Mammalian Biology (2022) 102:1005–1023
https://doi.org/10.1007/s42991-021-00221-3

DATA INTEGRATION AND MANAGEMENT

# Flukebook: an open-source AI platform for cetacean photo identification

Drew Blount[1] · Shane Gero[2,3,4] · Jon Van Oast[1] · Jason Parham[1,5] · Colin Kingen[1] · Ben Scheiner[1,6] · Tanya Stere[1] · Mark Fisher[1] · Gianna Minton[7] · Christin Khan[8] · Violaine Dulau[9] · Jaime Thompson[10] · Olga Moskvyak[11] · Tanya Berger-Wolf[12,13,14,15] · Charles V. Stewart[5] · Jason Holmberg[1] · J. Jacob Levenson[16,17]

Received: 4 June 2021 / Accepted: 27 December 2021 / Published online: 5 April 2022
© The Author(s) 2022, corrected publication 2022

## Abstract

Determining which species are at greatest risk, where they are most vulnerable, and what are the trajectories of their communities and populations is critical for conservation and management. Globally distributed, wide-ranging whales and dolphins present a particular challenge in data collection because no single research team can record data over biologically meaningful areas. Flukebook.org is an open-source web platform that addresses these gaps by providing researchers with the latest computational tools. It integrates photo-identification algorithms with data management, sharing, and privacy infrastructure for whale and dolphin research, enabling the global collaborative study of these global species. With seven automatic identification algorithms trained for 15 different species, resulting in 37 species-specific identification pipelines, Flukebook is an extensible foundation that continually incorporates emerging AI techniques and applies them to cetacean photo identification through continued collaboration between computer vision researchers, software engineers, and biologists. With over 2.0 million photos of over 52,000 identified individual animals submitted by over 250 researchers, the platform enables a comprehensive understanding of cetacean populations, fostering international and cross-institutional collaboration while respecting data ownership and privacy. We outline the technology stack and architecture of Flukebook, its performance on real-world cetacean imagery, and its development as an example of scalable, extensible, and reusable open-source conservation software. Flukebook is a step change in our ability to conduct large-scale research on cetaceans across biologically meaningful geographic ranges, to rapidly iterate population assessments and abundance trajectories, and engage the public in actions to protect them.

**Keywords** Community science · Computer vision · Deep learning · Individual identification · Machine learning · Mark recapture

## Introduction

Answering which species are at greatest risk, where they are most vulnerable, and what are the trajectories of their communities and populations is critical for conservation and management. Yet there are still large gaps in our ability to answer these questions, even for well-studied taxa (Rondinini et al. 2014). Wide-ranging species may be well known, and their biology and behavior well understood, but the appropriate data on their distribution and abundance remain lacking.

Globally distributed, wide-ranging, and highly mobile, marine species present a particular challenge in addressing these data gaps because no single research team can collect

Handling editors: Stephen C.Y. Chan and Leszek Karczmarski.

Jason Holmberg, J. Jacob Levenson: co-senior authors.

This article is a contribution to the special issue on "Individual Identification and Photographic Techniques in Mammalian Ecological and Behavioural Research – Part 1: Methods and Concepts" — Editors: Leszek Karczmarski, Stephen C.Y. Chan, Daniel I. Rubenstein, Scott Y.S. Chui and Elissa Z. Cameron.

✉ Drew Blount
drew@wildme.org

Extended author information available on the last page of the article

data over the entire range of many species. Cetaceans (the whales, dolphins, and porpoises) are a particularly acute example: 24% of all species are assigned to a threatened category (Vulnerable, Endangered, or Critically Endangered; 22 of 90) and 10% of all species are currently listed as 'Data Deficient' (nine of 90 species, IUCN 2020). To learn how these animals live and what influences their lives (either ecological or anthropogenic), we need to ask questions over huge geographical, and indeed logistical, scales.

From the boats that take researchers out to sea to the devices that help us access the depths, marine science has been and always will be a technology-driven field. It could be argued that the camera was the tool and photo-identification the technique that changed our approach to marine mammal research more than any. This pioneering non-invasive method of study drove a major shift in our approach to research and generated a huge amount of scientific knowledge on animals as individuals (see Würsig and Würsig 1977; Katona and Whitehead 1981; Bigg 1982; Payne 1986; Hammond et al. 1990). Numerous cetacean species have been particularly well studied by photo ID, generating large catalogs of photos of identified individuals, for example humpback whales (*Megaptera novaeangliae*, e.g. Barlow et al. 2011), bottlenose dolphins (*Tursiops* spp., e.g. Wells and Scott 1990), and North Atlantic right whales (*Eubalaena glacialis*, e.g. Hamilton and Martin 1999). These catalogs are the product of a huge amount of manual human labor, time, and cost; where each new photo must be matched against the rest of the catalog of known individuals in order to answer, "who is this animal?" and understand abundance, distribution, demographics, social behavior, health, and movement.

Because of the inherent costs of being at sea, cetacean field study is costly, and as a result, less than 1% of all cetacean species' ranges have been surveyed with enough frequency to estimate population trends (Kaschner et al. 2013). Outside of these larger transects, much cetacean research is underfunded and geographically small scale compared to the geographic spread of cetacean populations or species distributions. As such, much of the data management and analysis tools used for this research are developed in an ad hoc manner to fit current needs, under the current funding, to address the questions and actions at hand. As a result, protocols to curate and analyze photographic data often differ from project to project, creating serious challenges for collaboration between institutions or study sites, as well as data legacy concerns. Arriving at a critical mass of data for population analysis can take years, especially for rare or endangered species. After that, the enormous amount of manual data processing that is necessary for manual photo ID can create multi-year lags between data collection and scientific results, which can result in conclusions too coarse or too slow for effective and dynamic conservation and management actions. This limits the scope, scale, repeatability, continuity, and return-on-investment of these studies.

With modern advances in computer vision, especially those integrating machine learning, computers can increasingly curate these catalogs and match these photos for us. Yet despite algorithmic advances, marine mammal researchers often do not have the technical expertise or budget to take advantage of the latest computing tools. And even when cutting edge algorithms are employed, without a standard data format and platform, old problems of interoperability, data isolation, and data legacy remain. Closed management of data limits researchers from collaborating not only with each other, but with community scientists and members of the public. Despite collecting potentially valuable ecological data, engaged community scientists face many obstacles in connecting to researchers and contributing to scientific study.

Flukebook (https://www.flukebook.org) is a non-profit, open-source cetacean data archiving, analysis, and photo-identification tool developed as part of the larger Wildbook platform (https://wildme.org/#/wildbook). By bringing together software developers, machine learning researchers, and biologists, Flukebook bridges the gap between modern computing techniques and ecological field research to provide the latest computer vision technology to cetacean researchers around the world. Users upload data through a web interface, and the platform automatically detects and draws boxes around identifiable cetacean features in those photos, and identifies the individual animals therein using the latest algorithms for cetacean photographic ID. The platform also facilitates the management of these catalogs of photos and observations that are necessary to support these studies.

Flukebook allows historical catalogs to be imported as well as unprocessed data, and supports a large number of ecological observations including measurements, biological samples, effort data, and strandings. Users can export their data from Flukebook in a variety of formats, supporting third party tools for mark-recapture, social group analysis, and mapping, as well as data frameworks such as OBIS and GBIF for greater access and use by government and managers. Flukebook allows researchers, institutions, and community scientists to collaborate and share data on an opt-in basis, allowing for the big picture analysis that is necessary to study these long-lived and highly mobile species in a way that preserves data privacy and ownership.

Flukebook is a step change in our ability to conduct large-scale research on whales across biologically meaningful geographic ranges, to identify and manage their populations, and to engage the public in actions to protect them. Below, we introduce the platform; its data structure; its tools for managing, analyzing, importing, and exporting data; its model for data collaboration and privacy; and outline its advanced

capabilities for detecting and identifying individual animals with machine learning and computer vision tools.

## Platform overview

Flukebook is a web-based application. This breaks with a common paradigm of desktop-based photo ID tools for cetaceans (e.g. Whitehead 1990; Stanley 1995; Hillman et al. 2003; Crall et al. 2013; Weideman et al. 2017; Thompson et al. 2019; Moskvyak et al. 2019) to achieve myriad benefits. A web-based application allows users to access powerful machine learning algorithms running on high-powered servers without the need for any specialized equipment on their local machine. The site can be used from a web browser anywhere in the world with an internet connection reliable enough to upload and download images. Because it is a shared platform on a single server, it provides researchers with the ability not only to match photos within their own catalog, but to compare with other data collectors who might have seen the same individuals. Flukebook's server resources can be easily scaled-up by the developers at Wild Me (https://www.wildme.org), for example adding more processing power or database storage as the platform grows, and this server is automatically backed-up nightly ensuring catalogs cannot be lost by technological accidents. Fixes and improvements to the platform are instantly available to all users without any effort on their part.

The platform is free for approved users, and the Wild Me nonprofit organization which develops and manages Flukebook has no plans to charge for its use at any point in the future. The nonprofit's operating costs including Flukebook servers are covered primarily by grants (see this paper's funding statement), and to a lesser extent service contracts where the nonprofit is paid to develop and make available new features for the platform.

The most important organizational feature of Flukebook is its position as an open-source project in the larger family of Wildbook platforms. As of June 2021, there are 21 Wildbooks maintained by Wild Me, supporting the study of over 50 species. Each Wildbook instance runs a branch of the open-source Wildbook codebase (https://github.com/WildMeOrg/Wildbook), and can be understood as a computer vision-supported catalog for different species, such as sharks (https://www.sharkbook.ai/), giraffe (https://giraffespotter.org/), or giant sea bass (https://spottinggiantseabass.msi.ucsb.edu/), with Flukebook being the Wildbook for cetaceans. The open-source nature of Wildbook allows new features or algorithms developed for one platform to be easily deployed on all of them at virtually no cost. This provides a huge value for each investment and fundraising effort, and a significant effect in conservation where funding can be scarce. For example, the bulk data uploader on Flukebook was originally developed with funding from groups in the Indian Ocean and has since become widely adopted on other Wildbooks (e.g. the African Carnivore Wildbook, https://africancarnivore.wildbook.org; and the Amphibian and Reptile Wildbook, https://amphibian-reptile.wildbook.org) granting access to this feature to hundreds of researchers across the globe. This scales up the impact of conservation funding by rapidly sharing technological innovation with research communities for many terrestrial and marine species. The result is that small investment in one side of the platform's ecosystem results in improvements, bug fixes, and new features across all instances.

