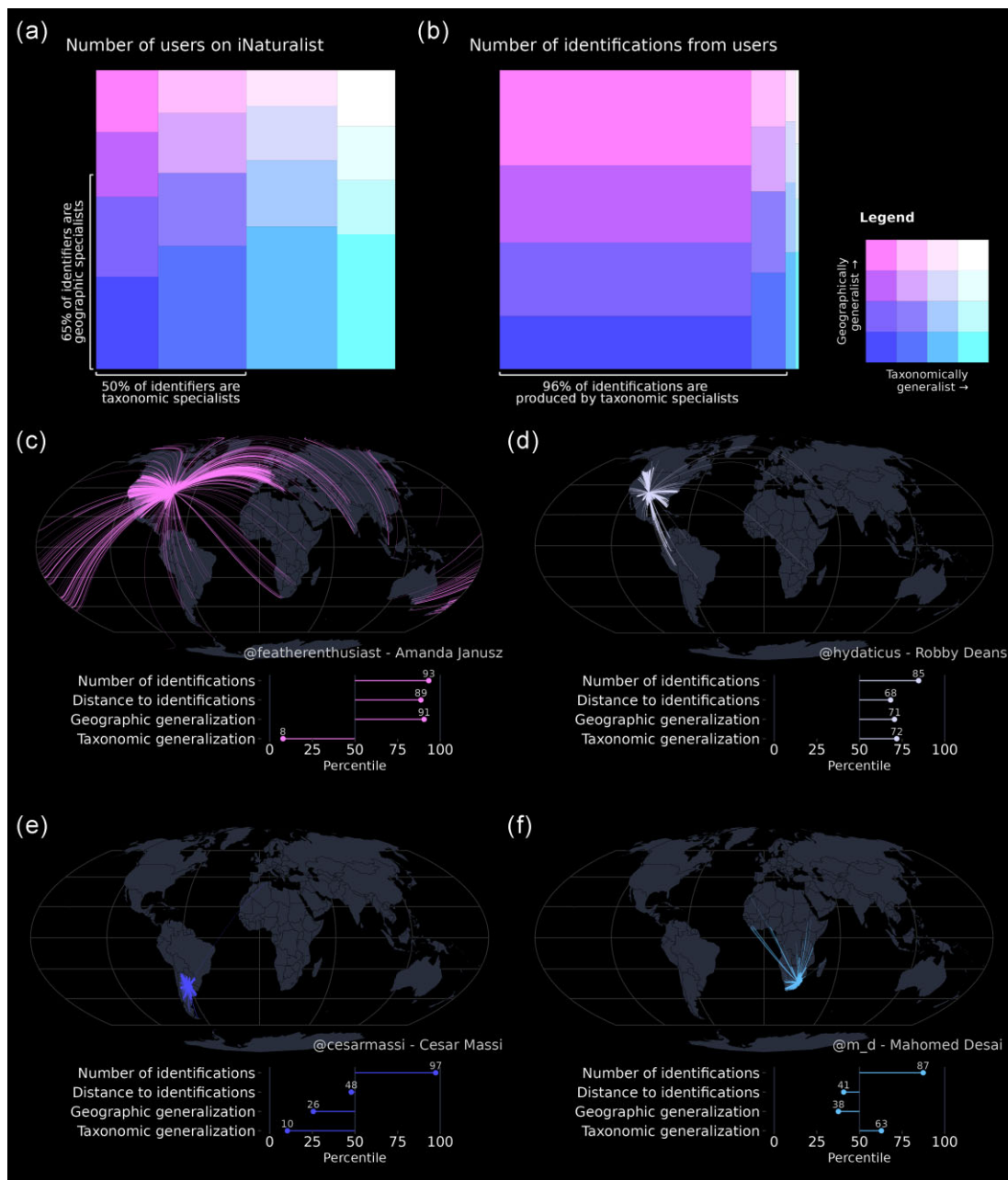
Figure 2. Tree maps (a, b) representing the proportion of identifications and identifiers that are taxonomically or geographically specialized (the horizontal and vertical axes, respectively) for identifiers with more than 100 identifications. The area represented by taxonomic specialists through generalists runs along the horizontal axis; the area of geographic specialization runs along the vertical axis (both move toward increasing generality; see the legend). Panels (c–f) showcase four different users with many identifications that vary in distance to identifications and geographic and taxonomic specialization; above, the user's percentile rank of each value is indicated by distance from the mean (50th percentile); below, the paths from observation centroid to 5000 random identifications are mapped.