The architecture of Flukebook and other Wildbook platforms is bipartite in nature. These two parts can be understood as the data management server and the image analysis server (Fig. 1). The data management server hosts the frontend web application on Flukebook.org, where users go to log in; inspect, manage, and upload data; start matching jobs; etc. This frontend contains all of the user interface (UI) components and interactions facing the users. The backend computer vision server, running a separate code repository on a separate machine, is called Wildbook image analysis (WBIA). These servers communicate via http requests, with images and computer vision queries sent from Flukebook to WBIA, and computer vision results sent back from WBIA to Flukebook. Both servers run SQL databases containing the data necessary for that server's function, which are synchronized by these calls and other nightly integrity scripts.

## Data management and curation

Though perhaps not as exciting as the computer vision algorithms, the data management web app's frontend of flukebook.org is equally if not more important in making the platform a valuable scientific tool. Tasks such as uploading data, inspecting photos, querying, reviewing, and confirming match results, bulk processing and exporting data, sharing data and collaborating, and producing a resolved catalog of unique individuals are all tasks separate from the computer vision processing that takes place on the image analysis server. In essence, this data management platform can be understood as a user interface on top of a database of ecological observations, with a suite of data processing and inspection tools. The database structure was designed with mark-recapture population studies in mind. As of June 2021, Flukebook hosts over 2 million photos spanning 266,038 sightings of 52,638 identified individual cetaceans. The user interface is available in four languages (English, French, Spanish and Italian).

The Flukebook frontend and data management server is an instance of Wildbook, specified in the 'flukebook' branch of the open-source Wildbook repository available on GitHub

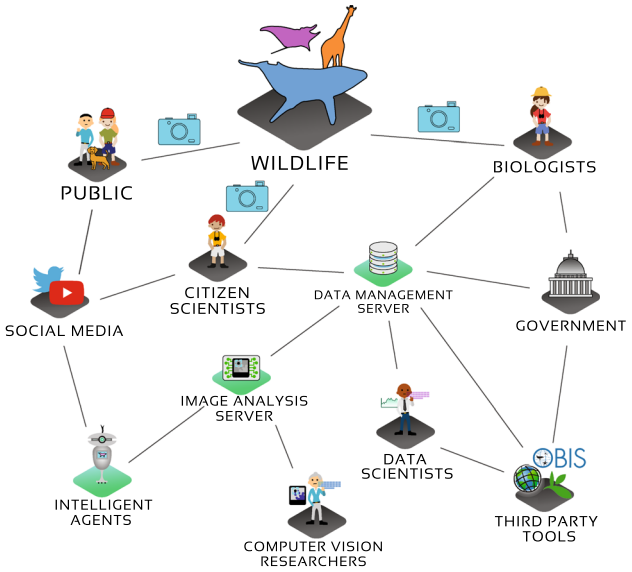


**Fig. 1** System diagram of Wildbook servers and stakeholders. Created by Wild Me

(https://github.com/WildMeOrg/Wildbook/tree/flukebook). Except for user interface themes and cetacean-specific algorithm configurations, Flukebook's code is largely identical to that of the master-branch Wildbook, a trait that all other Wildbook 7.x platforms share. Wildbook is a Java web application built on the Apache Tomcat framework (http://tomcat.apache.org/).

Table 1 provides an overview of the core high level data objects on Flukebook. The central data object is the *Encounter*, a recorded sighting of a single individual at a single place and time. In the capture-recapture framework, an *Encounter* is a capture of a single individual. The *Encounter* contains uploaded photos of the event in addition to observations researchers may have made in the field, such as behavior and morphological measurements. Attached to these photos are the derived data products such as labeled bounding boxes (highlighted rectangular regions in an image) around features such as dorsal fins and flukes that were found automatically, and links to identification results.

Other top-level data objects include the *Marked Individual*, an individual animal represented by its history of identified *Encounters*. The only constraint on these identifications is that each one must have been approved by the user, meaning that historical catalogs of *Encounters* can be uploaded with or without photos. New identifications made with the aid of matching algorithms on Flukebook also must be confirmed by the users, ensuring that data owners maintain ultimate control over data integrity and the *Marked Individual* objects, which contain the consolidated photos and observations of a single animal over time. The third major object is the *Sighting*, representing a sighting of a group of animals, containing observational fields which are not specific to a single individual such as group size, group behavior, and weather conditions.

The main objects have unique web endpoints, such that any *Encounter*, *Marked Individual*, and *Sighting* has a static URL that can be referenced and linked, where all data fields can be inspected and edited by approved users. These three objects are linked to each other in both the database and user interface (UI), so that one can navigate in a single click from a *Marked Individual* to one of its *Encounters*, then to the *Sighting* where one can find the other animals that were co-occurring during that observation, and so on. This connectivity is highlighted by the co-occurrence diagram on a *Marked Individual* page, which is dynamically computed to show at a glance the other individuals that have been observed in groups (via the *Sighting* objects) with the individual in question. For social cetaceans, Flukebook also displays visualizations of specific social relationships recorded by researchers, such as community membership and kin relations, as shown in Fig. 2.

It is important to note that while Flukebook dictates this data model, significant discretion is left to the users and data owners regarding data quality and standards. The goal is for all data to be interoperable without dictating particular research practices that may differ from study to study, and which may have already been established in ongoing studies that predate the platform itself. For example, the *Sighting* object concerns a group of animals observed in a single

**Table 1** The core high-level objects in the Flukebook data model

| Data object name | Description | Example subfields |
|---|---|---|
| *Encounter* | A sighting of a single individual at a single place and time | Photos, annotations (labeled photo regions), date and GPS coordinates, observational fields like behavior or size |
| *Marked individual* | An individual animal that has been identified across one or more *Encounters* | Species, sex, name/catalog ID, list of *Encounters* |
| *Sighting* | A sighting of a group of animals at a single place and time. A *Sighting* contains the *Encounters* for each individual observed in the sighting | Group size, group behavior, weather conditions |

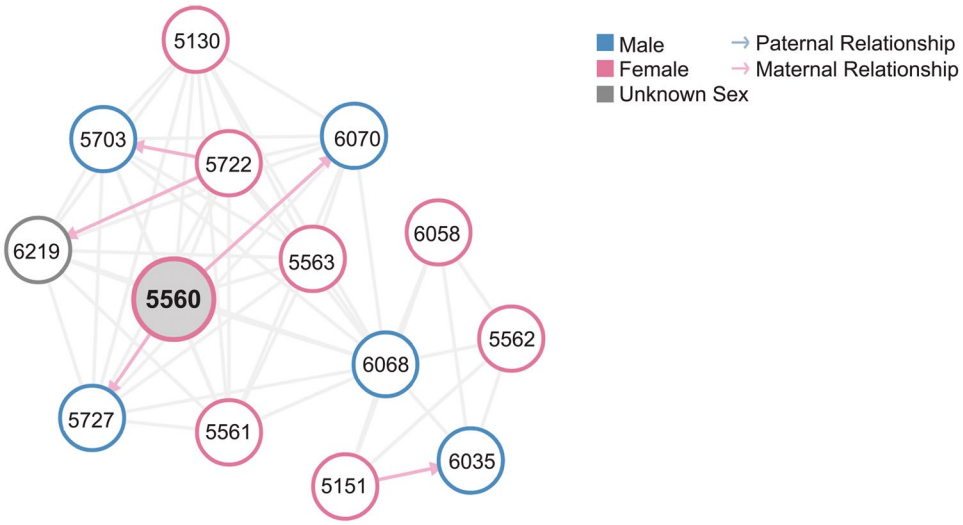


**Fig. 2** Social relationships visualized based on Flukebook data for "Pinchy", #5560, a sperm whale in the Eastern Caribbean. This illustrates all individuals with whom she co-occurred and maternal relationship within her unit, unit F (individuals #5130, 5703, 5722, 6070, 6219, 5727, 5563, 5561, and 6068) and bond-group unit U (individuals #6058, 5562, 6035, and 5151; Gero et al. 2014). Data courtesy of The Dominica Sperm Whale Project

place and time. But different researchers may use different notions of what constitutes a co-occurring group of animals, or even what constitutes a single time (is a continuous 6-h observation one event, or several?). By leaving these decisions to the user, the platform aims to be a neutral tool that facilitates diverse research. Simultaneously, the data and collaboration models encourage users to work together when doing so is beneficial.

## Data ownership and collaboration model

Flukebook has standard user accounts with securely stored passwords. Each account controls access to the data which they have uploaded. Users can be recorded as members of *Organization* objects, which represent real-world research groups and consortia. These *Organizations* serve to conform the Flukebook experience to a research project's data standards. Categorical observations (where the user inputs an observational value from a list of options), such as observed behavior on an *Encounter* or photo quality labels, are standardized on this *Organization*-level. These *Organizations* also serve to resolve the issue of conflicting names or ID numbers for *Marked Individuals* between different research groups, by allowing each *Organization* to record distinct individual names for a given animal. Standardizing data options such as these at the *Organization* level, rather than Flukebook-wide, is crucial for supporting the broad range of data collection standards used by cetacean researchers while still enforcing consistency within projects. For example, the metrics used to quantify photo quality or distinctiveness of an individual, both which are important to photo-identification work and the analyses they support, vary both between researchers studying the same species as well as between species; and often show observer bias between researchers using the same metric (see for example Urian et al. 2015 who found a remarkable degree of variation between experienced researchers' photo quality scores, even when using the same scale, and in addition they disagreed significantly on the total number of individuals in a controlled dataset). While this flexibility does result in different research teams using different labels in the metadata associated with photos, many users would simply not use the platform if they could not implement their existing data standards and definitions; so this flexibility was designed as a part of Flukebook as a key to enabling these collaborations. Simultaneously, researchers who intend to combine their observations in larger collaborative studies are, by having their data stored in Flukebook, provided with a conversion between two scales or metrics for the same parameter which ensures their data are fully consistent. Importantly, scientifically critical decisions can be made when analyses are undertaken off the platform after data are collated and exported. For example, researchers might apply a photo quality standard for inclusion in population level analyses, or their own definition of co-occurrence or association for social analysis.

Data ownership is a high priority factor for many Flukebook users and has often been the primary concern for researchers considering a shared data platform. Indeed, several large preexisting catalogs which have been added to Flukebook are constrained by existing privacy and ownership agreements that predate Flukebook itself, which must be respected while the data are hosted on the platform to facilitate automatic identification and other processing. Simultaneously, one of the core objectives of Flukebook is to enable and encourage collaboration between research groups and data providers. Valuable insights are gained by comparing and pooling data from multiple organizations, informing a more complete understanding of a population's abundance, distribution, and behavior. By design, Flukebook encourages

multi-institutional, international, collaborative research across large geographic scales while maintaining a respect for the privacy and original ownership of submissions.

The actual mechanics of Flukebook's security system are relatively simple and are the product of discussions with stakeholders and government agencies with diverse needs. First, every *Encounter* on Flukebook is owned by a single user account, by default the submitter of that data. This data owner determines who else can read and edit their data. Data visibility and access permissions are defined via pairwise *Collaborations* between users, which must be reciprocally approved at either "view-only" or "edit" levels.

These data sharing collaborations are initiated when a user sends a request to another user, usually prompted by an attempt by the requester to view the recipient's data and being denied access. When a collaboration request is sent, an email is sent to the recipient which includes an optional message from the requester and importantly, the requester's email address (user emails are otherwise hidden from other Flukebook users, and the requester does not see the recipient's email unless the recipient emails them). Sharing this contact information allows the two parties to establish communication directly and agree on how each other's data can be used; some users choose to establish, over email, formal memoranda of understanding regarding these agreements. Within Flukebook, a notification is sent to the user's account page where *Collaborations* can be managed, approved, and revoked.

Within this data ownership system Flukebook strongly encourages collaboration. While complete data records are only visible as prescribed by the *Collaborations*, summary information is just visible enough to other users, so that those users can determine if they would like to request access to those data (with the exception of certain locked-down datasets described in the next paragraph). In particular, the automated matching system compares across all pictures of a species, within the given search criteria, and will return a photo uploaded by another user if it is found to match. This allows the crucial discovery of two data collectors observing the same individual cetacean, perhaps separated by a large physical or temporal distance. In this case, upon seeing a match between their respective data, researchers can contact each other and decide whether they would like to collaborate and share more information. Besides during the matching process, users can also discover potentially relevant data to which they do not yet have access in the search functions (described below). For example, a sperm whale researcher might search for sperm whale sightings in the Caribbean in a certain time period to view summary data across users, and then send a *Collaboration* request to other users who have submitted such data. These two instances of partial data visibility—match results and search results—allow users to discover others studying the same research areas, species, and questions. This creates an environment in which users are consistently urged to learn how their data connect with that of other users on the platform, with the goal of building an increasingly collaborative research community among Flukebook users. However, full data records cannot be viewed nor exported from the platform without all users agreeing to established *Collaboration* agreements.

It is important to note that users can also opt out entirely from collaboration requests and even partial data visibility to other Flukebook users, essentially using the platform as a private matching tool; however, this option is hidden and can only be activated by site administrators. On the opposite end of the privacy spectrum, data submitted by members of the public without user accounts or organizational affiliation are visible to every user on Flukebook.

### Data validation

One of the standardization benefits of a shared data platform is providing automated validation and consistency checks on all data that are submitted. This can be as simple as constraining user input to ranges of values, such as positive integers, or a predefined list of study sites and recognized species. Flukebook also performs more complex data checks such as validating that a location's latitude/longitude lies within a fixed distance from a body of water. Invalid data can be rejected outright, and borderline data marked for further review.

## Search tools

Robust search and analysis tools are useful not only for navigating the catalogs on Flukebook but for generating useful data products. The Flukebook database can be searched by *Encounters*, *Marked Individuals*, and *Sightings*. Search forms translate a user's input on a web page into SQL queries of the database backend. Most data fields can be used as search filters, for example species, date, location, submitter, observed behavior, life stage, photographed features such as fluke or dorsal fin, whether a biological sample was taken, whether a satellite tag was deployed during the observation, and so on. Search results are presented in a tabular format where each row is one database record. Clicking that row takes the user to the record's web page; the table also includes hyperlinks to related data objects such as the *Marked Individual* page for a given *Encounter*. The SQL query which was derived from the search input is shown below the results, providing added clarity for technical users and developers.

Since search results effectively represent a dataset on Flukebook, for example "male sperm whales sighted by me and my collaborators in Dominica in 2018," search results

are where many of Flukebook's data analysis tools are located. In addition to the default tabular representation of these results, users can see them mapped geographically, distributed temporally on a calendar, or as a gallery of all matching images. There is a further 'Analysis' tab available on *Encounter* and *Marked Individual* search results that shows various computed analyses of that dataset, such as a discovery curve (Fig. 3), an annual seasonality histogram showing the number of *Encounters* per calendar week (Fig. 3), and pie charts of various distributions such as species, sex, and location of matching records.

### Data exports

In addition to the analysis and inspection tools provided in search results, users have the option of exporting the data represented by a search in a number of different formats for analysis using other third-party tools. These exports are also subjected to the Flukebook data security model, ensuring no user can export data they do not have permission to access.

The most basic format is a standard Excel sheet with one row per database object and one column per database field. This is the "complete" Flukebook export that shows all the data fields stored for a given object type and search query. This is the "Flukebook Standard Format" used by the Bulk Uploader described in the following section. This complete export allows for advanced filtering that may not be implemented in the user interface: for example, the *Marked Individual* page shows every confirmed *Encounter* of that individual, but a user might want to analyze only those *Encounters* that meet certain data quality or photo quality criteria; this is possible with search and export filtering.

There are also exports formatted specifically for compatibility with other databases, in particular the OBIS Ocean Biodiversity Information System. Other exports include a GIS shapefile of *Encounter* locations; a KML export that is compatible with Google Earth; and a configurable mark-recapture export compatible with RMark (Laake 2013), useful for using Flukebook data to generate abundance and trend estimates.

A benefit of being open source is that it is easy for developers and researchers outside of the Wild Me organization (which manages Flukebook) to produce tools to export or analyze Flukebook data. The most significant example of this is the RWildbook R package (Bonner and Huang 2018), which allows a user to pull data records directly from Flukebook into their R environment after providing their login credentials.

## Data input

Data submitted to Flukebook should include at a minimum three key pieces of information: what was seen, where it was observed, and when it was observed. Submissions can be uploaded through two pathways: encounter reporting and bulk upload. These distinct use cases are briefly described here, and more detailed step-by-step instructions for researcher data upload are available in the official software documentation at https://docs.wildme.org/docs/researchers/data_entry.

### Encounter reporting

A simple web form can be used to report an *Encounter* in cases where the user wishes to submit data from an observation of a single animal with one or more photos. When choosing this pathway, it is assumed that all photos contain only the same single animal, satisfying the definition of

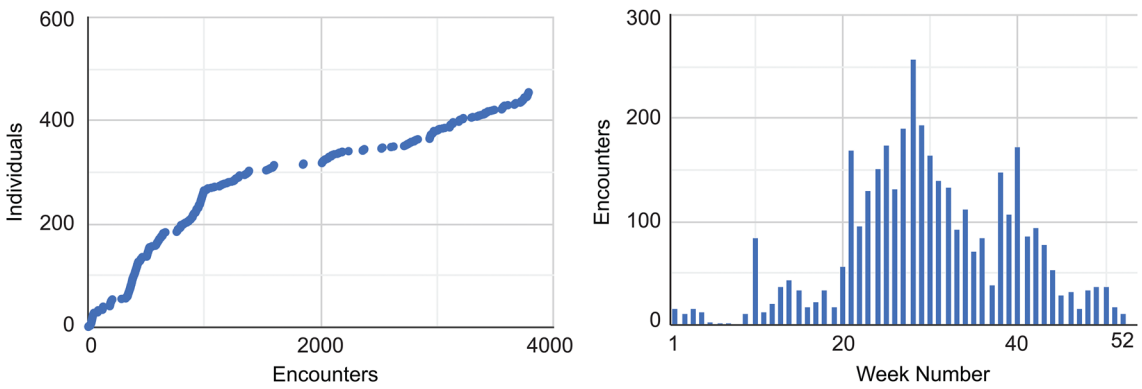


**Fig. 3** Two plots generated for Flukebook *Encounter* search results, in this case humpback whale *Encounters* off Iceland ($n = 455$ individuals). Left is the discovery curve of new individuals as a function of submitted and identified *Encounters*. Right is a seasonality histogram of *Encounters* per calendar week. Data courtesy of Elding Whale Watch

an *Encounter* (one animal at a location and point in time). This method is best used for more solitary animals and opportunistic sightings.

The simple online form guides the user to drag and drop or upload the photo files from their computer, submit the date and time the image(s) were taken (which can be automatically pulled from the EXIF metadata from the uploaded images themselves, which is presented to the user for their confirmation), describe the location the *Encounter* took place (including options to label country, field research sites listed in Flukebook already, GPS position which can also be pulled from the EXIF metadata from the images uploaded), label the species of whale or dolphin in the *Encounter* from a dropdown menu (which determines the computer vision pipeline that is then triggered, illustrated in Fig. 4), and some descriptors about themselves including name and email if they are not logged into their account.

There is an 'Additional Information' tab at the bottom which allows the submitter to include a larger more detailed array of fields connected to the *Encounter* including, but not limited to, *Organization*, project name, sex, observed behavior, metrics photo quality and markedness, other noticeable scarring, life stage and status, alternate names for the animal, *Sighting* ID, other email addresses, and additional comments, all of which can be edited after submission as well.

## Bulk upload

The Flukebook bulk upload submission pathway allows for hundreds or even thousands of *Encounters* to be submitted at once. The Bulk Uploader was designed to integrate a field season's worth of images or to allow for new users to upload their mid-sized, historical catalogs on their own.

During the Bulk Upload process, a folder of images is uploaded, along with metadata in an Excel file in a standardized format called the Wildbook Standard Format (available at https://docs.wildme.org/docs/researchers/bulk_import, or with a Flukebook login on the Bulk Upload page). The format is simple, with each row in the spreadsheet corresponding to one *Encounter*, and each column header corresponding to one data field. Each *Encounter* must include one field for each of the following: (1) location, (2) date, and (3) taxonomy. If multiple individuals were seen in a group, that co-occurrence information is recorded with optional *Sighting* columns. Many more detailed columns are included in the standard format, the use of which are optional. Automatic data quality checks are performed, such as ensuring that data fields are provided in the correct type and format, and the user reviews their submission before it is committed to the database. Ensuring data quality and integrity in this way has been found to be very important when allowing users to upload hundreds or thousands of records at once.