Australia and New Zealand, or North America. There were some exceptions (e.g., West Africa, which might be driven by the overall low numbers of observations; note the relatively large confidence intervals on the estimates for West or North Africa).

Identification pipelines produces stable, reliable taxonomic records

How does the status of an iNaturalist observation change after being uploaded to the platform? Above, we were able to deter-

mine which factors enhanced the likelihood that observations made it to research grade. But iNaturalist records can also have other outcomes, including being demoted to a coarser taxonomic level or deleted. We again used a linear model framework—in this case, using the binomial error distribution and logit link function (i.e., logistic regression) to determine the fate of observations over the course of a year. The status of an observation was the response variable, and the amount of time active on the platform was the predictor variable. We used all of the downloaded observations (104,092,966 in total) to fit this model.
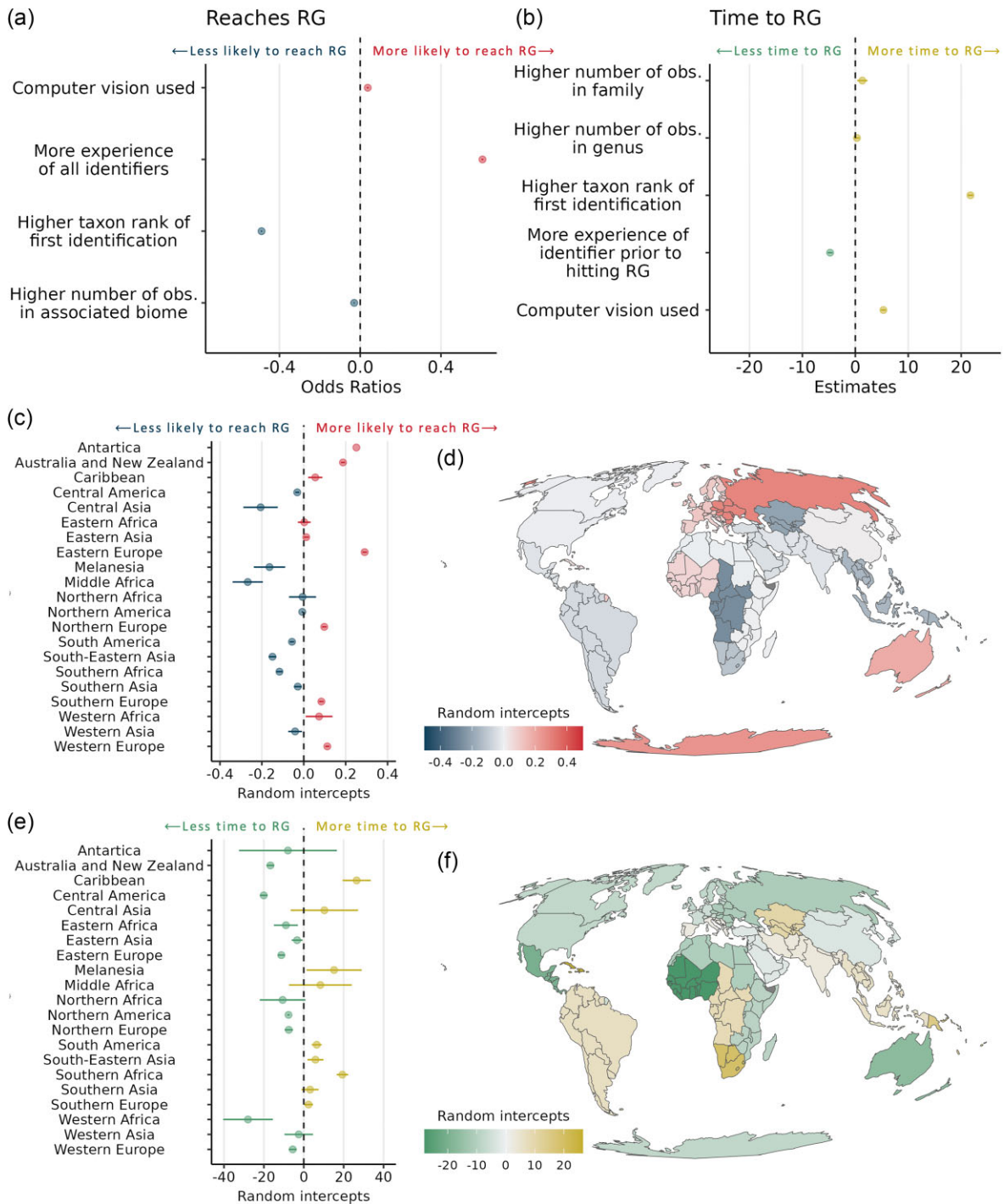
Figure 3. Model results showing top factors determining whether an observation reaches research-grade status (a) and time to research-grade status (b). Several top factors are related to the volume of iNaturalist observations in a region (e.g., "Higher number of observations in associated biome") or taxonomic group (e.g., "Higher number of observations in family"). Random effects of United Nations region were included in each model; the random effect estimates for likelihood to reach research grade are indicated in panel (c) and mapped in panel (d), and estimates for the time to reach research-grade status are indicated in panel (e) and mapped in panel (f).