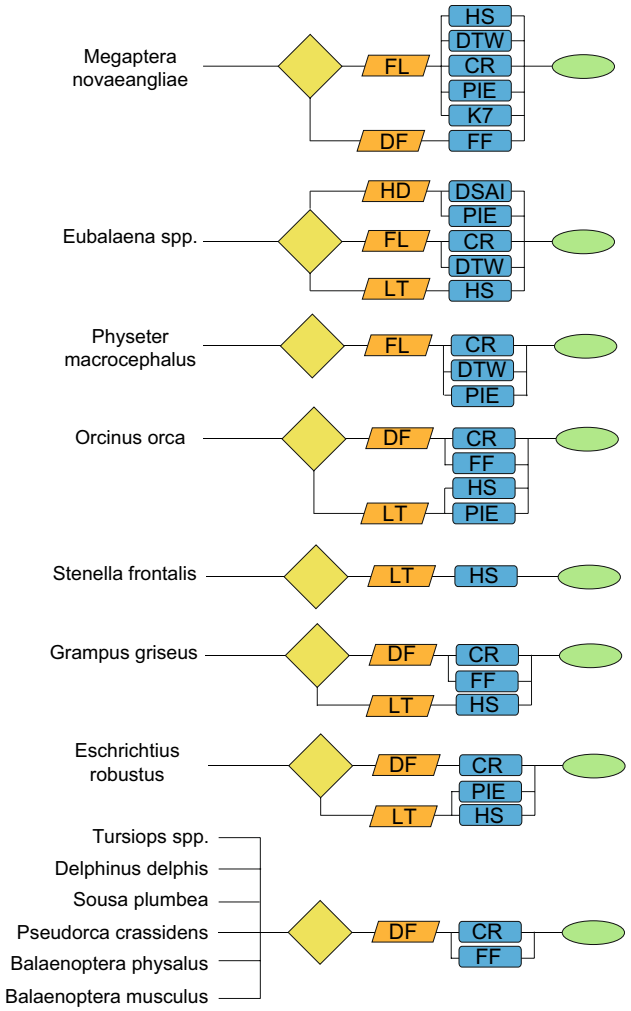


**Fig. 4** Flukebook species-specific image processing pipelines from automated detection (yellow diamonds) of body parts (orange parallelograms: *FL* fluke, *DF* dorsal fin, *HD* head, *LT* lateral side of body), through identification through multiple algorithms which have been trained for specific species (blue, rounded rectangles: *HS* HotSpotter, *DTW* dynamic time warping, *CR* CurvRank, *PIE* pose-invariant embeddings, *K7* Kaggle7, *FF* FinFindR, *DSAI* Deepsense.ai), and finally to defining Marked Individuals (green ovals). Fifteen of the 37 species-specific pipelines are new since the beginning of 2021, with more under active development

To avoid issues related to network connectivity, a user is allowed to upload photos in multiple web sessions, and all prior-uploaded photos are available for reference from subsequent Excel sheet uploads. To avoid over-taxing the server, a limit of 1000 rows per Excel file is encouraged.

For new users with exceptionally large historical catalogs that may be tens or hundreds of gigabytes in size, researchers may contact the developers at Wild Me to onboard these datasets on a contracted basis without going through the Bulk Uploading process themselves.

Importantly, Flukebook does not require photos to be associated with each *Encounter*, which allows users to bulk upload entire historical sightings catalogs on a purely presence-based basis. This allows researchers to use only the database management parts of the platform in cases where they do not have the data or motivation to use the computer vision features.

### Ocean alert mobile app

Currently, Flukebook has a front-end app for iOS and Android called Ocean Alert (Conserve.IO, https://apps.apple.com/us/app/ocean-alert/id1457113771) which allows community science and research users to submit data through their mobile devices. Supported data includes species, group size, behavioral data, and photos for identification, as well as effort in the form of vessel tracks that are saved on Flukebook. This project opened the possibility of connecting any app that collects this kind of data to Flukebook using an established application programming interface (API), which is a defined standard that allows external programs to communicate with the Flukebook server. This illustrates how Flukebook is not a closed platform, and this type of integration is an area of future development, with potential connections with other applications and projects. Of course, such efforts also conform with the privacy and ownership of user-submitted data on Flukebook.

## Computer vision: Wildbook image analysis

Access to cutting edge computer vision algorithms is the novel feature that draws in most users to Flukebook. The goal of this system is to decrease the amount of research time and effort needed to perform individual identification. This core computer vision functionality operates on a separate server than the frontend data management web application, on an instance of the Wildbook image analysis server (WBIA, available at https://github.com/WildMeOrg/wildbook-ia). This image analysis server is a Python application (https://www.python.org) that uses the Flask framework (https://flask.palletsprojects.com) to operate as a web server in communication with the frontend. The computer vision server operates a database containing primarily images and information algorithmically derived from those images. This image database structure is summarized in Table 2.

The image analysis server contains a large number of tools for image manipulation and processing, supporting the two primary phases of the image analysis pipeline: (1) detection, where the system takes an unprocessed image and draws bounding boxes around every animal in that image, each box and additional metadata forming an *Annotation*, and (2) identification, where one or more matching algorithms are run on each *Annotation* to identify the individual animal within it.

### Detection

The detection pipeline (presented in-depth in Parham et al. 2018) uses deep learning to process raw input images into labeled and cropped *Annotations*, the semantic image subregions described above, with the ultimate goal of finding the relevant regions to send as input to the subsequent identification phase. The WBIA detection pipeline is a cascade of deep convolutional neural networks (DCNNs) that apply a fully connected classifier on automatically extracted features. Two separate networks produce (1) *Annotation* bounding box locations, and (2) feature labels (such as fluke or dorsal fin), viewpoint labels (such as left vs. right dorsal views), and image quality for each candidate *Annotation*.

The image analysis server's detector is generally trained on a per-species basis, meaning that adding support for each new species requires a retraining process and a manually annotated training set of roughly 1,000–2,000 representative photos of that species (this number differs depending on how challenging the feature is to detect, as well as the availability of training data). Flukebook has a standardized web-based user interface for efficiently performing this data-labeling task, which involves users clicking to draw bounding boxes on source images and labeling viewpoint and orientation. Researchers work with the Flukebook development team to accomplish this task, providing expertise about identification for the newly supported species and how it is manually identified traditionally. Often Flukebook developers also participate in the labeling task, which gives them useful familiarity with the animals and features being studied. This is a concrete example of the biologist-programmer collaboration, which occurs across the Wildbook ecosystem. Examples of both a computer-generated annotation and hand-generated training annotations are shown in Fig. 5.

Flukebook's detectors allow researchers to submit their raw photos, no longer needing to manipulate and/or crop images one at a time to perform individual identification; and to submit one photo containing multiple individuals without duplication for each animal. This is a significant timesaver, removing much of the effort in image preprocessing, and tightens the loop between data collection and useful data products.

**Table 2** Top-level image-derived objects in the Wildbook image analysis data model

| Data name | Description | Data type |
|---|---|---|
| *Image* | The raw images that are the inputs of the whole system | Binary image files in supported formats, .jpg and .png |
| *Annotation* | A rectangular subset of an image containing an individual animal or part of an animal | Bounding box coordinates and pointer to the source image, optional theta for rotation |
| *Chip* | An *Annotation*, processed and resized to algorithm-specific dimensions; the direct input for identification algorithms | Pixel data as represented in a Python numpy array |

## Identification

Flukebook currently supports automatic matching for 15 cetacean species using seven major ID algorithms, some of which have multiple versions or trained models (Fig. 4). These algorithms target 30 species-specific features including flukes, dorsal fins, saddle patches, dorsal ridge, spots and scratches on flanks, and head callosities. This array of algorithms has grown consistently over the history of the project and at an accelerating pace, with 15 identification configurations added in the first 6 months of 2021. Including all of these matching methods in one place provides significant value, as many of them were initially developed for unique standalone desktop applications that are not interoperable (e.g. Crall et al. 2013; Thompson et al. 2019; Moskvyak et al. 2019).

Some of these algorithms, like HotSpotter (Crall et al. 2013) and CurvRank (Weideman et al. 2017, 2020), were developed directly by Wild Me's partners at Rensselaer Polytechnic Institute for use on Flukebook and related platforms. Others, such as finFindR (Thompson et al. 2019) and the pose-invariant embeddings algorithm (PIE, Moskvyak et al. 2019), were developed by independent computer vision researchers who then collaborated with Wild Me's machine learning engineers to integrate these algorithms within WBIA. A third category of algorithm is those that have won public machine learning competitions hosted on the popular competition platform Kaggle. These competition winners were made publicly available and were then integrated into Flukebook. This category includes the Deepsense.ai matcher for aerial right whale photos (Bogucki et al. 2018), and a humpback fluke matcher we have deemed "Kaggle7". The wide range of these algorithms' origins demonstrates the flexibility of the platform to rapidly integrate new methods. Algorithm development alone is not useful to biologists and managers

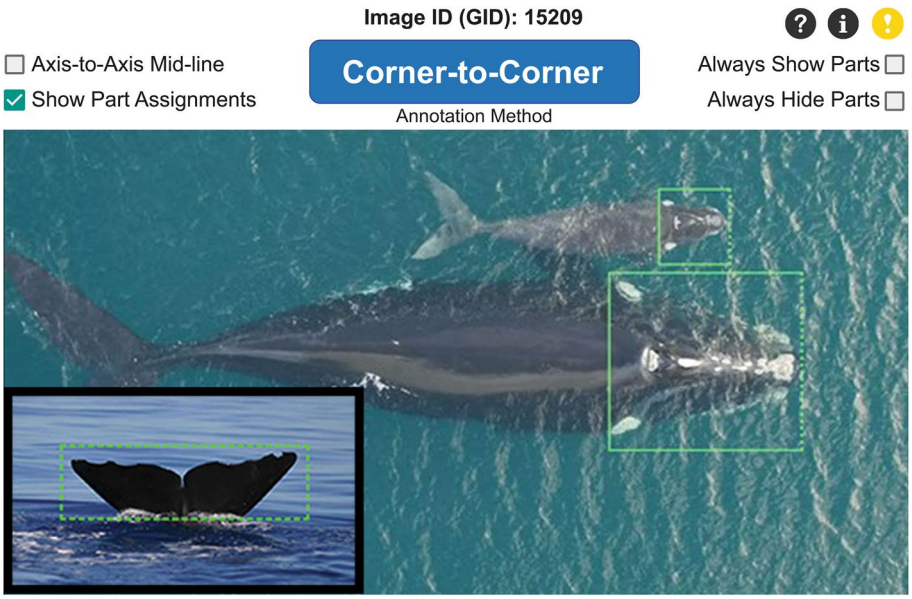


**Fig. 5** Using an interface for curating training data, this right whale image was manually processed by an experienced user to generate input for machine learning models that can then repeat the detection task automatically without human review. This process is required for any new species. Note the dotted line indicating the front of the whales, which is manually labeled in the training data so that orientation can be extracted automatically once operational on the platform. Inset: Once trained, this sperm whale fluke was detected automatically in Flukebook, allowing downstream machine learning identification tasks to focus on only areas of interest in a photo

without a software interface to host data, query the system, and to store and export results. The architecture of WBIA's identification pipeline is modular, allowing these algorithms to be used interchangeably and further algorithms to be integrated in a standardized way. An internal identification API uses the image analysis server's *Chips* and *Annotations* as inputs and produces a standardized match-score output format.

One challenge in the development and integration of novel algorithms is that, like much in academic computer science, computer vision and machine learning research often take the form of developing cutting-edge algorithms based on the latest theory and technology in order to publish in leading field-specific journals. This often leads to "academic software", which as a result of systemic constraints rather than any failure on the researcher's part, are hard to scale to production and can have relatively little focus on the usability of the tool for conservation. One of the underlying principles of Flukebook is to bridge this gap between academic algorithm development and applied conservation utility by bringing professional software design and development expertise together to package and deliver the latest computer vision algorithms as useful tools to researchers. In a real sense, Flukebook becomes the long-term steward of these algorithms, which ensures their curation, access, and source code legacy while also supporting their utility through software updates, re-training with new data, and changes to protocols. Below, we broadly describe each identification algorithm available in Flukebook and maintained by the Flukebook developers.

## HotSpotter

HotSpotter (Crall et al. 2013) is based on extracting and matching small, distinctive regions between query images and reference images in the catalog. These regions are found throughout each image using a Hessian-based keypoint extractor and described using the well-known SIFT descriptor (Lowe 1999). HotSpotter is not trained on a per-species basis, as it is a general pattern matcher. Though this means the algorithm does not have for example humpback whale-specific knowledge when matching humpback flukes, it is advantageous that HotSpotter does not need an extensive catalog of training images in order to be applied to a new species. Matching these SIFT descriptor vectors between images results in particular regions being found by the algorithm to correspond between the query and candidate matches. These corresponding regions are dubbed "hotspots", and the likelihood of a correct match grows quickly with the number and distinctiveness of hotspots that are found. Post-processing eliminates hotspots that are spatially inconsistent with the others before a final ranking of potential catalog matches. In Flukebook, HotSpotter is applied to humpback whale flukes, where it is very effective at matching pigmentation patterns (Flynn et al. 2017), illustrated in Fig. 6. It is also deployed here for other cetaceans with pigment and scar patterns, such as Atlantic spotted dolphins (*Stenella frontalis*), killer whales (*Orcinus orca*), Risso's dolphins (*Grampus griseus*), North Atlantic and southern right whales (*Eubaleana glacialis* and *E. australis*), and gray whales (*Eschrichtius robustus*).

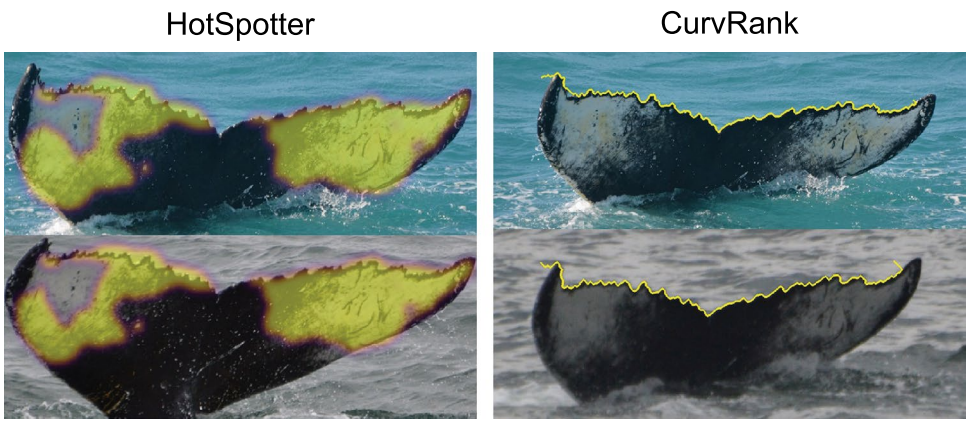


**Fig. 6** The same humpback whale fluke matched using two algorithms. On top of both comparisons is the new unidentified image, with best-matching database photo at the bottom. Left match made by HotSpotter. The highlighted regions show areas of "hotspots" distinctively matched between the two images. Right, trailing edges matched and extracted by CurvRank. The highlighted segment is the extracted trailing edge. Note that each algorithm returns a different best-matching image, reinforcing the benefit of multiple algorithms. While both algorithms were provided with the same *Annotation*, the results from CurvRank show a wider cropped version of the image for user interface reasons only. Colored algorithm highlights generated in Flukebook, photos courtesy Elding Whale Watch

### CurvRank and dynamic time warping

Three generations of related curvature-based matching algorithms are used in Flukebook to match individuals based on the distinctive trailing edges of their flukes and dorsal fins. These types of algorithms have been widely used for whale flukes and dolphin dorsal fins since the onset of computer-assisted photo-identification (e.g., Whitehead 1990; Beekmans et al. 2005). This family of algorithms includes a cetacean-ID-specific implementation (Jablons 2016) of dynamic time warping (DTW; Berndt and Clifford 1994), as well as two versions of CurvRank (Weideman et al. 2017, 2020). The latter two algorithms use a digital-curvature-based encoding of the trailing edge contour of flukes or dorsal fins, and match these using a local Naive Bayes nearest neighbor algorithm (McCann and Lowe 2012). CurvRank learns a contour appearance model from training data that requires only coarse hand-tracing of edges, which is hand-annotated by users in a web interface similar to the one used to train detection models (Fig. 5). These algorithms were originally applied to humpback whale flukes (Fig. 6), but more recently additional CurvRank models have been trained and deployed on Flukebook, matching (1) dolphin and baleen whale dorsal fin trailing edges and (2) the distinct dorsal ridge bumps on gray whales. We deploy multiple generations of these related techniques side-by-side because, while newer versions have the latest technology, earlier versions are better-studied and understood by the research community and are valuable in comparison to the newest methods.

### FinFindR

While it is another trailing edge matcher, finFindR has a distinct architecture from the CurvRank family of algorithms. Developed by Thompson et al. (2019) in R (R Core Team 2020), the Flukebook team collaborated with the developer to integrate it into Flukebook. It runs in a separate Docker container from WBIA and is accessed via a REST API. FinFindR is deployed on bottlenose dolphin dorsal fins where notches and scars, and to a lesser extent the overall shape of the fin, are used to identify individuals. Flukebook also uses this algorithm for other delphinid species including common dolphins (*Delphinus delphis*), spinner dolphins (*Stenella longirostris*), Indian Ocean humpback dolphins (*Sousa plumbea*), false killer whales (*Pseudorca crassidens*), and Risso's dolphins; as well as baleen whale dorsal fins for both fin (*Balaenoptera physalus*) and blue whales (*B. musculus*).

### Deepsense.ai

The U.S. National Oceanic and Atmospheric Administration (NOAA), in partnership with the New England Aquarium, hosted a competition for automated North Atlantic right whale ID algorithms on the popular machine learning competition platform Kaggle. The winning method was developed by Polish machine learning firm Deepsense.ai using a subset of images from the North Atlantic Right Whale Catalogue (Hamilton and Martin 1999). This algorithm was found to match whales from this catalog quickly and with 87% top-1 accuracy, using a series of convolutional neural networks (Bogucki et al. 2018). Special-purpose feature detectors orient and draw a bounding box around the head of each whale, which are passed to a classifier with one class for each individual. This algorithm has also been trained separately by developers at Wild Me on a catalog of southern right whales, resulting in separate North Atlantic and southern right whale classifiers.

### Kaggle7

Another Kaggle competition solution, dubbed "Kaggle7", was selected for its relatively straightforward implementation and ability to train quickly (original GitHub repository at https://github.com/ducha-aiki/whale-identification-2018). It uses components of the DTW algorithms to predict the end points of the fluke and the notch, which help to precondition the image prior to training and ID. Like the Deepsense.ai algorithm (and unlike other algorithms listed here), this algorithm is a fixed classifier, meaning it is only able to match those individuals that were in its training set. In the case of Kaggle7, this training set consists of pacific humpback whales uploaded by the Cascadia Research Collective (CRC), a set of 1701 individuals. Despite training only on CRC data, all users have access to this matcher.

### Pose-invariant embeddings

An algorithm dubbed PIE developed by Olga Moskvyak and colleagues (Moskvyak et al. 2019) uses triplet loss methods (Schroff et al. 2015) to train a deep learning encoding method that maps an image into a feature vector. Query and catalog image are matched based on distances between their associated feature vectors. This allows PIE to accommodate new individuals as they are added to the database. PIE is trained on a per-species basis, and originally two models were developed for manta rays and humpback whale flukes (two separately trained models). It has now been trained by both Flukebook developers on killer whales, sperm whales, and North Atlantic right whales (this model is applied on both North Atlantic and southern right whales), and by the original algorithm developer (O.M.) on gray whales. This is

a promising new technique whose use will continue to grow as more species-specific models are trained.

## Confirming matches for all algorithms: humans in the loop

Since individual identification is the crucial data point in photographic mark-recapture, which generates abundance estimates that can impact local, national, and international conservation policy, it is important that researchers use their expert judgement to confirm or reject potential matches identified through computer vision. As a result, each matching algorithm in Flukebook returns a ranked list of candidate matches that are displayed to the user, who then selects and confirms the correct match. Thus, the data owners and expert researchers who use Flukebook maintain ultimate authority and responsibility for the data products, as well as a continuous familiarity with the catalogs. Because matches can only be confirmed by users, cutting edge algorithms can be introduced to the platform for testing by users, alongside more established algorithms whose performance has been rigorously evaluated by biologists, without compromising the integrity of the results that are used in population analyses.