Our model results are presented visually in figure 4. This shows that, over the course of a year, the majority of observations hit research grade—and almost all that made it remained there. Among the 3% of records that change within a year after reaching research grade, nearly half of those were due to taxonomic reclassifications (swaps) or identification to a finer taxonomic level (e.g., subspecies), which, ultimately, are improvements. This leaves a small proportion of records (1.6%) that do leave research grade, some of which eventually return (approximately 0.4%). A significant proportion of observations (28.4%) remain in the "needs identification" category, likely in part because they are not identifiable and in part because of a lack of identifiers. A sizable minority of observations (approximately 13%) become inactive, because they were either annotated to a lower quality grade (i.e., "casual") or
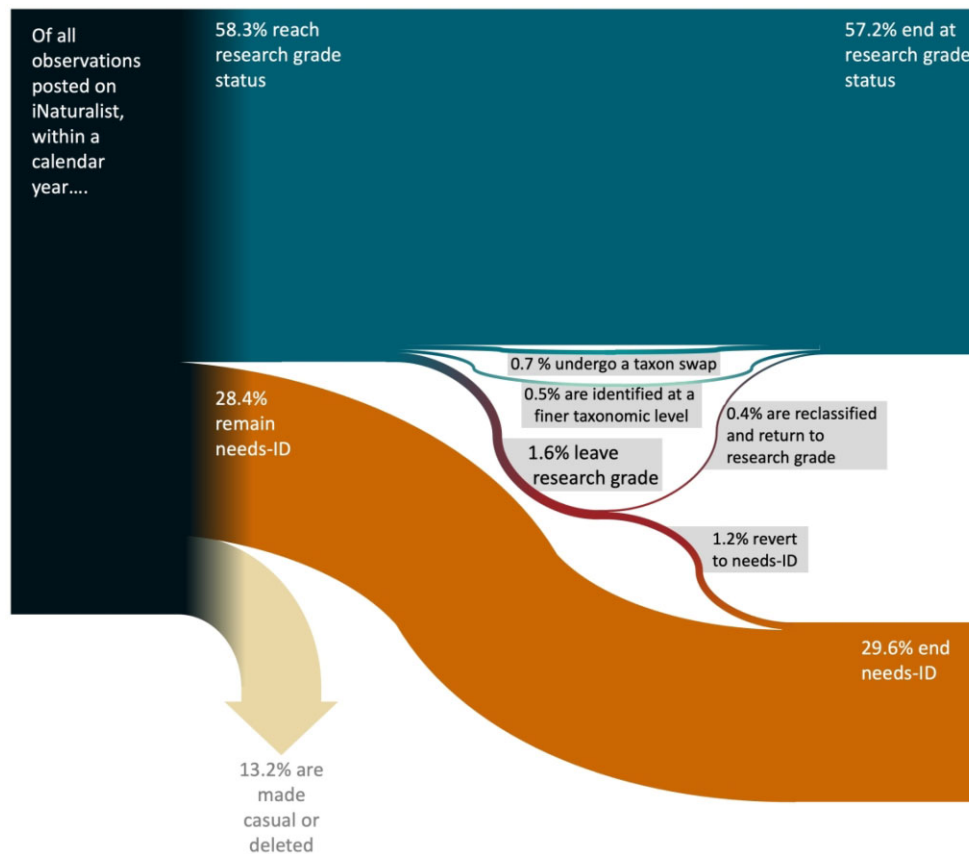
Figure 4. Sankey plot showcasing the fate of observations with regards to identification during a calendar year. The majority of records hit research-grade status and stay there, with some amount of flux due to taxon swaps, new identifications that remove records from research grade, and some that then return back to research grade.