Because each match is confirmed by a user, and because many users intend to identify every possible animal that was observed during a mark-recapture study with the highest degree of confidence possible, algorithm accuracy figures (discussed further in the following section as well as Table 3) are better understood as a measure of time saved by researchers than as a measure of the accuracy of those researchers' ultimate data products. If a researcher finds no matches to a photo using the available algorithms on Flukebook, they may choose to confirm that negative result via manual photo identification, either using gallery and search tools on Flukebook, including a side-by-side photo comparison tool for manual matching, or existing tools that the researcher used before adoption of this platform. It is important to understand Flukebook and automated photo ID in general as a tool used by researchers, rather than a tool that produces research on its own; the platform places the ultimate responsibility for data accuracy on its users.

For species and features with multiple algorithms available for matching, such as humpback whale flukes, the results from each algorithm are presented separately but on the same result page for comparison. Algorithm blending, which takes the ranked result lists of multiple algorithms and combines them into a unified list of candidates, has been explored but not implemented to date. This is an interesting avenue of future research. At present, there is value in users getting to know each algorithm's strengths and weaknesses and weighing those, along with the distinct features visible in each photo (for example, a pattern matching algorithm is not terribly useful on an all-black humpback fluke, but very useful on a half-white half-black fluke). Future efforts in blending should then consider the accuracy of each algorithm on a given species, and possibly the characteristics of a given query image as well.

**Table 3** Reported performance numbers for algorithms in the Flukebook identification pipeline

| Species | Feature | Algorithm | Performance (top-*k* score) | Reference |
|---|---|---|---|---|
| *Megaptera novaeangliae* | Tail fluke | HotSpotter | 69%, 90% (top-1, top-2) | Flynn et al. (2017) |
| *Megaptera novaeangliae* | Tail fluke | Dynamic time warping | 40%, 49%, 53% (top-1, top-5, top-12) | Wiedeman et al. (2017) |
| *Megaptera novaeangliae* | Tail fluke | CurvRank v2 | 85%, 89%, 91% (top-1, top-5, top-12) | Wiedeman et al. (2020) |
| *Megaptera novaeangliae* | Tail fluke | Pose-invariant embeddings (PIE) | 63%, 88%, 93% (top-1, top-5, top-10) | Moskvyak et al. (2019) |
| *Megaptera novaeangliae* | Tail fluke | CurvRank v1 and Hotspotter Combined | 88%, 92%, 98% (top-1, top-5, top-12) | Wiedeman et al. (2017) |
| *Tursiops truncatus* | Dorsal fin | CurvRank v1 | 95%, 97%, 98% (top-1, top-5, top-12) | Wiedeman et al. (2017) |
| *Tursiops truncatus* | Dorsal fin | finFindR | 88%, 94%, 97% (top-1, top-10, top-50) | Thompson et al. (2019) |
| *Eubalaena glacialis* | Head, aerial (drone or plane) photo | Deepsense.ai | 87%, 95% (top-1, top-5) | Bogucki et al. (2018) |

The performance column includes only previously published figures when available. Because results are returned in the format of a ranked candidate list, "top-*k* score" means the percentage of results where the correct match is returned at rank *k* or lower, e.g. top-3 score indicates the percentage of correct matches returned in the top 3 candidates. Different publications use different values for *k*, specified in parentheses in the table

### Performance

It should be noted that a computer vision algorithm does not have an inherent level of accuracy because its accuracy is always dependent on the data that are provided. A top performing algorithm will not be as accurate on blurry images as on perfectly framed and in-focus exemplar images. This is why in the broader field of computer vision, there are established reference datasets such as the MNIST dataset of hand-written digits (LeCun et al. 1998, dataset available at http://yann.lecun.com/exdb/mnist) which are used to compare algorithms against each other. These do not yet exist in conservation-focused computer vision yet; and as such, the reported accuracy numbers in Table 3 reflect both the algorithms being presented as well as the datasets described in the cited sources. Due to the ongoing development of several of the species-specific pipelines, performance metrics cannot be reported here. See the Discussion section for further comment on this.

## Discussion

If we are to achieve data-driven decision making based on accurate abundance trajectories for the conservation of threatened species, then we need the capacity to rapidly iterate population estimates to ensure regular evaluation and adjustment of our action plans. This requires reconciling disparate datasets, which for many species must occur across wide geographic regions and with the integration of opportunistic data from various stakeholders, in order to assess local, regional, and international population trends in support of national policy makers and multinational conservation agreements to protect those species.

In order to achieve these goals, Flukebook intercedes with three different stakeholder groups. The platform most directly benefits scientists through tools for photo-identification and data management. Automation of once-manual tasks enables greater efficiency in processing and curating photo-ID data. These lead to both cost savings in undertaking research and time savings in reporting on field studies. In addition, the collaborative platform allows questions to be asked on larger datasets than would otherwise be possible without significant logistical challenges, a noteworthy example of which can be seen in the collaborative global research done on the Wildbook for whale sharks (Norman et al. 2017) and the early-stage collaborative work already underway on Flukebook by consortia such as the Arabian Sea Whale Network and (Blount et al. 2020) the Caribbean Marine Mammals Preservation Network (CARI'MAM; Vail and Borobia 2020). Larger labs or organizations can enable multiple graduate students, interns, or researchers to work on the same datasets from remote locations. Furthermore, the cloud-based platform is backed up nightly and stored on multiple servers, limiting data loss or corruption and providing a long-term data repository that is accessible from anywhere in the world.

For naturalists and community scientists, Flukebook enables contribution to the active process of research, interaction with the scientists undertaking it, and engagement with the feedback of which individual they saw while at sea. The platform also provides them the search functions to explore more of the species they have sighted and visualizations such as the social network diagram enable the users to learn more about the social lives of the individual they have sighted and submitted.

Finally, Flukebook creates a data pipeline which reaches managers and governments in formats that they are already using for population assessment or understanding changing distributions which may result from global climate change or localized anthropogenic impacts, such as those which may be associated with coastal construction or offshore energy development. Many management agencies, including governments, seek to implement adaptive management, a systematic process for improving management by learning from outcomes and subsequently adjusting practices accordingly. A major challenge to the realization of adaptive management is the requirement for rapid iteration of population or ecological models and to ultimately demonstrate trade-offs in various decision scenarios. Machine learning based workflows contained within Flukebook allow for the collection and accelerated analysis of large volumes of data to advance science-informed decision-making and enable dynamic management.

On the whole, Flukebook is building a community of recreational wildlife enthusiasts, community scientists, naturalists, dedicated researchers, policy makers, and managers through a platform that enables all to collaborate to support the conservation and management of highly mobile marine species.

## Limitations

Flukebook, as with any toolkit built to save the user time and costs, empowers the user to use advanced methods but does not replace researchers entirely. Flukebook accelerates photo identification, but in order to undertake more detailed analysis such as modelling population estimates and social structure, a knowledgeable researcher must make important decisions regarding experiment design, data standards, and statistical inference offline and outside Flukebook. Importantly, they must be familiar with the data on the platform in order to make these decisions. As such, users conducting scientific research based on data from Flukebook should be careful to understand Flukebook and use its tools correctly.

While Flukebook's standard data architecture allows for researchers to share data, match against each other's photos, and even export each other's data for outside analysis (when collaborations are in place); there are limits to the amount of data standardization that can be achieved when hosting such diverse data as this platform. Data fields such as categorical behavior observations (e.g., feeding, breaching, mating), metrics of photo quality or heterogeneity of identifiability (e.g. scales from 0 to 5 vs 0 to 10 or those which include markedness when other users do not), and individual naming conventions (e.g. individual animals in Flukebook may have different nicknames given by different research groups) can differ significantly between research teams. Flukebook offers tools for users to efficiently standardize these fields across one study or even a consortium of researchers, such as the *Organizations* described in the *Data Ownership and Collaboration Model* section above, but a user is not forced to accept conventions which they might not share. This has the benefit of allowing diverse data to be imported onto Flukebook, with the potential drawback that multiple data standards will co-exist on the platform as separate fields in the database. In some senses, this is a limitation which is not new to this platform. It has always been true that if multiple catalogs of cetacean observations are to be combined into a single study, there may be some data reconciliation and standardization required to make that analysis possible given that field-wide standards do not always exist. Flukebook provides data sharing tools, a flexible database structure to import varied metadata fields, standardized export formats, and photo ID algorithms that empower and encourage researchers to collaborate in the first place. In some ways, Flukebook reflects the diversity of approaches which exist in cetacean photo-identification, and the analyses which rely upon it, while facilitating efforts without enforcing data standards which are not universal accepted within the field.

Determining the accuracy of a novel algorithm when applied to a particular species and specific body part comes with several challenges. First, every potential match must be confirmed by a researcher in order to be recorded, and to determine the rank at which an algorithm returned the correct result (for example whether the matching individual was the first, or tenth suggested candidate). Unless a historical manually matched catalog can be used as reference, this process requires dedicated effort from expert biologists who upload and process their data on the platform in partnership with the algorithm developers or Flukebook team. Second, algorithm accuracy is always deeply dependent on the data on which it is queried. A catalog with one photo per individual will generally have lower matching accuracy than one with ten photos per individual; a catalog with higher quality data will see better algorithm performance than one of lower quality; a population with fifty total individuals will be easier to match within than one with thousands of individuals. Thus, any accuracy metric should be contextualized by the data which produced those values, in contrast to the data one would like to query moving forward. It would not be wise for any researcher to assume an algorithm will have the same accuracy on their data as it does on another group's data without testing this assumption. While we present previously published accuracies in Table 3, we made a deliberate decision to deemphasize accuracy as a metric for the algorithms presented in this paper for these reasons. Performing a robust analysis for all 37 species-feature-algorithm pipelines shown in Fig. 4 was outside the scope of this paper, though some of these pipelines are under active analysis by researcher-users which will hopefully be published in the future. Flukebook is an ever-growing platform and when users with large existing catalogs join Flukebook, especially when using newly developed ID pipelines or species new to the platform, we encourage these biologists (and often work with them) to perform accuracy testing, querying a sample of their own previously identified photos to empirically generate an accuracy estimate on their own data.

A final point regarding accuracy is that since every match suggested by Flukebook's algorithms must be confirmed by a user who is looking at both the query and candidate photos, an algorithm will not introduce errors into the data regardless of its purported accuracy rating. When a user confirms negative matches (potentially new individuals from photos queried which found no positive match in the database) using manual matching methods, these tools will save time with no accuracy detriment compared to previous fully manual methods. Flukebook enables the user to invest the majority of their time matching on these rare challenging matches by rapidly identifying the majority of well known, unchanged individuals which are identified regularly, thereby creating a more efficient use of expert time. Both efficiency and accuracy improvements have been robustly documented using various automatic cetacean matching systems over the years (e.g. Mizroch et al. 1990; Whitehead 1990; Beekmans et al. 2005; Cheeseman et al. 2022). As with our discussion of the difficulties in extrapolating algorithm accuracy above, it is important that researchers measure these factors with the datasets, algorithms, and species relevant to their own work. In the context of Flukebook, this will require further study to quantify expected improvements in efficiency and accuracy for each specific pipeline on the platform, some of which is already underway. Automated identification is not a black box that produces scientific results on its own, but a tool that must be understood and used carefully by the scientist.