deleted from the site. The key take home from this analysis is that most records on iNaturalist reach research grade and stay there, which implies that identification pipelines generally produce stable and likely reliable taxonomic records.

Why does this matter for biodiversity science?

One of the most essential aspects of iNaturalist is the collaboration between those documenting organisms and those helping to identify them. This collaboration is multifaceted in terms of benefits (Mesaglio and Callaghan 2021, Callaghan et al. 2022). Observers are seeking either a confirmation that their initial identifications are correct or a more refined identification at a finer taxonomic level. Therefore, identifiers provide a service and motivate observers to further add new content and become more proficient in recognizing what they see. Identifiers also fundamentally drive the research value of iNaturalist because identification confirmations are required for a record to reach research-grade status, the level at which most researchers use in their analyses. This value is made more explicit via the connections between iNaturalist and GBIF, which publishes all the research-grade observations as a data set that includes the identifier who first added the eventual research-grade identification on iNaturalist (iNaturalist contributors 2023). Because that identifier is included with each record, GBIF also provides one key mechanism to support credit to identifiers via its provision of digital object identifiers for downloaded data sets (www.gbif.org/citation-guidelines). We advocate

for a more inclusive identification credit model in which all identifiers of research-grade observations are listed in GBIF metadata. Finally, identifiers are critical for providing the labels that are needed for training computer vision models that support automating identification, which, itself, is part of a virtuous feedback loop. Many observers use computer vision taxonomic predictions in their first identification, and some identifiers also use these models, although our results suggest this is less common than we expected. We cover in more detail below identification stability and quality, biases in identifier efforts, and how computer vision has and may continue to affect the identification processes, before closing with more on the importance of iNaturalist's open and inclusive process for building a community of users.

Identification stability and quality

Identification processes sit at the center of an ongoing question of the fitness for use of iNaturalist data as the resource continues to rapidly grow. The research community has not wholeheartedly embraced using iNaturalist records in downstream research, and there is something of a cottage industry developing to assess identification quality for various taxonomic groups (Goodwin et al. 2015, Hochmair et al. 2020, Barbato et al. 2021, Koo et al. 2022, McMullin and Allen 2022, Rosa et al. 2022). This mistrust of iNaturalist data is not entirely unfounded; there are many cases in which research-grade digital vouchers are misidentified or simply cannot be identified to species given the quality of the photograph (or other media captured). One of the key results from our work that should allay some concerns regarding identification

quality is that the research-grade pipeline is remarkably stable. Most observations that are identified to research grade stay in that category, including those that get highly active identifier attention. We also note that many of the top identifiers who contribute the majority of identifications on the site are classically qualified taxonomic specialists. Because iNaturalist hosts such a high volume of user activity and promotes consensus-based identifications, the chance for revisions toward correct identification may be higher than in natural history collections, where specimen misidentifications are not uncommon. Furthermore, taxonomic updates are far less frequent in natural history collections databases than in iNaturalist, which means that there is a high number of specimens in natural history collections without currently valid names—partially contributing to misidentification estimates as high as 50% (Goodwin et al. 2015). The living nature of iNaturalist records and the frequent updates of existing records to the GBIF pipeline provide a mechanism for ongoing correction and revision of both observation and backbone taxonomies.

We still recognize the need for further vetting of research-grade specimens, and downstream users concerned about identification quality have a number of options depending on their tolerance for error. For many applications, we simply recommend accepting an often very low error rate and directly using research-grade records without further filtering. For those who require more stringency, we propose a couple of simple filtering approaches: filtering data to observations verified by trusted identifiers, or filtering out records that are identified once by a user and have only one identification confirming that identification. Both of these recommendations will help remove cases for which there are most likely to be misidentifications. However, such filters may remove many well-identified and perfectly usable data (Gaier and Resasco 2023).

Identifier taxonomic and geographic specialization

Di Cecco and colleagues (2021) showed that observation effort is highly spatially and taxonomically uneven and that most observers observe locally and tend to be specialists. We find that identifiers in general are much less likely to be constrained to local geographic areas, although there is quite a bit of variation in their geographic reach. The most active identifiers are as likely to be geographically specialist as generalist. Because many identifiers are less geographically specialized, most parts of the world are getting identifications and identifier biases are not strongly driving geographic unevenness. The bottleneck, at least at the scale of all iNaturalist records, is likely due to limited observations in many regions of the world.

Also unsurprisingly, identifiers are often more likely to be taxonomic specialists than observers, and this means that there are likely taxonomic groups on iNaturalist with some observation effort but limited identifications. Part of this lack of coverage may be due to limited expertise on the platform but equally as much is likely due to groups for which identifications by photograph are more difficult. Underlying all these results is a strong feedback loop between observers and identifiers. Regions and taxa for which identifiers are active are likely to attract observers because they know that their digital vouchers will be seen and improved (Mesaglio and Callaghan 2021). Finally, it may be possible that continued growth of computer vision approaches helps shift the needle in interesting ways. Although some taxa are hard to identify to species because of a lack of traditional diagnostic characters being visible, there is some indication that automated approaches can still provide accurate identifications even when traditional diagnostic features are not fully visible in photographs (Vidal et al. 2021). Continued exploration of how to improve identifications for certain taxonomic groups that are underrepresented (e.g., many small insects, macrofungi) will likely prove fruitful and may overcome barriers currently limiting more research-grade observations of these groups.

The importance of computer vision in identification

One of the unique aspects of iNaturalist was its early adoption of machine learning–based computer vision to help with identification across the platform. Computer vision is primarily used by a rapidly growing proportion of observers identifying their own observations, rather than by those making identifications. However it is deployed, its use increases the likelihood of an observation hitting research grade and presumably supports initial identification at a finer taxonomic level—one of the most important factors in shortening the time to research grade. Crucially, the computer-vision algorithm is trained on research-grade observations, themselves produced by iNaturalist identifier activity. The iNaturalist computer-vision system provides one of the key entry points to iNaturalist observers through its quick and user-friendly program and is so useful that it's been used as the backbone to a sister mobile app called Seek. However, by far the single most important factor supporting the movement of observations to research grade is having observations viewed by expert identifiers. This effect is far more important than computer vision for now, although the increasing use and accuracy of computer vision may shift this in the future. Along with computer vision, iNaturalist continues to make improvements to the tools they provide identifiers, which may also increase efficiency (Callaghan et al. 2022).

iNaturalist supports an open and inclusive community

Di Cecco and colleagues (2021) and others (Mesaglio and Callaghan 2021) have noted the many uses of iNaturalist data in support of biodiversity and allied domains, along with some of the pitfalls and challenges with data biases along multiple dimensions. Identification processes are crucial for the utility of iNaturalist to the broader research community but are still subject to biases resulting from limited taxonomic expertise in certain regions and taxa. Still, these same issues plague all biodiversity science, and iNaturalist provides an open and friendly platform for experts and novices alike to bring their love of natural history and help each other for the benefit of all (Harmon 2022). We hope that models for allocating credit for the amazing work done by iNaturalist identifiers can be further increased and valued in multiple contexts into the future.

Acknowledgments

References cited

Amano T, Lamming JD, Sutherland WJ. 2016. Spatial gaps in global biodiversity information and the role of citizen science. *BioScience* 66: 393–400.

Barbato D, Benocci A, Guasconi M, Manganelli G. 2021. Light and shade of citizen science for less charismatic invertebrate groups: Quality assessment of iNaturalist nonmarine mollusc observations in central Italy. *Journal of Molluscan Studies* 87: eyab033.

Barve V, Hart E. 2022. Package "Rinat." R Foundation. https://cran.r-project.org/web/packages/rinat/rinat.pdf.