Lastly, Flukebook is a cloud-based, online platform and we acknowledge the limitations of using such a platform for those who are working in the majority of countries with emerging economies, and who may have limited or no access to high-speed internet. Given that image processing and the algorithms

run on servers and not the user's device, the system can be used as long as images can by uploaded and downloaded.

## Future directions

An active area of development in the Wildbook software ecosystem is that of *intelligent agents*, or web-crawling bots designed to extract relevant sightings data, including photos or video captures, from publicly available posts on social media. This flips the previous approach in that the platform actively seeks out potential data contributors rather than passively expecting the users to submit. Tourists and ecologically minded members of the community are eager to upload photos and videos of animals they have seen in the wild, without realizing that these might represent valuable ecological data. Flukebook and its sister platforms like Sharkbook.ai provide not only the technological backbone to process such data, but an authoritative catalog of the individual animals who might appear in this content, linked to the researchers who would be interested in these sightings.

The first Wildbook intelligent agent deployed was a YouTube crawler integrated with Sharkbook.ai (then under the name "Whaleshark.org") in 2018. This agent searches YouTube for videos of the focal species, then processes key frames from the video to find clear photos of the animals and identify the individuals depicted. The agent also extracts available information from the video file and metadata; as well as reading the text available on YouTube to get the date, time, and location of the sighting. If the location, date and time cannot be automatically extracted, the bot actually asks the submitter for this information via the YouTube comment section, in multiple languages based on the language of the original post. This feature has yet to be implemented in Flukebook but is coming in a future release.

An intelligent agent active on Flukebook is 'Tweet-a-whale' which is currently in beta release (the phase of software development when a feature is fully implemented but undergoing usability testing). Twitter users can submit a photo for identification directly to Flukebook by tweeting a photo and tagging the account @tweetawhale. The agent running the account will extract available metadata from the text of the tweet and submit the image to Flukebook. The image goes through detection and identification, and resulting candidate matches are tweeted back to the original submitter. Users are then required to log in to Flukebook to approve these potential matches, meaning a member of the public who is not a Flukebook user cannot confirm match results on their own. Data submitted this way are then available to all researchers on Flukebook and labeled as data submitted by the public, unless the twitter user is already affiliated with a Flukebook user account, in which case the data are controlled by that user. To ensure privacy as well as conform with regulations such as GDPR (https://gdpr-info.eu/), publicly submitted data are anonymized with respect to their submitter. Since these publicly submitted data are labeled as such on Flukebook, researcher users can choose whether these data are included in analyses or exports they perform. As many scientists who have worked with community science and especially social media submissions know, the quality of data submitted by the general public might not be of the same caliber as that submitted by researchers.

The data and matching tools on Flukebook provide opportunities to perform studies regarding bias in mark recapture and related concepts such as individual matchability. For example, work is currently underway to study the relationship between changing patterns on gray whales and the matchability of their photos separated by large amounts of time. Because gray whale patterns change considerably as scars and marks accumulate and fade on their skin, one expects that photos of the same individual at close periods in time will be more easily matched than those separated by a long period, which will be quantified using Flukebook data in a partnership between Flukebook developers and the data owners. We hope other studies such as those relating matchability to subjective image quality rankings will be performed by our users in the future. While Flukebook does not answer these questions of recapture bias on its own, nor does it make them obsolete, it makes them more tractable to its researcher-users.

Next generation features will also include species detection for community science submissions, and an overhaul of the user interface and experience designed for a modern web. This entirely new frontend to Wildbook is under active development, and will be deployed first on the Wildbook for Zebras (https://zebra.wildbook.org/) in early 2022. There will subsequently be a gradual rollout to other Wildbooks, such that when the new user interface is deployed on Flukebook, it will have already been extensively used and tested on other platforms.

## Conclusions

By leveraging cutting-edge computer vision with a global data management platform, Flukebook allows for easier storage of data and more rapid photo-identification in cetacean studies while bringing together researchers and community scientists to create bigger datasets. Flukebook's position within the open-source family of Wildbook platforms provides major benefits for conservation technology development, where returns on technology investment are multiplied across every Wildbook platform. This computer vision pipeline works with real-world conditions, allowing for broad contributions from both research professionals and amateur naturalists. This process gives greater confidence that we know who the animals are and where they have been. Getting this baseline information correct allows government and managers to focus on policies that can bring about actionable change.

## Appendix

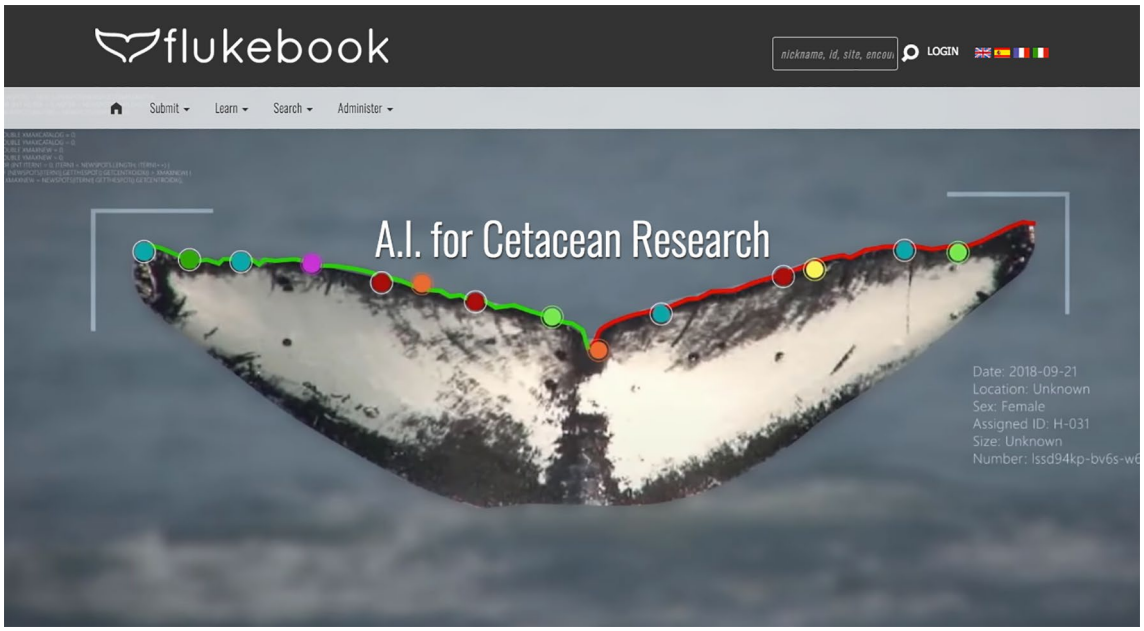


**Fig. A1** The landing page for the Flukebook platform (https://www.flukebook.org), which applies computer vision algorithms and deep learning to identify and track individual cetaceans in space and time

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42991-021-00221-3.

**Acknowledgements** The authors wish to thank the many volunteers, researchers, NGOs and government partners, whale watch companies, and community naturalists who make up the users of Flukebook. Dr. Ann Pabst of UNCW and Dr. Doug Nowacek of Duke University who led the BOEM cooperative agreement which funded the creation of this manuscript. Virgil Zetterlind and the team at Conserve.IO who developed the Ocean Alert app. Shauna Bingham and the Channel Islands Naturalist Corps at the Channel Islands National Marine Sanctuary for piloting the new Ocean Alert app-based pipeline. We also thank major collaborators and contributors of datasets which were foundational to algorithm development and website functionality: Arabian Sea Whale Network, CARI'MAM, Cascadia Research Collective, The Dominica Sperm Whale Project, Elding Whale Watch, IndoCet, Globice, MarEcoTel, New England Aquarium, NOAA, Norwegian Orca Survey, Oceania Project, Sarasota Dolphin Research Program, Wild Dolphin Project, and Whale Watching Akureyri. We also wish to thank our anonymous reviewers for feedback which strengthened this manuscript greatly, as well as the editors of Mammalian Biology for putting together this special issue.

**Author contributions** SG, JH and JJL conceived the Flukebook platform based on the Wildbook platform; Flukebook's codebase was developed by DB, JVO, JP, CKi, BS, TS, MF, JH; algorithms were developed by DB, JP, CKh, JT, OM, TBW, CVS; funding for the development of Flukebook was secured by SG, GM, VD, CKh, TBW, CVS, JH, and JJL; DB and SG wrote the manuscript; all authors reviewed, edited, and approved the final manuscript.

**Funding** Funding for the development of Flukebook was provided by the Gordon and Betty Moore Foundation, NOAA, Microsoft, the National Science Foundation, the Pineapple Fund, H2O.ai, IndoCet, Arabian Sea Whale Network, the Waitt Foundation, the Adobe Foundation, NVIDIA, U.S. Department of Interior, Bureau of Ocean Energy Management, Environmental Studies Program through cooperative agreement #M17AC00013 as well as the Organization of American States Western Hemisphere Migratory Species Initiative.

**Code availability** Code for the Flukebook platform is available on the 'flukebook' branch of the open-source Wildbook repository available on GitHub at https://github.com/WildMeOrg/Wildbook/tree/flukebook. The Wilbook Image Analysis computer vision server is available at https://github.com/WildMeOrg/wildbook-ia.