Bonney R, Shirk JL, Phillips TB, Wiggins A, Ballard HL, Miller-Rushing AJ, Parrish JK. 2014. Next Steps for citizen science. *Science* 343: 1436–1437.

Brown ED, Williams BK. 2019. The potential for citizen science to produce reliable and useful information in ecology. *Conservation Biology* 33: 561–569.

Callaghan CT et al. 2022. The benefits of contributing to the citizen science platform iNaturalist as an identifier. *PLOS Biology* 20: e3001843.

Deck J, Guralnick R, Walls R, Blum S, Haendel M, Matsunaga A, Wieczorek J. 2015. Meeting report: Identifying practical applications of ontologies for biodiversity informatics. *Standards in Genomic Sciences* 10: 1–6.

Di Cecco GJ, Barve V, Belitz MW, Stucky BJ, Guralnick RP, Hurlbert AH. 2021. Observing the observers: How participants contribute data to iNaturalist and implications for biodiversity science. *BioScience* 71: 1179–1188,

Dowle M, Srinivasan A. 2021. data.table. GitLab. https://rdatatable.gitlab.io/data.table.

Gaier AG, Resasco J. 2023. Does adding community science observations to museum records improve distribution modeling of a rare endemic plant? *Ecosphere* 14: e4419.

Goodwin ZA, Harris DJ, Filer D, Wood JRI, Scotland RW. 2015. Widespread mistaken identity in tropical plant collections. *Current Biology* 25: R1066–R1067.

Groll A, Tutz G. 2014. Variable selection for generalized linear mixed models by L1-penalized estimation. *Statistics and Computing* 24: 137–154.

Harmon A. 2022. Can humans find common ground? Sure. Just Start With Sea Slugs. New York Times (9 December 2022).

Hochmair HH, Scheffrahn RH, Basille M, Boone M. 2020. Evaluating the data quality of iNaturalist termite records. *PLOS ONE* 15: e0226534.

iNaturalist contributors. 2023. iNaturalist Research-Grade Observations. Global Biodiversity Information Facility. https://doi.org/10.15468/ab3s5x.

Kelling S, Gerbracht J, Fink D, Lagoze C, Wong W-K, Yu J, Damoulas T, Gomes C. 2012. ebird: A human/computer learning network for biodiversity conservation and research. Pages 2229–2236 in EDITORS, eds. *AAAI'12: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. AAAI Press.

Kelling S, et al. 2015. Can observation skills of citizen scientists be estimated using species accumulation curves? *PLOS ONE* 10: e0139600.

Kelling S et al. 2019. Using semistructured surveys to improve citizen science data for monitoring biodiversity. *BioScience* 69: 170–179.

Koo K-S, Oh J-M, Park S-J, Im J-Y. 2022. Accessing the accuracy of citizen science data based on iNaturalist data. *Diversity* 14: 316.

McMullin RT, Allen JL. 2022. An assessment of data accuracy and best practice recommendations for observations of lichens and other taxonomically difficult taxa on iNaturalist. *Botany* 100: 491–497.

Mesaglio T, Callaghan CT. 2021. An overview of the history, current contributions and future outlook of iNaturalist in Australia. *Wildlife Research* 48: 289–303.

Olson DM et al. 2001. Terrestrial ecoregions of the world: A new map of life on Earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience* 51: 933–938.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation.

Rosa RM, Cavallari DC, Salvador RB. 2022. iNaturalist as a tool in the study of tropical molluscs. *PLOS ONE* 17: e0268048.

Shirey V, Belitz MW, Barve V, Guralnick R. 2021. A complete inventory of North American butterfly occurrence data: Narrowing data gaps, but increasing bias. *Ecography* 44: 537–547.

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.

Ueda K. 2016. iNaturalist Blog: Identify. iNaturalist. www.inaturalist.org/blog/6475-identify.

Vidal M, Wolf N, Rosenberg B, Harris BP, Mathis A. 2021. Perspectives on individual animal identification from biology and computer vision. *Integrative and Comparative Biology* 61: 900–916.

R Special Interest Group on Databases, Wickham H, Müller K, R Consortium. 2021. DBI. DBI. https://dbi.r-dbi.org.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J. 2019. Welcome to the Tidyverse. *Journal of Open Source Software* 4: 1686.