## Declarations

**Conflict of interest** The authors declare no conflicting or competing interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References


Barlow J, Calambokidis J, Falcone EA, Baker CS, Burdin AM, Clapham PJ, Ford JK, Gabriele CM, LeDuc R, Mattila DK, Quinn TJ (2011) Humpback whale abundance in the North Pacific estimated by photographic capture-recapture with bias correction from simulation studies. Mar Mamm Sci 27(4):793–818. https://doi.org/10.1111/j.1748-7692.2010.00444.x

Beekmans BWPM, Whitehead H, Huele R, Steiner L, Steenbeek AG (2005) Comparison of two computer-assisted photo-identification methods applied to Sperm Whales (*Physeter macrocephalus*). Aquat Mamm 31(2):243–247. https://doi.org/10.1578/AM.31.2.2005.243

Berndt D, Clifford J (1994) Using dynamic time warping to find patterns in time series. In: Proc 3rd Int Conf on Knowl Discov and Data Min, KDD workshop, vol 10, issue 16, pp 359–370

Bigg M (1982) An assessment of killer whale (*Orcinus orca*) stocks off Vancouver Island, British Columbia. Rep Int Whal Comm 32:655–666

Blount D, Minton G, Khan C, Levenson J, Dulau V, Gero S, Parham J, Holmberg J (2020) Flukebook: continuing growth and technical advancement for cetacean photo identification and data archiving, including automated fin, fluke, and body matching. Rep Int Whal Comm SC/68A/SH/07

Bogucki R, Cygan M, Khan CB, Klimek M, Milczek JK, Mucha M (2018) Applying deep learning to right whale photo identification. Conserv Biol 33(3):676–684. https://doi.org/10.1111/cobi.13226

Bonner S, Huang X (2018) RWildbook: interface for the 'wildbook' wildlife data management framework. R package: version 0.9.3. https://CRAN.R-project.org/package=RWildbook

Cheeseman T, Southerland K, Park J, Olio M, Flynn K, Calambokidis J, Jones L, Garrigue C, Jordán AF, Howard A, Reade W, Neilson J, Gabriele C, Clapham P (2022) Advanced image recognition: a fully automated, high-accuracy photo-identification matching system for humpback whales. Mamm Biol (Special Issue) 102(2). https://doi.org/10.1007/s42991-021-00180-9

Crall J, Stewart C, Berger-Wolf T, Rubenstein D, Sundaresan S (2013) HotSpotter: patterned species instance recognition. In: IEEE Workshop Appl Comput Vis (WACV) 2013, pp 230–237. https://doi.org/10.1109/WACV.2013.6475023

Flynn K, Calambokidis J, Weideman H, Crall J, Jablons Z, Stewart C, Kingen C, Van Oast J, Holmberg J (2017) Testing of two new automated fluke identification algorithms and comparison to non-automated methods for humpback whales. Proc 22nd Bienn Biol Mar Mamm

Gero S, Milligan M, Rinaldi C, Francis P, Gordon J, Carlson C, Steffen A, Tyack P, Evans P, Whitehead H (2014) Behavior and social structure of the Sperm Whales of Dominica, West Indies. Mar Mamm Sci 30:905–922. https://doi.org/10.1111/mms.12086

Hamilton PK, Martin SM (1999) A catalog of identified right whales from the western North Atlantic: 1935–1997. N Engl Aquar

Hammond PS, Mizroch SA, Donovan GP (1990) Individual recognition of cetaceans: use of photo identification and other techniques to estimate population parameters. Rep Int Whal Comm (Special Issue 12)

Hillman GR, Wursig B, Gailey GA, Kehtarnavaz N, Drobyshevsky A, Araabi BN, Tagare HD, Weller DW (2003) Computer-assisted photo-identification of individual marine vertebrates: a multi-species system. Aquat Mamm 29(1):117–123. https://doi.org/10.1578/016754203101023960

IUCN (2020) The IUCN red list of threatened species. Version 2020-3. https://www.iucnredlist.org. Accessed 26 Jan 2021

Jablons Z (2016) Identifying humpback whale flukes by sequence matching of trailing edge curvature. Master's thesis, Rensselaer Polytechnic Inst

Kaschner K, Quick NQ, Jewell R, Williams R, Harris CM (2013) Global coverage of cetacean line-transect surveys: status quo, data gaps and future challenges. PLoS One 7(9):e44075. https://doi.org/10.1371/journal.pone.0044075

Katona SK, Whitehead HP (1981) Identifying humpback whales using their natural markings. Polar Rec 20(128):439–444. https://doi.org/10.1017/S003224740000365X

Laake J (2013) RMark: an R interface for analysis of capture-recapture data with MARK. Alaska Fish Sci Cent, NOAA Natl Mar Fish Serv. http://www.afsc.noaa.gov/Publications/ProcRpt/PR2013-01.pdf

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791

Lowe DG (1999) Object recognition from local scale-invariant features. Proc Int Conf Comput vis 2:1150–1157. https://doi.org/10.1109/ICCV.1999.790410

McCann S, Lowe DG (2012) Local naive Bayes nearest neighbor for image classification. In: IEEE Conf on Comput Vis and Pattern Recognit, pp 3650–3656. https://doi.org/10.1109/CVPR.2012.6248111

Mizroch S, Beard JA, Lynde M (1990) Computer assisted photo-identification of humpback whales. Rep Int Whal Comm 12:63–70

Moskvyak O, Maire F, Armstrong AO, Dayoub F, Baktashmotlagh M (2019) Robust re-identification of manta rays from natural markings by learning pose invariant embeddings. ArXiv. https://arxiv.org/pdf/1902.10847.pdf

Norman BM, Holmberg JA, Arzoumanian Z, Reynolds SD, Wilson RP, Rob D, Pierce SJ, Gleiss AC, de la Parra R, Galvan B, Ramirez-Macias D, Robinson D, Fox S, Graham R, Rowat D, Potenski M, Levine M, Mckinney JA, Hoffmayer E, Dove ADM, Hueter R, Ponzo A, Araujo G, Aca E, David D, Rees R, Duncan A, Rohner CA, Prebble CEM, Hearn A, Acuna D, Berumen ML, Vázquez A, Green J, Bach SS, Schmidt JV, Beatty SJ, Morgan DL (2017) Undersea constellations: the global biology of an endangered marine megavertebrate further informed through citizen science. Bioscience 67(12):1029–1043. https://doi.org/10.1093/biosci/bix127

Parham J, Stewart C, Crall J, Rubenstein D, Holmberg J, Berger-Wolf T (2018) An animal detection pipeline for identification. In: IEEE Winter Conf on Appl Comput Vis (WACV), pp 1075–1083. https://doi.org/10.1109/WACV.2018.00123

Payne R (1986) Long term behavioral studies of the southern right whale (*Eubalaena australis*). Rep Int Whal Comm 10:161–167

R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Rondinini C, Di Marco M, Visconti P, Butchart SH, Boitani L (2014) Update or outdate: long-term viability of the IUCN red list. Conserv Lett 7(2):126–130. https://doi.org/10.1111/conl.12040

Schroff F, Kalenichenko D, Philbin J (2015) FaceNet: a unified embedding for face recognition and clustering. In: IEEE Conf on Comput Vis and Pattern Recognit (CVPR), pp 815–823, https://doi.org/10.1109/CVPR.2015.7298682

Stanley R (1995) Darwin: identifying dolphins from dorsal fin images. Senior Thesis, Eckerd Coll

Thompson JW, Zero VH, Schwacke LH, Speakman TR, Quigley BM, Morey JSM, McDonald TL (2019) finFindR: computer-assisted recognition and identification of bottlenose dolphin photos in R. BioRxiv. https://doi.org/10.1101/825661

Urian K, Gorgone A, Read A, Balmer B, Wells RS, Berggren P, Durban J, Eguchi T, Rayment W, Hammond PS (2015) Recommendations for photo-identification methods used in capture-recapture models with cetaceans. Mar Mamm Sci 31:298–321. https://doi.org/10.1111/mms.12141

Vail C, Borobia M (2020) Implementation of the action plan for marine mammals in the wider Caribbean region: a scientific and technical analysis. Report to UN Environment, Caribbean Environment Programme, Specially Protected Areas and Wildlife Regional Activity Centre. http://gefcrew.org/carrcu/SPAWSTAC9/InfoDocs/WG.42-INF.29_Addendum1-en.pdf. Accessed 11 Dec 2021

Weideman HJ, Jablons Z, Holmberg J, Flynn K, Calambokidis J, Tyson RB, Allen JB, Wells RS, Hupman K, Urian K, Stewart CV (2017) Integral curvature representation and matching algorithms for identification of dolphins and whales. In: IEEE Int Conf on Comput Vis, pp 2831–2839. https://doi.org/10.1109/ICCVW.2017.334

Weideman HJ, Stewart C, Parham J, Holmberg J, Flynn K, Calambokidis J, Paul DB, Bedetti A, Henley M, Pope F, Lepirei J (2020) Extracting identifying contours for African elephants and humpback whales using a learned appearance model. In: IEEE/CVF Winter Conf on Appl Comput Vis (WACV), pp 1276–1285. https://doi.org/10.1109/WACV45572.2020.9093266

Wells RS, Scott MD (1990) Estimating bottlenose dolphin population parameters from individual identification and capture-release techniques. Rep Int Whal Comm 12:407–415

Whitehead H (1990) Computer assisted individual identification of sperm whale flukes. Rep Int Whal Comm 12:71–77

Würsig B, Würsig M (1977) The photographic determination of group size, composition, and stability of coastal porpoises (*Tursiops truncatus*). Science 198(4318):755–756. https://doi.org/10.1126/science.198.4318.755

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

**Drew Blount[1] · Shane Gero[2,3,4] · Jon Van Oast[1] · Jason Parham[1,5] · Colin Kingen[1] · Ben Scheiner[1,6] · Tanya Stere[1] · Mark Fisher[1] · Gianna Minton[7] · Christin Khan[8] · Violaine Dulau[9] · Jaime Thompson[10] · Olga Moskvyak[11] · Tanya Berger-Wolf[12,13,14,15] · Charles V. Stewart[5] · Jason Holmberg[1] · J. Jacob Levenson[16,17]**

1 Wild Me, Portland, OR 97217, USA

2 Department of Biology, Carleton University, Ottawa, ON K1S 5B6, Canada

3 Zoophysiology, Institute for BioScience, Aarhus University, Aarhus C, 8000 Aarhus, Denmark

4 The Dominica Sperm Whale Project (http://www.thespermwhaleproject.org), Roseau, Dominica

5 Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

6 H2O.ai, Mountain View, CA 94043, USA

7 Megaptera Marine Conservation, The Hague, The Netherlands

8 National Oceanic and Atmospheric Administration, Northeast Fisheries Science Center, Woods Hole, MA 02543, USA

9 GLOBICE-Reunion, Reunion, 97410 Saint Pierre, France

10 Valdez AI Consulting, LLC, Laramie, WY 82072, USA

11 School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD 4000, Australia

12 Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA

13 Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210, USA

14 Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210, USA

15 Translational Data Analytics Institute, The Ohio State University, Columbus, OH 43210, USA

16 U.S. Department of Interior, Bureau of Ocean Energy Management, Sterling, VA 20166, USA

17 School of the Environment, University of Massachusetts Boston, Boston, MA 02125, USA